NPMI를 이용한 어휘의 감성분석 연구

류기곤, 김현철
고려대학교 컴퓨터학과
e-mail:gon0121@korea.ac.kr

A Study on Sentiment Analysis of Words using Normalized PMI

Ki-Gon Lyu, Hyeon-Cheol Kim
Dept of Computer Science and Engineering, Korea University

요 약

감성분석은 최근 오피니언 마이닝에서 주목받고 있는 분야로써, 특정 주제, 상품, 유명인사 등에 대한 사람들의 반응을 긍정 또는 부정으로 구분하거나 점수를 이용하여 긍정 또는 부정의 강도를 분석하는데 이용되고 있다. PMI(pointwise mutual information)와 SO-PMI(semantic orientation from pointwise mutual information)는 비교적 빠르고 간편하게 극성을 판단할 수 있다는 장점이 있지만, 어휘와 기준 어휘 사이의 극성 값이 넓은 범위를 갖는다는 단점이 있다. 본 논문에서는 일상적인 언어 사용 환경에서 나타나는 어휘로부터 감성을 분석하고자 하였다. 특히 어휘의 극성 값 편차로 인해 나타날 수 있는 어려움을 보완하기 위해 NPMI(normalized pointwise mutual information)를 이용하여 어휘의 감성을 분석하였다. PMI와 NPMI를 비교 분석한 결과 어휘의 감성 강도를 나타내는 데 있어서 밀집도에서 큰 차이를 보였다.

1. 서론

감성분석(sentiment analysis)은 최근 오피니언 마이닝(opinion mining)에서 가장 중요하게 다뤄지고 있는 분야로써, 특정 주제, 상품, 유명인사 등에 대한 사람들의 반응을 긍정 또는 부정으로 구분하거나 점수를 이용하여 긍정 또는 부정의 강도(strength)를 분석하는데 이용되고 있다. 감성의 강도는 어떤 어휘가 특정 감성에서 얼마큼의 값을 갖는지를 수치로 표현할 수 있기 때문에 단순하게 긍정, 부정에 대한 분석 뿐 아니라 아주 긍정 또는 100점 만점 중 80정도의 부정을 표현할 수 있다. 이러한 감성의 강도를 분석하기 위해 대표적으로 어휘의 극성(word polarity)을 이용하는 연구가 있다.

PMI(pointwise mutual information)는 어휘의 극성을 판단하는데 일반적으로 사용하는 방법이다. 두 이산확률변수(discrete random variable)의 연관성을 표현하며, 극성이 높을수록 동일 문서에 출현할 확률이 높다고 가정할 수 있다. 따라서 어떤 어휘와 특정 감성으로 분류되는 기준 어휘의 PMI가 양수일수록 어휘 극성이 높기 때문에, 기준 어휘의 감성과 연관성이 높다고 판단할 수 있다. PMI를 통해 특정 감성으로 분류되는 기준 어휘 집합과 어휘 극성을 분석한 후 SO-PMI(semantic orientation from pointwise mutual information)를 이용하여 어휘의 감성을 판단할 수 있다. SO-PMI는 어떤 어휘와 특정 감성으로 분류되는 기준 어휘 집합과의 어휘 극성을 이용하여 감성을 결정한다. 감성을 판단하고자 하는 각각의 어휘에 대해 긍정 감성 기준 어휘 집합과 부정 감성 기준 어휘 집합과의 PMI 합의 차이를 통해 양수일수록 긍정, 음수일수록 부정 감성이라고 판단할 수 있다.

PMI와 SO-PMI는 비교적 빠르고 간편하게 극성을 판단할 수 있다는 장점이 있지만, 어휘와 기준 어휘 사이의 극성 값이 넓은 범위를 갖는다는 단점이 있다. PMI 결과는 최소 음수 무한대부터 최대 어휘나 기준 어휘의 출현확률 중 낮은 범위 내의 값으로 표현되고, 출현 확률에 따라 편차가 매우 크게 나타날 수 있다. 편차가 크면 클수록 SO-PMI를 이용한 감성 분석에 영향을 주게 되어 감성을 잘못 판단할 수 있고, 감성과 어휘들 간의 극성 차이를 세밀하게 분석하기 어려울 수 있다. 하지만, 어휘 극성 값의 넓은 범위는 정규화를 통해 동일한 범위 내의 값으로 변환할 수 있다.

본 논문에서는 일상적인 언어 사용 환경에서 나타나는 어휘로부터 감성을 분석하고자 하였다. 감성 분석을 위해 감성 별로 균등하게 어휘가 포함되어 있는 Greg Siegle의 BAWL(The ORIGINAL Balanced Affective Word List)를 기준 어휘로 사용하였다. 기준 어휘를 검색어로 이용하여 플리커로부터 대량의 어휘 자원을 수집하였고, 특히 어휘의 극성 값 편차로 인해 나타날 수 있는 어려움을 보완하기 위해, Gerlof Bouma의 NPMI(normalized pointwise mutual information)를 이용하여 어휘의 감성을 분석하였다.

2. 본론

감성 분석을 위해 사진공유 사이트인 플리커에서 대량의 소셜미디어를 수집하였다. 플리커에서 공유되는 사진을 감성을 수집하기 위한 공통된 외부자극, 사진에 달린 댓글을 집단지성을 통해 표현된 감성 어휘로 가정하였다. 사진을 수집하기 위한 검색어는 BAWL의 긍정과 부정 감성으로 분류된 어휘 전체를 사용하였고, SO-NPMI를 이용하여 감성을 분석하였다.

어휘의 극성을 분석하기 위해서는 어휘의 출현 확률을 구해야 한다. 하지만, 어휘의 출현 확률을 알 수 없기 때문에, Web-PMI의 어휘의 출현빈도를 이용한 출현 확률 추정 방법을 이용하였다. [식 1]처럼 어휘 x의 출현 확률 $p(x)$는 전체 문헌 N 중 x가 나타난 문헌의 수 $hits(x)$로 구할 수 있다.

$$p(x) \simeq \frac{hits(x)}{N} \qquad [식\ 1]$$

어휘와 기준 어휘와의 극성을 분석하기 위해, [식 1]에서 계산된 출현 확률과 PMI를 이용하였다. [식 2]처럼 어휘 x와 기준 어휘 y의 어휘 극성 $pmi(x;y)$는 어휘 x의 출현 확률 $p(x)$와 기준 어휘 y의 출현 확률 $p(y)$ 중 어휘 x와 기준 어휘 y가 동시에 나타난 출현확률 $p(x,y)$에 로그를 취함으로써 구할 수 있다.

$$pmi(x;y) = log\ \frac{p(x,y)}{p(x)p(y)} \qquad [식\ 2]$$

어휘 x와 기준 어휘 y의 어휘 극성 $pmi(x;y)$는 최소 $-\infty$부터 최대 $-log\,p(x)$와 $-log\,p(y)$ 중 작은 값을 갖기 때문에, 편차가 매우 크게 나타날 수 있다. [식 1]처럼 어휘의 빈도를 이용하여 출현 확률을 계산하면, 수집된 어휘 자원에 의해 출현 확률이 달라질 수 있고, 더 나아가서는 어휘극성 분석에 영향을 준다. 따라서 어휘의 빈도에 의한 어휘극성의 편차를 보완할 수 있도록, NPMI를 이용하여 정규화 하였다.

어휘 x와 기준 어휘 y의 정규화 된 어휘 극성 $npmi(x;y)$는 [식 3]처럼 어휘 x와 기준 어휘 y의 어휘 극성 $pmi(x;y)$를 어휘 x와 기준 어휘 y가 동시에 나타난 출현확률 $p(x,y)$의 로그 값으로 정규화 한다. 특히, $npmi(x;y)$는 최소 -1부터 최대 1 사이의 값으로 정규화 되기 때문에, 빈도의 차이로 인해 손해 볼 수 있는 어휘극성을 편차를 보완할 수 있다.

$$npmi(x;y) = \frac{pmi(x;y)}{-log\ p(x,y)} \qquad [식\ 3]$$

SO-PMI는 감성을 판단하고자 하는 어휘와 감성 기준 어휘들과의 차이를 통해 감성을 판단할 수 있다. 감성을 판단하고자 하는 어휘를 x, 긍정 감성 기준 어휘를 $pw \in PW$, 부정 감성 기준 어휘를 $nw \in NW$라 할 때, 어휘 x의 감성은 [식 4]처럼 구할 수 있다.

$$so\text{-}pmi(x) = \sum_{pw \in PW} pmi(x,pw) - \sum_{nw \in NW} pmi(x,nw) \qquad [식\ 4]$$

어휘 x의 감성이 긍정일수록 $pmi(x;pw)$가 $pmi(x;nw)$보다 크기 때문에 $SO-PMI(x)$는 양수일 것이고, 반대로 부정일수록 $pmi(x;nw)$가 $pmi(x;pw)$보다 크기 때문에 $SO-PMI(x)$는 음수일 것이다. 따라서 $SO-PMI(x)$가 양수인지 음수인지에 따라 어휘 x의 감성을 상대적으로 판단할 수 있다.

본 논문에서는 어휘극성 판단을 위해 NPMI를 적용했기 때문에, [식 5]처럼 다시 쓸 수 있다.

$$so\text{-}npmi = \sum_{pw \in PW} npmi(x,pw) - \sum_{nw \in NW} npmi(x,nw) \qquad [식\ 5]$$

다음 <표 1>과 <표 2>는 [식 5]를 통해 분석된 어휘의 긍정과 부정 감성분석 결과이다.

<표 1> SO-NPMI 감성분석 결과 (긍정 상위 10개)

단어	빈도	긍정	부정
carefree	561	**2.523251**	−0.967672
humerous	414	**2.522450**	−0.807261
eager	508	**2.506373**	−0.543314
vitality	500	**2.473954**	−0.954172
stamina	500	**2.437379**	−0.356946
pleased	503	**2.400760**	−0.642280
devoted	508	**2.379411**	−1.056447
cheerful	756	**2.361089**	−0.528244
faithful	578	**2.341097**	−0.784312
relieved	235	**2.309600**	−0.736311

<표 2> SO-NPMI 감성분석 결과 (부정 상위 10개)

단어	빈도	긍정	부정
solemn	502	−0.621858	**2.271372**
feeble	499	−0.621154	**2.269260**
bankrupt	499	−0.621154	**2.269260**
alienation	501	−0.307240	**2.164670**
rejected	501	−0.307240	**2.164670**
cramp	493	−0.741475	**2.139603**
pathetic	502	−0.109240	**2.096887**
dud	501	−0.108869	**2.096133**
ashamed	499	−0.407162	**2.060828**
weakness	502	−0.428962	**2.040384**

3. 실험

감성 강도의 편차 크기에 따른 감성분석 결과를 비교하기 위해, SO-PMI와 SO-NPMI의 감성 강도와 밀집도를 분석하였다.

<표 3> SO-PMI 감성분석 결과 (상위 10개)

단어	빈도	긍정	부정
humerous	414	**26.951499**	−8.511194
eager	509	**25.262761**	−1.808055
relieved	235	**25.252658**	−7.944914
carefree	563	**24.867332**	−11.810782
vitality	502	**24.521993**	−11.695669
solemn	502	− 6.555039	**25.380812**
feeble	499	− 6.549045	**25.362830**
bankrupt	499	− 6.549045	**25.362830**
alienation	501	0.356708	**23.066251**
rejected	501	0.356708	**23.066251**

<표 4> SO-NPMI 감성분석 결과 (상위 10개)

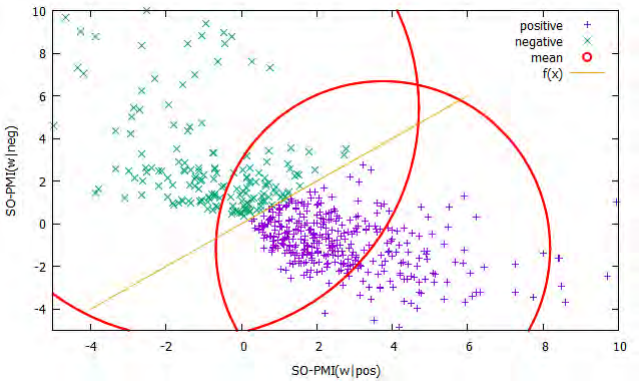
단어	빈도	긍정	부정
carefree	563	**2.523251**	− 0.967672
humerous	414	**2.522450**	− 0.807261
eager	509	**2.506373**	− 0.543314
vitality	502	**2.473954**	− 0.954172
stamina	503	**2.437379**	− 0.356946
solemn	502	− 0.621858	**2.271372**
feeble	499	− 0.621154	**2.269260**
bankrupt	499	− 0.621154	**2.269260**
alienation	500	− 0.307240	**2.164670**
rejected	500	− 0.307240	**2.164670**

<표 3>은 긍정과 부정 감성 강도가 높은 순으로 내림차순 정렬한 상위 10개의 SO-PMI 결과이고, <표 4>는 SO-NPMI 결과이다. 감성 강도에 따른 어휘의 순위가 감성에서의 절대적인 위치를 나타낸다고 판단하기는 어렵지만, 상대적인 차이를 통해 어휘가 어느 감성에 더 가까운지 또는 다른 어휘들과의 감성 거리가 어느 정도인지 판단하는데 이용할 수 있다.

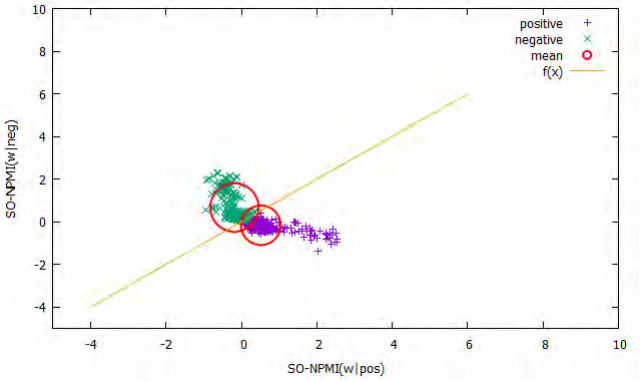
<표 5> SO-PMI와 SO-NPMI의 밀집도

	SO−PMI		SO−NPMI	
	긍정	부정	긍정	부정
최소	0.273936	0.369434	0.037874	0.051775
최대	26.951499	25.380812	2.523251	2.271372
평균	3.754108	5.421363	0.515642	0.668727
분산	19.660304	37.227373	0.274214	0.409706
표준편차	4.433994	6.101423	0.523654	0.640082

<표 5>에서 볼 수 있듯이, SO-PMI는 빈도로 인한 어휘의 극성 차이가 감성 강도에 그대로 반영되기 때문에 감성 강도의 밀집도가 떨어지는 것을 볼 수 있다. 반대로 SO-NPMI는 감성 강도가 일정한 범위 내에 분포하기 때문에 빈도로 인한 감성 강도의 손실을 줄일 수 있다.



(그림 1) SO-PMI의 어휘 분포 (가로 긍정, 세로 부정)



(그림 2) SO-PNMI의 어휘 분포 (가로 긍정, 세로 부정)

위의 (그림 1), (그림 2)에서 볼 수 있듯이, 기울기가 1인 정비례 1차 함수를 기준으로 긍정과 부정이 구분되는 것을 볼 수 있다. 특히 평균을 중심, 표준편차를 반지름으로 그린 원의 크기를 보면 동일한 축적에서 SO-PMI에 비해 SO-NPMI의 감성 강도가 더 밀집되어 있는 것을 볼 수 있다. 밀집도가 높으면 각 어휘의 감성 강도가 작은 원안에 포함될 확률이 높기 때문에 감성에 대한 추상화 모델을 만들기가 쉬울 뿐 아니라 높은 성능을 기대할 수 있다.

4. 결론

본 연구는 일상적인 언어 사용 환경에서 나타나는 어휘의 감성을 분석하기 위해, 소셜미디어에서 수집한 대량의 어휘 자원에 대한 감성분석을 분석하였다. 특히, PMI와 SO-PMI를 이용한 감성 분석에서 나타나는 극성 값의 편차로 인한 문제점을 NPMI를 통해 정규화 된 어휘 극성으로 해결하고자 하였다.

감성 분석한 결과 SO-PMI 보다 SO-NPMI를 이용한 감성 강도 분석 결과가 더 높은 밀집도를 보였다. 높은 밀집도는 감성을 추상화 하는데 유리하기 때문에 감성을 구분할 수 있는 계산 모델의 가능성을 볼 수 있었다.

본 연구는 빈도에 의한 극성 값의 편차를 고려하여 일상적으로 사용하는 어휘의 감성을 분석하기 위한 것으로, 일반화하기 위해서는 감성 강도의 밀집도를 고려하여, 밀집된 감성 강도를 구분하여 감성을 분류할 수 있는 선형 분류 모델을 연구하고 더 나아가 감성 내에 강도가 유사한 어휘들을 군집하고 분류할 수 있는 분류 모델을 연구할 필요가 있다.

사사

이 논문은 2010년도 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임 (2010-0022973)

참고문헌

[1] C.E. Osgood, G. Suci and P. Tannenbaum, “The Measurement of Meaning”, University of Illinois Press, 1957.

[2] Esuli, Andrea and S. Fabrizio, “Determining Term Subjectivity and Term Orientation for Opinion Mining”, EACL, Vol. 6, 2006.

[3] G. Bouma, “Normalized (Pointwise) Mutual Information in Collocation Extraction”, Proceedings of the Biennial GSCL Conference, pp. 31-4-, 2009.

[4] G. Siegle, “The ORIGINAL Balanced Affective Word List Project”, 1994. Retrieved 2014, http://www.sci.sdsu.edu/CAL/wordlist/origwordlist.html

[5] K. W. Church, and P. Hanks, “Word association norms, mutual information, and lexicography”, Journal of Computational Linguistics, Vol. 16, Issue 1, pp. 22-29, 1990.

[6] M. Thelwall, K. Buckley, G. Paltogloum D. Cai, and A. Kappas, “Sentiment strength detection in short informal text”. Journal of the American Society for Information Science and Technology, Vol. 61, No. 12, pp. 2544-2558, 2010.

[7] M.M. Bradley, and P.J. Lang, “Affective norms for English words (ANEW): Instruction manual and affective ratings”. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[8] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews”, Proceedings ACL ’02 Proceeding of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424, 2002.

[9] P.D. Turney and L. L. Michael, “Measuring praise and criticism: Inference of semantic orientation from association”, ACM Transactions on Information Systems (TOIS), Vol. 21, No. 4, pp. 315-346, 2003.