Latent Structural SVM을 활용한 감성 분석기

양승원O 이창기
강원대학교 컴퓨터과학과
swyang@kangwon.ac.kr, leeck@kangwon.ac.kr

Sentiment Analysis using Latent Structural SVM

Seung-Won YangO Changki Lee
Kangwon National University Dept. of Computer Science

요 약

본 연구에서는 댓글(음식점/영화/모바일제품) 및 도메인이 없는 트위터 데이터에 대한 감성 분석을 수행하고, 각 문장에 대한 object(or aspect)와 opinion word를 추출하는 시스템을 개발하고 평가한다. 감성 분석을 수행하기 위해 Structural SVM 알고리즘과 Latent Structural SVM 알고리즘을 사용하여 비교 평가하였으며, 실험 결과 Latent Structural SVM이 더 좋은 성능을 보였으며, 구문 분석을 통해 분석된 VP, NP정보를 활용하여 object(aspect)와 opinion word를 추출할 수 있음을 보였다.

1. 서 론

어떤 주제의 긍정적이거나 부정적인 의견 표출에 대한 요약된 정보를 제시해 주거나 어떤 주제에 대한 보다 상세한 항목에 대한 평가를 요약해서 제시해 주는 분석의 응용은 '감성 분석(sentiment analysis)', '의견 분석(opinion mining)', '감정 분석(emotion analysis)' 등으로 불린다[1].

특정 상품에 대한 댓글 혹은 특정 주제에 대한 트위터 데이터들은 개인의 의견을 포함하고 있는 감성 데이터로 볼 수 있으며, 감성 데이터 분석을 통해 마케팅에 활용할 수 있는 좋은 지표가 될 수 있다. 그러나, 날로 증가하는 감성 데이터를 일일이 읽어 의미를 파악하는 것은 실제 불가능한 일이다. 따라서, 대용량 감성 데이터를 자동으로 분석/활용할 수 있는 방법이 필수적이다.

감성 분석은 통상 3단계로 이뤄 진다. 첫 번째는 각종 소셜 미디어 매체에서 정보를 수집하는 '데이터 수집(data collection)' 단계이다. 두 번째는 이렇게 총체적으로 수집된 정보에서 사용자의 주관이 드러난 부분만을 걸러 내는 '주관성 탐지(subjectivity detection)' 과정이다. 마지막 세 번째 단계에서 '극성 탐지(polarity detection)' 작업이 이뤄지는데, 이는 추출된 감성 데이터를 '좋음'과 '싫음'의 양극단으로 분류하는 과정이다. 본 연구에서는 데이터 수집에 대한 단계는 다루지 않는다. 이러한 감성 데이터는 다음과 같은 구성 요소를 가지고 있다고 정의할 수 있다.

Basic components of an opinion

Opinion holder: The person or organization that holds a specific opinion on a particular object.

Object (or aspect): on which an opinion is expressed

Opinion: a view, attitude, or appraisal on an object from an opinion holder.

아래의 그림은 간단한 감성 분석의 예시를 보여 주고 있다.

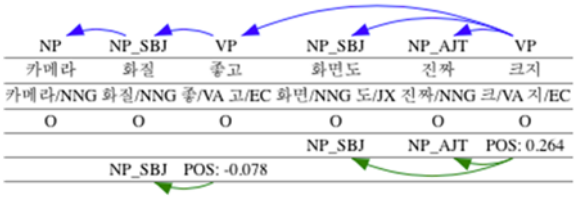

그림 1 감성 분석 예시

감성 데이터 구성 요소 관점에서 봤을 때, 그림 1과 같이 감성 분석 결과를 얻었을 경우 해당 글의 작성자를 opinion holder라고 할 수 있으며, "카메라 화질"은 object(aspect), "좋은(좋다)"를 opinion word로 정의할 수 있다. 또한, 동일한 방식으로 "화면"은 object, "크지(크다)"를 opinion word로 정의할 수 있다. 예와 같이 하나의 문장은 단문으로만 이루어지지 않고, 중문 혹은 복문으로 구성될 수 있으며, 단문이 아닐 경우 각각의 object와 opinion word를 따로 추출해야 분석의 결과에 혼란을 제거할 수 있다.

본 연구에서는 댓글(음식점/영화/모바일) 데이터와 도메인이 없는 트위터 데이터에 대한 감성 분석을 수행하고, 각 문장에 대한 object와 opinion word을 추출하는 시스템을 개발하고 평가한다. Opinion holder는 댓글의 경우 댓글을 작성한 작성자로, 트위터의 경우 트윗을 작성한 작성자로 정의하였으며 본 연구에서 해당 주제는 다루지 않는다. 감성 분석을 수행하기 위해 Structural SVM 알고리즘과 Latent Structural SVM 알고리즘을 사용하여 비교 평가하였다.

2. 관련 연구

감성 분석 작업은 그 대상을 기준으로 다양한 접근 방법을 제시하고 있다. 우선, 문서(혹은 특정 리뷰) 단위의 감성 분석 작업의 경우 모든 문서는 단일한 주제(object, aspect)에 대해서 작성될 것을 가정하고, 문서들의 분류 문제로 정의한 후 각 문서 단위로 긍정/부정/중립 중 하나의 클래스로 매칭하도록 한다. 그러나, 실제 단일한 주제를 다루고 있지

않다는 문제점을 가지고 있다. 두 번째로, 문장(sentence) 단위의 감성 분석 작업이 있을 수 있는데, 각 문장들이 주관(subjectivity)을 가지고 있는지 판별(Subjectivity Detection)하는 문제로 접근하고, 주관을 가지고 있는 문장이라고 판단될 경우에 감성 분류 문제로 본다. 이는, 하나의 문장은 하나의 주관(의견)을 가지고 있다고 가정하고 있지만, 실제 단일한 주관을 가지고 있지 않은 문제가 있다. 따라서, 문장 단위가 아니라, 구(phrases)나 절(clauses) 단위로 접근해야 할 필요성이 있다.

[5]에서는 의존 문법(dependency grammer)과 DP(double propagation)을 이용하여 aspects와 opinion words를 추출하도록 접근하였으며, [6]에서는 WTM(Word-Based Translation Model)을 바탕으로 opinion target과 opinion word간의 그래프 문제로 접근하였다.

본 논문에서는 감성 분류를 위해서 Latent Structural SVM 모델을 사용하였으며, 구문 분석 결과를 활용하여 object(aspect)와 opinion word를 추출하였다.

3. Latent Structural SVM을 이용한 감성 분류기

본 연구에서는 감성 분석 문제를 해결하기 위해서 Latent Structural SVM을 활용하였다[2][7]. 3.1에서 Latent Structual SVM에 대해서 간단하게 살펴본 후, 3.2에서 실제 감성 분석기 학습 및 분석 절차에 대해서 알아본다.

3.1 Latent Structural SVM

Latent Structural SVM은 기존의 Structural SVM에 은닉 변수(hidden variable) $\mathbf{h}$ 를 추가한 것으로, 다음과 같이 정의된다[2][7].

$$\min_{\mathbf{w}} \frac{1}{2} \| \mathbf{w} \|^2 + \frac{C}{n} \sum_{i}^{n} \left(\max_{(\mathbf{y},\mathbf{h}) \in \mathbf{Y} \times \mathbf{H}} \left(\Delta(\mathbf{y}_i, \mathbf{y}, \mathbf{h}) + \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \right) \right)$$

$$- \frac{C}{n} \sum_{i}^{n} \left(\max_{\mathbf{h} \in \mathbf{H}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \right) \quad (1)$$

위 식에서 $\Delta(\mathbf{y}_i, \mathbf{y}, \mathbf{h})$는 정답 태그 열 $\mathbf{y}_i$와 예측 결과 태그 열 $\mathbf{y}$와 은닉 변수 $\mathbf{h}$를 입력으로 받는 loss 함수로 일반적으로 은닉 변수 $\mathbf{h}$ 는 무시하여 $\Delta(\mathbf{y}_i, \mathbf{y})$ 와 같아 지고(즉, $\mathbf{y}_i$ 와 $\mathbf{y}$사이의 다른 태그 개수), $\Phi(\mathbf{x}_i, \mathbf{y}, \mathbf{h})$는 은닉 변수 $\mathbf{h}$가 추가된 자질(feature) 벡터 함수를 나타낸다. 은닉 변수 $\mathbf{h}$를 무엇으로 정의하냐에 따라 자질 벡터 함수 $\Phi(\mathbf{x}_i, \mathbf{y}, \mathbf{h})$가 달라지게 되며 이를 사용자가 정의해 주어야 한다.

Latent Structural SVM에 감성 분석 문제를 적용하기 위해서 $\mathbf{x}$=문장, $\mathbf{y}$=감성 분류명(긍정/부정), $\mathbf{h}$=트리거 후보(동사 혹은 문장의 맨 뒤 명사구)로 정의했다.

Structural SVM

$\mathbf{x}$=문장, $\mathbf{y}$=감성 분류명(긍정/부정),

Latent Structural SVM

$\mathbf{x}$ =문장, $\mathbf{y}$ =감성 분류명(긍정/부정), $\mathbf{h}$ =트리거 후보(명사(주어/목적어 등의 aspect) + 용언(동사/형용사 등의 opinion word))

3.2 감성 분석기 학습 및 분석 절차

간단한 감성 분석기의 학습 및 분석 절차는 아래와 같다. 문서 수집은 특정 도메인에 대한 댓글 수집(음식점/영화/모바일제품) 및 트위터 데이터에 대한 수집을 수작업으로 진행하였다.

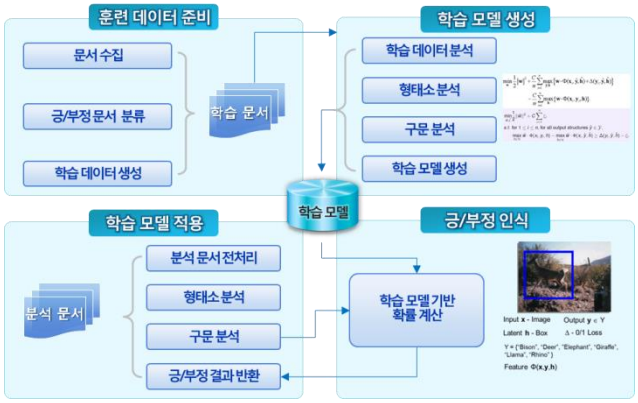

그림 2 감성 분석 절차

입력된 문장에 대해서 형태소 분석/구문 분석을 수행, 해당 자질을 사용하도록 하였으며, 실제 감성 분석에서는 Latent Structural SVM 알고리즘을 사용하여 분석을 수행하였다. 감성 분석에 사용된 자질은 3.2.2에서 자세하게 설명하도록 한다. 학습 및 테스트에 사용된 데이터는 아래와 같다.

표 1 학습 데이터

Data Set	Train Data	Test Data
영화	300	100
음식점	300	100
모바일	4,543	500
트위터	16,127	1,793

3.2.1 전처리 단계

정제(문장 분리, 아이디/RT/URL 제거 등) 된 데이터를 사용하여 각각 형태소 분석, 구문 분석을 수행하여 자질을 생성한다. 본 연구에서 사용한 형태소 분석 및 구문 분석은 Structural SVM 모델을 사용한 분석기를 사용하였다[3].

3.2.2 감성 분석을 위한 자질

감성 분석의 자질은 다음과 같은 자질을 사용하였다.

표 2 감성 분석 자질 예

카메라 화질 좋고 화면도 진짜 크지
Structural SVM Feature
POS 카메라/NNG 화질/NNG 좋/VA 화면/NNG 진짜/NNG 크/VA
Latent Structural SVM Feature
$\Phi(x, y, h_1)$: pred=좋 pred=좋고 pred=VP pred=좋_VP pred_dic pred_dic=1 hd_pred=크 hd_pred=좋_크 arg=화질 pred_arg=좋_화질 pred_arg=좋_화질 pred_arg2=좋_카메라_화질 카메라/NNG 화질/NNG 좋/VA 화면/NNG 진짜/NNG 크/VA
$\Phi(x, y, h_2)$: pred=크 pred=크지 pred=VP pred=크_VP pred_dic pred_dic=0 arg=화면 pred_arg=크_화면 pred_arg=크_화면 arg=진짜 pred_arg=크_진짜 pred_arg=크_진짜 카메라/NNG 화질/NNG 좋/VA 화면/NNG 진짜/NNG 크/VA

Structural SVM과 Latent Structural SVM 모두 형태소 분석 결과의 어휘를 사용하였으며, 본 연구에서는 체언, 용언,

수식언에 대한 어휘만을 추출하여 그 특성으로 활용하였다. Latent Structural SVM의 경우 추가적으로 구문 분석의 결과를 사용하였으며, 결과 중 NP/VP/VCP를 HEAD로 간주하여, 각 HEAD단위로 주어/목적어/수식어 등을 포함하여 hidden variables로 정의하였다. 또한, 간단한 감성 사전을 구축하여 그 특성을 추가하였다.

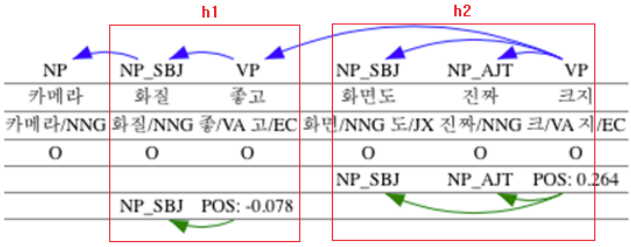

그림 3 hidden variable

3.2.3 감성 분석 결과 도출

문장 분석은 학습과 동일한 방식으로 전처리 과정을 수행하고, 형태소 분석/구문 분석의 과정을 거쳐 자질을 추출하게 된다. 이 과정에서 구문 분석의 결과 중 HEAD(즉, hidden variable)로 간주되는 영역별로 분석을 수행하여 긍/부정 여부를 판단하였으며, 추출된 HEAD 정보를 바탕으로 object와 opinion word를 생성하였다.

즉, *"카메라 화질 좋고 화면도 진짜 크지"*라는 문장을 분석하면 2개의 HEAD 정보를 얻을 수 있으며(*"좋고"*, *"크지"*) 각각의 긍/부정 결과 (POS/POS)를 얻을 수 있다. 또한, HEAD 정보와 구문 분석 결과를 활용하여, 다음의 결과를 도출할 수 있다.

opinion1 = {Polarity:POS Object:카메라 화질 Opinion:좋다}

opinion2 = {Polarity:POS Object:화면 Opinion:크다}

opinion word의 경우 원형 복원 단계를 거쳐 정규화하는 작업을 수행하였으며, 간단한 규칙을 생성하여 표현력을 강화하였다.

예를 들어 *"사과하기로 했습니다"*는 *"사과하다"*로, *"통과되지 않아"*는 *"통과되지 않다"* 등으로 원형 복원을 수행하고, predicate에 대한 표현력을 강화하도록 하였다. aspect(or object)는 NP_SBJ/NP_OBJ를 기본으로 구문 분석 결과를 활용하여 구 묶음(청크)을 생성하였다.

표 3 Opinion 생성 규칙(일부)

규칙	예시
MAG+VA	더 귀엽다
MAG+VA+VX	안 예뻐지다
NNG+VCP/VCN	병신 이다
NNG+VV	짜증나다
NNG+XSV	사랑하다

4. 실험 및 결과

표 4 성능 평가

data	Accuracy		NEG (F1)	POS (F1)	aspect
	SSVM	LSSVM	LSSVM	LSSVM	매칭률
영화	76.00	**79.00**	78.79	79.21	63.9
음식점	82.00	**90.00**	90.57	89.36	67.3
모바일	85.58	**90.45**	87.96	92.08	75.0
트위터	84.66	**84.21**	80.46	83.45	-

각 학습 데이터를 도메인 단위로 구분하여 성능을 측정하였으며, 상대적으로 학습 데이터 수가 부족한 도메인에서 점수가 낮게 나왔다.

aspect(or object)의 성능 테스트는 각 테스트 데이터에 대한 정답을 설정하고, 매칭률를 계산하였다. 설정한 값과 동일하거나 값을 포함할 경우 정답으로 하였으며, 추출하지 못할 경우 오답으로 계산하였다. 모바일의 경우 "반응속도, 그립감" 등과 같이 명확히 목표를 명시하고 감정을 표현하는 방식이 많아 비교적 높은 매칭률을 보였다.

Structural SVM의 경우 특성을 추출하는 과정 및 분석 시간이 비교적 빠르지만, 성능 측면에서 전체적으로 Latent Structural SVM 보다 낮았으며, Latent Structural SVM과 달리 aspect와 opinion후보를 찾아 주지 못했다.

5. 결론

본 연구에서는 댓글(음식점/영화/모바일) 데이터와 도메인이 없는 트위터 데이터에 대한 감성 분석을 수행하고, 각 문장에 대한 object와 opinion word를 추출하는 시스템을 개발하고 평가했다. 실험은 Structural SVM과 Latent Structural SVM 모델을 사용하여 테스트하였으며, 테스트 결과 Latent Structural SVM 모델을 사용할 경우 성능이 우수했으며, 입력 문장에 대한 극성 판단만 하지 않고, 극성에 영향을 준 object와 opinion word를 추출 할 수 있다.

연구 결과, 댓글 및 트위터 데이터의 특성(띄어쓰기 오류, 맞춤법 오류 등)에 의해서 전처리 단계인 언어 분석 결과가 좋지 않아서 성능 하락이 발생한 것으로 보인다. 향후 Deep Learning 알고리즘을 적용하여 해당 문제에 접근할 수 있을 것을 보인다. 특히 RNN이나 CNN의 경우 단어의 특성뿐만 아니라, 구 및 문장에 대한 특성을 결정할 수 있어, 감성 분석에 좋은 성능을 보일 것으로 예상한다.

References

[1] 조은경, 감성 분석 연구의 현황과 말뭉치에 기반한 사례 분석, 언어과학연구 61, 2012.

[2] 이창기, 잠재 구조적 SVM을 확장한 결합 학습 모델, 정보과학회논문지: 소프트웨어 및 응용 41권 6호, 2014.6

[3] 이창기, Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델, 정보과학회논문지: 소프트웨어 및 응용 제 40권 12호, 2013.12

[4]Khairullah Khan, Mining opinion components from unstructured review, Journal of King Saud University, 2014

[5] Guang Qiu, Opinion Word Expansion and Target Extraction through Double Propagation, Computational Linguistics Volume 37, Number 1, 2011

[6] Kang Liu, Opinion Target Extraction Using Word-Based Translation Model, EMNLP-CoNLL '12, 2012

[7] Chun-Nam John Yu and Thorsten Joachims, Learning Structural SVMs with Latent Variables, In Proceedings of the ICML, 2009