한국어 감정분석을 위한 말뭉치 구축 가이드라인 및 말뭉치 구축 도구

하은주○, 오진영, 차정원
창원대학교
fraga8604@changwon.ac.kr, psycheojy@changwon.ac.kr, jcha@changwon.ac.kr

Annotation Guidelines for Korean Sentiment Analysis and Annotation Tool

Eun-Ju Ha○, Jin-Young Oh, Jeong-Won Cha
Changwon University

요 약

한국어 감정분석에 대한 연구는 활발하게 진행되고 있다. 그렇지만 학습 및 평가 말뭉치 표현에 대한 논의가 부족하다. 본 논문은 한국어 감정분석에 대해 정의하고, 말뭉치 제작을 위한 가이드라인을 제시한다. 또한, 태깅 가이드라인에 따라 말뭉치를 구축하였으며 한국어 감정분석을 위한 반자동 태깅 도구를 구현하였다.

주제어: 한국어 감정분석, 의견마이닝, 감성분석, 감정분석 태깅 가이드라인

1. 서론

자연어처리(Natural Language Processing)의 한 분야인 감정분석(Sentiment Analysis)은 텍스트에 나타나 있는 어떤 대상에 대한 누군가의 의견이나 평가, 감정 따위를 자동으로 분석하는 것을 목표로 한다. 감정분석은 “Sentiment”라는 영어 단어를 한국어로 어떻게 번역하느냐에 따라서 감정분석이나 감성분석이라는 표현으로 사용되기도 한다. 그리고 의견(Opinion)을 추출한다는 의미에서 의견마이닝이라는 표현으로도 사용한다.

감정분석에서 분석하고자 하는 대상을 의견대상(Opinion Target)이라 한다. 상품이나 제품, 기업을 포함한 조직, 정치인이나 연예인 같은 인물 등 다양한 것들이 의견대상이 될 수 있다. 최근에는 소셜네트워크 서비스가 등장하면서 과거보다 훨씬 많은 사람이 자신의 의견을 더 쉽고, 자유롭게 표출할 수 있게 되었다. 그 결과, 의견대상에 대한 의견이 담겨 있는 의견 텍스트(Opinion Text)는 지금 이 시각에도 폭발적으로 증가하고 있다.

영어권에서는 [1]이 손으로 의견, 감정 태깅을 부착할 때 발생하는 문제점을 조사하였다. 이 연구는 신문 10,000 문장의 말뭉치를 제작한 첫 번째 시도이다. [2]는 문장 단위의 감정 태깅 말뭉치를 만들기 위해 22 Grimms’ tale에서 1,580 문장을 손으로 태깅을 달았다. [3]은 인도어 감정분석을 위한 말뭉치를 작성하였다.

하지만 감정분석을 위한 가이드라인의 부재로 인하여 학습과 평가를 위한 데이터 생성이 이루어지지 않고 있다. 이로 인하여 연구자마다 서로 다른 기준으로 연구결과를 보고하고 있는 실정이다[4-9].

본 논문에서는 한국어 감정분석 연구를 위한 태깅 말뭉치 구축 가이드라인을 제시하고 연구용 태깅 말뭉치를 제작한다.

구성은 다음과 같다. 2장에서는 한국어 감정분석 가이드라인을 제시하고, 3장에서는 태깅 가이드라인을 설명하며, 4장에서는 정의한 의견 말뭉치를 구축하는 과정에 관해서 설명한다. 마지막으로 5장에서 결론을 기술한다.

2. 한국어 감정분석 가이드라인

본 논문에서는 한국어 감정분석을 위한 가이드라인을 제안한다.

2.1 기본 원칙

감정분석에 사용되는 의견텍스트는 주로 리뷰 문서이다. 감정분석은 세 가지의 의견구성요소가 있다. 다음 의견문장(Opinion Sentence)를 살펴보자.

1) “나는 호텔의 서비스가 만족스러웠다.”

예문 1)을 분석한 의견구성요소는 표 1에 나타나 있다.

표 1. 의견 구성요소

단어	구성요소	
나	의견주체 (Opinion Holder)	
호텔	의견대상 (Opinion Target)	주제 (Topic)
서비스		부주제 (Subtopic)
만족스러웠다.	의견단어 (Opinion Word)	

- 의견주체(Opinion Holder)는 의견을 피력하는 주체로서 예문에서는 글쓴이 자신이 의견주체에 해당한다.
- 의견대상(Opinion Target)은 구분 없이 쓰이기도 하고 주제(Topic)와 부주제(Subtopic)로 나뉘기도 하는데 주제는 의견 표출의 대상이 되는 사람, 조직, 상품 등이고 부주제는 주제가 가지고 있는 특징이나 속성이다.

- 의견단어(Opinion Word)는 의견대상에 대한 의견을 나타내는 표현으로써 긍정 혹은 부정의 극성을 띄고 있다. 일반적으로 극성을 가진 형용사가 의견단어로 취급된다.
-

2.2 의견말뭉치의 메타 정보

그림1은 본 논문에서 구축한 말뭉치의 예제이다.

```
#DOC_ID:73482019
#AUTHOR:tripadvisor member
#LOCATION:
#AUTHOR_TYPE:
#HELP_COUNT:0
#LANGUAGE:ko
#HOTEL:프레그런스 호텔 오아시스
#TITLE:합리적인가격과 친절한 직원
#RATINGS:4|가격4|장소  3|객실  3|청결도4|서비스5
#CONTENT
<#:S_OBJ>      공항에서 택시를 타고 들어갔는데 규모가 큰 호텔이 아니다 보니 정확한 주소를 보여줘야 합니다

<#:S_OBJ>      근처에 비슷한 호텔들이 많아서 그런거 같아요

<#:S_POS>      <방:OT>은 <작지만:OW_NEG> <깨끗했습니다:OW_POS>
<REL>has_opinion;방;작지만;1;2</REL>
<REL>has_opinion;방;깨끗했습니다;1;3</REL>
#END_OF_REVIEW
```

그림1. 말뭉치 예제

문서의 형식은 간단히 텍스트 처리를 통해 메타정보를 얻을 수 있도록 설계하였다. 상단에 '#' 기호로 시작되는 것들이 메타정보이며, 트립어드바이저 사이트에서 자체적으로 제공하는 것이다. 메타정보에 대한 설명을 표2로 정리하였다.

표2. 의견말뭉치의 메타 정보

메타태그	설명
DOC_ID	리뷰 문서 아이디
AUTHOR	리뷰 작성자
LOCATION	리뷰 작성자의 국적
AUTHOR_TYPE	리뷰 작성자의 유형
HELP_COUNT	리뷰 추천 수
LANGUAGE	리뷰 작성 언어
HOTEL	호텔명
TITLE	리뷰 제목
RATINGS	호텔 평점
CONTENT	리뷰 본문
END_OF_REVIEW	리뷰의 마지막 줄

2.3 말뭉치 구축 태그셋

감정분석은 문서에 나타난 글쓴이의 의견이나 감정이 긍정적인지, 부정적인지 분류하는 것을 말한다. 본 논문에서는 문장을 극성에 따라 크게 긍정(POS), 중립(NEU), 부정(NEG), 의견이 담겨 있지 않은 객관 문장(OBJ)으로 구분한다. 표3에 문장 태그를 정리하였다.

표3. 문장 태그

태그	정의
S_POS	긍정적 의견을 담고 있는 문장
S_NEU	중립적 의견을 담고 있는 문장
S_NEG	부정적 의견을 담고 있는 문장
S_OBJ	의견이 담겨 있지 않은 문장

개체형 의견정보는 문장에 나타난 명사나 명사구를 대상으로 도메인과 관련성에 따라 의견대상을 태깅하며 의견단어는 어절 단위로 묶어서 극성(긍정/부정/중립) 정보를 함께 태깅한다. 표4에 개체형 의견정보를 정리하였다.

표4. 개체형 의견정보

구분	태그	정의
의견대상	OT	도메인과 관련이 있는 의견대상
	OT_EXT	도메인과 관련 없는 의견대상
의견단어	OW_POS	긍정적 의견단어 및 표현
	OW_NEU	극성을 명확히 알 수 없는 의견단어 및 표현
	OW_NEG	부정적 의견단어 및 표현

관계형 의견정보는 의견대상 간 또는 의견대상과 의견단어와의 관계 형태의 정보를 의미한다. 정관계는 문서의 기준이 되는 것이 먼저 기술된 경우이고 역관계는 기준이 되는 것이 나중에 기술된 경우이다. 표5에 관계형 의견정보를 정리하였다.

표5. 관계형 의견정보

구분	관계명	대상
의견대상과 의견단어 간의 관계	has_opinion	의견대상을 기준으로 의견단어와 정관계
	is_opinion_of	의견대상을 기준으로 의견단어와의 역관계
의견대상 간의 관계	equal	동일한 의견대상을 다른 명칭으로 사용
	has_aspect	주제를 기준으로 부주제와 정관계
	is_aspect_of	주제를 기준으로 부주제와 역관계

3. 태깅 가이드라인

3.1 문장 태깅

문장의 시작 부분에 위치하며 문장의 극성에 따라 <#;문장태그명>을 표기해준다.

- S_POS(긍정적 문장): 문장에 나타난 글쓴이의 의견이 긍정적으로 나타나는 경우 S_POS로 분석함.

  예) **<#:S_POS>**나는 호텔의 서비스가 **만족스러웠다.**

- S_NEU(중립적 문장)

  - 중립적 의견을 담고 있는 경우 S_NEU로 분석함.

  예) **<#:S_NEU>**나는 호텔 서비스는 **나쁘지 않았다.**

  - 한 문장에 긍정과 부정이 함께 나타나는 경우도 S_NEU로 분석함.

  예) **<#:S_NEU>**호텔 서비스는 **좋지도 나쁘지 않았다.**

- S_NEG(부정적 문장): 문장에 나타난 글쓴이의 의견이 부정적으로 나타나는 경우 S_NEG로 분석함.

  예) **<#:S_NEG>**나는 호텔의 서비스가 **맘에 들지 않았다.**

- S_OBJ(객관 문장): 문장에 글쓴이의 의견이 나타나지 않는 객관적인 내용을 담고 있는 경우 S_OBJ로 분석함.

  예) **<#:S_OBJ>**호텔에 수영장이 있다.

3.2 개체형 의견정보

개체형 의견정보는 개체명 인식(Named Entity Recognition) 말뭉치에서 개체명에 해당하는 것이다. 의견대상은 단어 단위, 의견단어는 어절 단위로 태깅한다.

- 의견대상

  (1) 의견대상은 도메인과의 관련성 여부에 따라 구분.

  (2) OT는 도메인과 관련 있는 의견대상.

  (3) OT_EXT는 도메인과 관련 없는 의견대상.

  예) 한 번 더 **<한국:OT_EXT>**에 가고 싶고, 다음에 가게 되면 이 **<호텔:OT>**에 묵고 싶습니다.

- OW_POS(긍정적 의견단어)

  (1) 긍정적인 의견단어나 표현이 담겨 있는 경우.

  예) 호텔의 서비스가 **<좋다:OW_POS>**.

- OW_NEU(중립적 의견단어)

  (1) 극성을 명확히 알 수 없는 의견단어 및 표현.

  예) 호텔의 서비스가 **<좋지도 싫지도 않다:OW_NEU>**

- OW_NEG(부정적 의견단어)

  (1) 부정적인 의견단어나 표현이 담겨 있는 경우.

  예) 호텔의 서비스가 **<싫다:OW_NEG>**

3.3 관계형 의견정보

관계형 의견정보는 의견대상과 의견단어, 의견대상과 의견대상 간의 관계 형태 두 가지로 분류된다. <REL>로 시작하여 </REL>로 닫는다. 의견대상과 의견단어의 위치정보는 1부터 시작하여 표기한다.

> <#:문장태깅명> 문장태깅
> <REL>관계명;의견대상/의견단어;위치정보;</REL>

- has_opinion

  - 서로 관련 있는 의견대상과 의견단어

  - 문장에 의견대상이 먼저 기술되어 있는 정관계의 경우 태깅

  예) <서비스:OT>가 <좋았다:OW_POS>.

  <REL>**has_opinion**;서비스;좋았다;1;2</REL>

- is_opinion_of

  - 서로 관련 있는 의견대상과 의견단어

  - 문장에 의견대상이 나중에 기술되어 있는 역관계의 경우 태깅

  예) <좋은:OW_POS> <서비스:OT>였다.

  <REL>**is_opinion_of**;좋은;서비스;1;2</REL>

- equal

  - 두 의견대상이 동일하지만 서로 다른 명칭으로 사용된 경우

  예) <호텔:OT>은 산책하기에 <좋은:OW_POS> <거점:OT>이었다.

  <REL>**equal**;호텔;거점;1;3</REL>

- has_aspect

  - 의견대상과 의견대상 간의 관계 형태 정보

  - 주제를 기준으로 부주제와 정관계

  예) <호텔:OT>의 <서비스:OT>가 좋았다.

  <REL>**has_aspect**;호텔;서비스;1;2</REL>

- is_aspect_of

  - 의견대상과 의견대상 간의 관계 형태 정보

  - 주제를 기준으로 부주제와 역관계

  예) 최고의 <서비스:OT>를 제공하는 <호텔:OT>

  <REL>**is_aspect_of**;서비스;호텔;1;2</REL>

4. 의견 말뭉치 구축

본 논문에서 사용한 말뭉치는 트립어드바이저[10]에서 제공하는 호텔 리뷰 문서로 전 세계의 호텔에 대한 다국어 리뷰를 제공하고 있다. 본 논문에서는 이 사이트에서

인기 있는 15개 도시를 대상으로 도시별로 최대 150개의 호텔 리뷰 전체를 수집하였다. 수집된 리뷰 중에서 한국어 리뷰는 총 6,853개이다. 그중에서 2,000개의 리뷰를 대상으로 의견대상을 포함한 몇몇 정서 정보를 부착하여 태깅된 말뭉치(Labeled Corpus)를 구축하였다. 본 논문에서 사용한 실험 말뭉치의 통계량은 표6과 같다.

표6. 한국어 감정분석을 위한 트립어드바이저 말뭉치

구분	태깅 안 된 말뭉치	태깅된 말뭉치	합계
리뷰	4,853	2,000	6,853
문장	41,934	14,495	56,429
어절	372,478	125,369	497,847

본 말뭉치는 딥러닝 학습을 위한 충분한 양의 말뭉치는 아니지만, 한국어 감정분석 연구를 위한 기초 자료로서 역할은 할 수 있을 것으로 기대한다.

한국어 감정분석을 위한 태깅을 위하여 반자동 태깅 툴인 Kasen을 개발하였다. 그림2의 화면에서 편집하며, 각각의 기능은 아래와 같다.

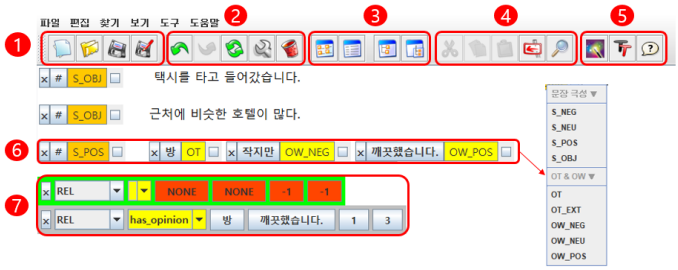

그림2. 한국어 감정분석 반자동 태깅 도구 : Kasen

① 신규/불러오기/저장

② 되돌리기/다시 실행/다시 출력/재배열/패널 제거

③ 보기 모드 선택/관계 태깅 추가

④ 문서 편집/삭제

⑤ 마술봉[1)]/글꼴 변경/툴 정보

⑥ 문장태그/개체형 의견정보

⑦ 관계형 의견정보

문장태그와 개체형 의견정보 태그 후 위치 정보만 배치해주면 관계형 의견정보는 자동생성된다. Kasen은 감정분석을 위한 말뭉치 구축을 보다 효율적으로 할 수 있는 기능을 갖췄다.

5. 결론

본 논문에서는 한국어 감정분석을 위한 말뭉치 구축 가이드라인을 제시하였다. 이를 이용하여 딥러닝에 사용할 수 있는 대량의 말뭉치를 구축 중이다. 향후 이 가이드라인을 보완하고 이를 표준화할 예정이다. 또한, GAN 기술을 활용하여 대량의 말뭉치를 고속 구축할 예정이다.

1) 사전을 이용한 자동 초기 태깅 기능

사사

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1D1A1B03033534).

참고문헌

[1] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. Language resources and evaluation 39(2):165-210.

[2] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, pages 579-586.

[3] Piyush Arora. 2013. Sentiment analysis for hindi language. MS by Research in Computer Science, IIIT Hyderabad .

[4] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, vol. 37, no. 1, pp. 9-27, 2011.

[5] Ana-Maria Popescu and Oren Etzioni, "Extracting Product Features and Opinions from Reviews," In Proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing, pp. 339-346, 2005.

[6] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto, "Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining," In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1065-1074, 2007.

[7] Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu, "Phrase Dependency Parsing for Opinion Mining," In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1533-1541, 2009.

[8] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O' Brien-Strain, "Extracting and Ranking Product Features in Opinion Documents," In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1462-1470, 2010.

[9] 배원식, "연쇄반응 알고리즘을 이용한 의견대상 추출," 창원대학교 컴퓨터공학과 박사학위 논문, 2013.

[10] http://www.tripadvisor.co.kr, 2018.9.4.(참조)