KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스

김문형 ○ 장하연 ○ 조유미 ○ 신효필

서울대학교 언어학과

{likerainsun, hyan05, jmocy84, hpshin}@snu.ac.kr

KOSAC: Korean Sentiment Analysis Corpus

Munhyong Kim ○ Ha-Yeon Jang ○ Yu-Mi Jo ○ Hyopil Shin

Seoul National University, Linguistics Department

요 약

본 연구는 자동 감정 및 의견 분석 연구에 필수적인 한국어 감정 코퍼스를 구축하는 방법을 제안하고 이에 따라 실제 구축된 코퍼스를 분석한 결과를 소개한다. 세종 구문분석 코퍼스로부터 선별한 332개 신문 기사, 7744 문장을 주석 대상으로 삼아 본 연구에서 설정한 한국어 감정 주석 스키마를 이용하여 총 7744개의 문장에 17582개의 감정 표현을 주석하였다. 주석된 감정 표현의 속성들을 사용하여 전체 문장의 주관성과 극성 분류 실험을 SVM 모델을 사용하여 실험한 결과 각각 65.72%, 82.52%의 정확도를 보여 한국어 감정 코퍼스의 활용 가능성을 보여주었다.

1. 서 론

현재 자연언어처리 분야에서는 화자의 감정이나 의견을 자동으로 인식하고 추출하려는 연구가 활발히 이루어지고 있다. 이런 연구는 일반적으로 한 문서의 전반적인 감정을 인식하고 분류하는 연구와 한 문장의 감정을 대상으로 하는 연구로 나뉠 수 있다. 예를 들어 영화평에서 평가자가 영화 전반에 대해 긍정적으로 평가하는지 혹은 부정적으로 평가하는 지를 자동으로 분류하는 연구는 문서 단위의 감정 분석의 하나이고 한 문장이 어떤 대상에 대해 긍정적인지 부정적인지 판단하는 연구는 문장을 대상으로 하는 감정 분석 연구이다[1].

자동 감정분석 연구는 긍정, 부정의 어휘를 분류한 후 문서 내에서 이런 극성 어휘들의 빈도를 연산하는 방법이 주를 이루어 왔으나 포괄적인 감정 표현을 학습하거나 감정 분석의 기초자료로 활용될 수 있도록 실제 감정표현들이 주석된 대량의 코퍼스 구축이 필요하다는 것이 주지되었다. 이는 텍스트 자동 감정 분석 시스템을 개발에 있어서 필요한 전자사전의 구축이나 기계 학습을 위한 트레이닝 데이터로 사용될 수 있을 뿐만 아니라 인간이 감정을 표현하는 방식에 대한 언어학적 연구를 위한 토대가 되기도 한다.

영어를 대상으로 구축된 감정 분석 코퍼스인 Multi-perspective Question Answering (MPQA) [1][2][3]는 약 10000 문장 안의 나타나는 감정 표현들에 감정 표현의 의미를 잘 나타낼 수 있는 주석 언어를 이용하여 주석하였다. 이 코퍼스는 감정, 오피니언 자동 분석 연구에서 기계 학습을 위한 학습 데이터로써 그리고 골드 스탠더드로써 중요한 역할을 하고 있다.

본 연구는 한국연구재단의 2년간(2011.5-2013.4)의 지원을 받아 한국어 감정분석 코퍼스 구축을 시작하였다. 한국어 감정 분석 코퍼스를 개발하기 위해서 먼저 감정 주석 체계의 개발이 이루어졌다. 이는 MPQA 코퍼스를 구축하는데 사용된 감정 표상 체계인 [1][2]에 기반하여 한국어의 특성에 알맞도록 새로운 속성들을 추가하고 주석에 용이하도록 전체적인 구조에 수정을 가하여 만들어졌다[4].

본 논문의 2장에서는 한국어 감정 주석 체계에 대해 설명하고, 3장에서 한국어 감정 코퍼스 구축 방식에 대해 소개한다.. 그리고 4장에서는 구축된 코퍼스의 주석 결과와 활용도에 대하여 논의한다. 마지막 5장은 결론 및 향후 과제이다.

2. 한국어 감정 주석 언어

한국어 감정 주석 언어(Korean Sentiment Markup Language, 이하 KSML)는 문장 전체에서 드러나는 서술자의 주관성을 표시하는 SUBJECTIVITY 태그, OBJECTIVITY 태그와 문장보다 작은 단위의 핵심 주관 표현들을 주석하는 SEED 태그로 구성된다. 주관 표현의 의미적 특성들은 SEED 태그의 다양한 속성값을 통해 최대한 상세히 주석된다.

표 1. SEED 태그 속성 목록

```
anchor: morpheme id(s)
id: tag id
nested-source: w-(morpheme id(s)|implicit|out)-...
          -(morpheme id(s)|implicit|out)
target: morpheme id(s)
type: direct-explicit, direct-speech, direct-action,
      indirect, writing-device
subjectivity-type: emotion-{pos,neg,complex,neutral},
            judgment-{pos,neg,complex,neutral},
            argument-{pos,neg,complex,neutral},
```

```
            agreement-{pos,neg,neutral},
            intention-{pos,neg},
            speculation-{pos,neg}, others
polarity: positive, negative, neutral, complex
intensity: low, medium, high
```

SEED 태그의 속성값을 간단히 소개하면, anchor 속성은 주석 대상이 되는 표현을 형태소 id를 통해 지시하고, nested-source 속성은 해당 주관 표현의 표출자와 그 주관성의 전달 경로를 기술한다. 기본적으로 동일 문서 내 모든 문장은 동일 서술자(writer)에 의해 기술되므로 'w' 출처가 항상 포함된다. target은 주관적 표현이 기술하는 대상 및 주제를 표시한다. type 속성은 주관성 표현 방식을 주석하는 것으로 크게 주관성이 술어를 통해 명시적으로 표현되는지(direct), 간접적으로 드러나는지(indirect), 일종의 텍스트 상의 장치를 이용해 주관성을 반영하고 있는지(writing-device)로 구분된다. 주관 표현의 구체적인 의미와 표출자의 태도는 subjectivity-type 속성으로 주석되며, 각 유형은 의해 분류되며, 각 유형은 방향성 표지(pos, neg, complex, neutral)를 통해 더욱 상세한 의미의 방향성을 반영한다. Polarity와 intensity는 각각 주관 표현의 극성과 강도를 나타내는 속성이다.

문장 전체의 주관성을 주석하는 SUBJECTIVITY 태그는 anchor, id, Polarity, intensity 만을, OBJECTIVITY 태그는 anchor, id 속성값만을 포함한다.

3. 한국어 감정 코퍼스 구축

코퍼스 구축을 위한 텍스트는 세종 구문 분석 코퍼스 중에서 조선 일보 생활, 사회면과 한국일보 한겨레 신문에서 332개 기사, 7744 문장을 선정하여 주석하였다. 구문 분석 코퍼스를 선택한 이유는 주석된 감정 표현 패턴의 구문 정보가 자동 감정 분석 시스템의 구축에 유용하게 사용될 수 있기 때문이다.

신문 기사에서 KSML을 사용하여 주석하는 것은 매우 많은 시간과 노동력을 요구하는 일이다. 따라서 이 연구에서는 한국어 감정 코퍼스 주석을 위해서 주석 텍스트를 클릭하고 선택된 영역에 대해 속성을 체크하여 데이터 베이스에 저장하는 그래픽 인터페이스의 툴을 개발했다[5]. 이 툴로 인해서 주석의 정확성을 매우 높일 수 있었고 주석자 간의 교차 분석도 수월하게 진행할 수 있었다.

주석하는 코퍼스의 신뢰도를 높이기 위해서는 보통 두 명 이상의 주석자가 하나의 텍스트를 주석하고 주석 결과 중에서 일치하지 않는 부분에 대하여 제 삼의 연구자가 판단하여 하나의 주석으로 결정하는 과정을 거친다. 하지만 본 연구에서는 세 명의 주석자가 각자 다른 텍스트를 주석하였고, 그 주석한 결과물을 서로 상호 교차 검토하고 잘못된 부분에 대하여 회의를 통하여 수정하는 방식을 택하여 좀 더 빠른 시간 안에 많은 양의 주석이 가능하도록 하였다.

4. 코퍼스 주석 결과 분석 및 활용

이렇게 구축된 한국어 감정분석 코퍼스의 주석 결과를 분석하고 이 데이터를 향후 어떻게 활용할 수 있는지를 살펴보기 위한 실험 및 그 그 결과는 다음과 같다.

4.1 코퍼스 주석 결과 분석

주석된 전체 7744 문장 중에서 2654개 문장이 주관적인 문장으로 그리고 5090개 문장이 객관적인 문장으로 주석됐다. 또한 전체 SEED 태그는 17582개로 하나의 문장에 평균 2.3개의 SEED 표현이 주석된 것을 의미한다. 주석된 코퍼스가 어느 정도 신뢰할 수 있을지 살펴보기 위한 척도로 빈도가 높은 감정 표현들 중에서 22개를 골라 그 표현들이 텍스트에서 나타난 빈도와 주석된 빈도 사이의 비율을 측정하였다. 그 결과 그 표현들이 평균 81%의 정확도로 주석된 것을 알 수 있었다. 이것은 주석자들이 얼마나 일관적으로 주석을 했는지 판단할 수 있는 하나의 척도라 할 수 있다.

주석된 SEED 태그에서 type과 Subjectivity-type의 속성의 비율을 확인할 수 있는 빈도 교차표는 다음과 같다.

표 2. type, Subjectivity-type의 교차 빈도표

	Agr.	Argu.	Emo.	Int.	Judg.	Oth.	Sp.
D.A.	1	8	73	8	41	1	0
D.E.	156	276	344	276	2740	40	157
D.S.	8	1150	22	28	86	7	13
IND	252	321	714	406	6079	22	61
W.D.	4	98	9	305	770	2935	171

위 표에서 보는 바와 같이 Judgment 타입의 SEED 태그가 다른 종류에 비해 많이 주석된 것을 볼 수 있다. 그리고 Direct 타입 중에서는 Direct-explicit이 가장 큰 비중을 차지하고 Direct 타입보다는 Indirect 타입의 태그가 더 많은 비중을 차지하고 있음을 살펴볼 수 있다.

4.2 코퍼스 활용

이번 연구를 통해 구축된 감정 코퍼스는 기계 학습을 통한 감정 분석 시스템 연구 시 학습 자료로 사용될 수 있다. 또한 한국어 감정 연구를 시행할 때 감정 분석 알고리즘 및 시스템의 효과 검증을 정확히 비교할 수 있는 공통 코퍼스로서 사용될 수 있을 것으로 생각된다.

그렇게 활용하기 위해서는 해당 코퍼스의 주석 정보가 실제로 문장의 감정을 분석하는 데에 유용한지를 확인해야 한다. 따라서 한 문장에 포함된 SEED 태그들에서 추출한 자질을 통하여 그 문장이 주관적인지 또는 객관적인지 판단하거나 긍정적인지 부정적인지 자동으로 판단하는 실험을 SVM light를 이용해 (linear kernel, default option) 시행했다. 실험의 트레이닝 데이터와 테스트 데이터는 10묶음 교차 검증법을 사용했다. 각 문장에 속한 SEED 태그의 속성 자질들 중 어떤 자질들이 전체 문장의 주관성 분류나 극성 분류에서

영향을 많이 주는지 가늠하기 위해 유력해 보이는 자질 쌍으로부터 시작하여 다른 자질들을 하나씩 더해 가장 좋은 성능을 보이는 자질 조합을 새로운 시작점으로 삼는 방식으로 실험을 진행했다. 그 중 가장 좋은 성능을 보인 실험 결과를 그래프로 보이면 다음과 같다.

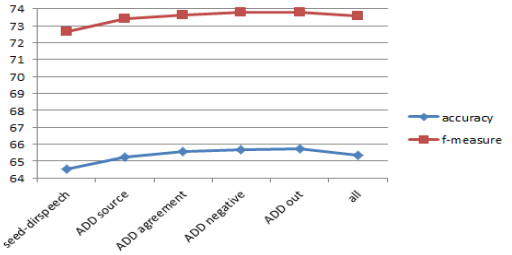

그림 1. 문장 주관성 SVM 분류 실험 결과

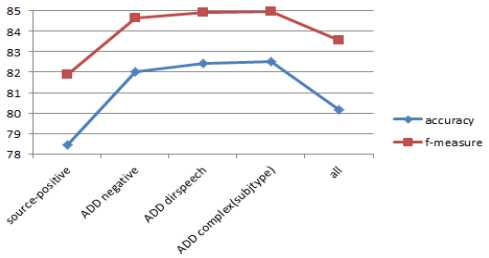

그림 2. 주관적 문장 극성 SVM 분류 실험 결과

주관성 분류 실험에서는 각 문장에 포함된 SEED 태그와 direct-speech (type), agreement (subjectivity-type), negative (polarity)의 수와 전체 nested-source와 out (nested-source) 수만을 분류 기준으로 사용했을 때 가장 좋은 성능을 보였다 (정확도 65.72%, 정밀도 59.76%, 재현율 96.41%, F-척도 73.78%). 특히 SEED 태그와 전체 nested-source 수의 사용은 주관적인 문장과 객관적인 문장이 포함하는 주관 표현의 수와 주관 표현 표출자 및 경로의 수 차이가 확연하다는 것을 보여준다. 이는 주관적인 문장이 일반적으로 많은 주관 표현을 포함하고 있으며, 객관적인 문장의 경우에 주관 표현이 상대적으로 많은 경로를 통해 표현된다는 점을 반영하는 것으로 보인다. 또한 nested-source 속성 중 사람 일반이나 출처가 없는 것을 나타내는 out 값이 중요한 기준으로 사용되었다는 점은 객관적 문장에 포함된 주관 표현의 경우에 "좋은 카메라는 가벼운 카메라다"의 '좋은'과 같이 구체적인 표출자가 없는 일반적인 주관 표현일 가능성이 높음을 시사한다. negative 극성값이 주관성 분류 기준으로 사용되었다는 것이 눈에 띄는데, 이는 뉴스 기사를 토대로 구축된 감정 코퍼스의 특성상 부정적인 주관성이 더 높은 비율로 포함되었을 가능성이 높고, 사설의 경우 부정적인 의견을 강도 높은 표현을 통해 전달하는 것이 대부분이라는 점에서 자연스럽게 이해된다.

문장의 극성 분류 실험에서 가장 좋은 성능을 (정확도 82.52%, 정밀도: 77.64%, 재현율 93.93%, F-척도 84.96%) 이끌어낸 5가지 분류 기준은 direct-speech (type), complex (subjectivity-type의 방향성 표지), positive (polarity), negative (polarity), 전체 nested-source 수였다. 우선 subjectivity-type 중에서도 emotion, judgment, argument 유형과만 결합하는 complex 방향성 표지가 극성 분류의 중요한 기준 중 하나라는 점에 주목할 수 있다. 주관 표현 유형 중 특정 대상에 대해 표출자의 기분을 표현하거나 대상을 평가하고, 어떤 상황에 대해 자신의 주장을 직접적으로 표현하는 주관 표현 유형이 긍정적인 문장과 부정적인 문장을 분류하는 기준 중 하나인 것과 다른 주장에 대해 동의하거나 반대하는 주관 표현인 agreement 유형의 경우에는 주관적인 문장과 객관적인 문장을 구분하는 효과적인 기준으로 사용되었다. 구축된 감정 코퍼스의 기초 자료가 뉴스 기사라는 점을 감안할 때, 제 3자의 특성 대상이나 상황에 대한 태도를 묘사하는 객관적 문장에서 agreement 유형이 자주 사용되고, 직접적인 의견을 표현하는 사설이나 기사 내부에 직접 인용된 감정 표현에서 emotion, judgment, argument처럼 좀 더 직접적인 주관 표현을 사용하는 경향이 있을 것이라고 정리할 수 있다.

5. 결론 및 향후 연구 과제

이 논문은 한국어 텍스트 자동 감정 분석 연구의 기초가 되는 감정 분석 코퍼스를 구축하기 위해 진행한 일련의 연구들을 소개하고 구축된 코퍼스에 대한 분석과 활용도에 대한 가능성을 제시하였다.

이렇게 구축된 코퍼스는 자동 감정 분석 연구를 위한 사전 구축, 기계 학습을 위한 학습 데이터로 사용될 수 있다. 이 자료를 바탕으로 감정분석, 의견분석을 위한 의의 있는 표현들이 학습될 수 있을 뿐만 아니라 한국어 감정분석 연구에 골드 스탠다드로 활용될 수 있어 그 의의는 크다. 앞으로 주석된 표현들의 일반화 및 일반 공개 형태의 결정 등을 통해 이 분야 연구에 널리 활용될 수 있는 자료가 될 수 있으리라 기대한다.

참 고 문 헌

[1] Wiebe, Janyce, Theresa Wilson, and Claire Cardie. *Annotating Expressions of Opinions and Emotions in Language*. Language Resources and Evaluation (formerly Computers and the Humanities), 39(2/3):164–210, 2005.

[2] Wiebe, Janyce. *Instructions for annotating opinions in newspaper articles*. Department of Computer Science Technical Report TR-02-101, University of Pittsburgh. 2002.

[3] Wilson, Theresa Ann. *Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D Dissertation, University of Pittsburgh. 2008.

[4] Shin, Hyopil, Munhyong Kim, Yu-Mi Jo, Hayeon Jang, and Andrew Cattle. Annotation Scheme for Constructing Sentiment Corpus in Korean. *In Proceedings of PACLIC 26*. 181-190. 2012.

[5] Cattle, Andrew, Munhyong Kim, and Hyopil Shin. Morpheme-based Annotation Tool for Korean Text. *In Proceedings of AAC-2013*. San Diego State University. 2013.