의견어중심의 의존트리패턴자질을 이용한 기계학습기반 한국어 블로그 문서 의견분류시스템

곽동민*, 이승욱**
*고려대학교 컴퓨터 정보통신대학원
**고려대학교 컴퓨터·전파통신공학과
e-mail:gsclub@gmail.com*, swlee@nlp.korea.ac.kr**

Machine Learning Based Blog Text Opinion Classification System Using Opinion Word Centered-Dependency Tree Pattern Features

Dong-Min Kwak*, Seung-Wook Lee**
*Graduate School of Computer and Information Technology, Korea University
** Dept. of Computer and Radio Communications Engineering, Korea University

요 약

블로그문서의 의견극성분류 연구는 주로 기계학습기법에 기반한 방법이었고, 이때 주로 활용된 자질은 명사, 동사 등의 품사정보와 의견어 어휘정보였다. 하지만 하나의 의견어 어휘만을 고려한다면 그 극성을 판별하는데 필요한 정보가 충분하지 않아 부정확한 결과를 도출하는 경우가 발생할 수 있다. 본 논문에서는 여러 어휘를 동시에 고려하였을 때 보다 정확한 의견분류를 수행할 수 있을 것이라는 가정을 세웠다. 본 논문에서는 효과적인 의견어휘자질의 추출을 위하여 의견이 내포될 가능성이 높은 의견어휘를 기반으로 의존구문분석을 통해 의존트리패턴을 추출하였고, 제안하는 PF-IDF가중치를 적용하여 지지벡터기계(SVM)와 다항시행접근 단순베이지안(MNNB)알고리즘으로 비교 실험을 수행하였다. 기준시스템인 TF-IDF가중치 기법에 비해 정확도(accuracy)가 지지벡터기계에서 5%, 다항시행접근 단순베이지안에서 8.9% 향상된 성능을 보였다.

1. 서론

블로그 문서에 표현된 의견을 기반으로 문서를 긍정과 부정으로 분류하는 것은 다양한 서비스와 연계 가능하다는 장점이 있어 최근 활발한 연구가 이루어지고 있다. 그 예로, 의견이 포함된 문서를 분류하기 위해서 PMI(pointwise mutual information)을 통한 SO(semantic orientation)을 예측하여 분류하는 시도가 있었고[1], 의견이 내포된 어휘자질기반으로 지지벡터기계를 활용한 기법도 연구되었다[2][3]. 단순히 문서에 포함된 단어들의 명사, 동사 등의 품사자질과 의견이 내포된 어휘자질을 기반으로 기계학습기법을 적용한 결과, 후자의 의견이 내포된 어휘자질을 기반으로 실험한 분류가 좀 더 높은 정확도를 보였다. 하지만 단순히 하나의 어휘에 의견이 내포된 경우 그 극성이 판단하는데 정보가 불충분하다. 이 방법에서는 여러 어휘와 결합이 되어야만 뚜렷한 극성을 갖게 되는 점을 고려하지 못하기 때문에 효과적인 분석이 힘들다.

본 연구에서는 의견극성분류 성능을 향상시키기 위해 다음 두가지 방법론을 제시하고 분석 평가한다. 첫째, 의견의 극성이 명확하게 드러난 다양한 의견표현패턴을 추출한다. 즉, 의견이 내포되거나 내포 될 가능성이 있는 기본어휘를 기준으로 의존문법을 적용하여 생성된 의존트리를 활용하여 다양한 의견표현패턴들을 생성한다. 둘째, 블로그문서에서 자주 나타나는 의존트리패턴을 위한 패턴가중치 기법(PF-IDF)을 제안한다.

2. 의견어 중심의 의존트리패턴자질을 이용한 의견분류시스템

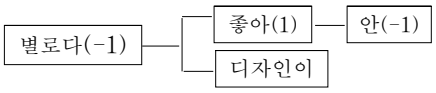
2.1 의견어휘사전 생성

의견이 내포되어 극성이 뚜렷한 어휘와 내포될 가능성이 있는 후보어휘를 대상으로 한다. 의견이 내포될 가능성이 있는 어휘는 단독으로는 극성의 여부를 판단하기 불충분하나 주변어와의 결합을 통해 극성이 강해지거나 발생할 수 있는 어휘이다. 예를 들어 부정보조용언인 '-않 '은 극성이 모호하나 '좋지 않다', '나쁘지 않다' 와 같이 주변어와의 결합을 통해 극성이 발생할 수 있다. 또한 '-가능성있다.'와 같은 경우도 '수상할 가능성있다.'와 같이 긍정의 의미를 더욱 명확히 하는 경향이 있다.

이런 의견어 어휘들을 활용하기에 앞서 수집 단계가 필수적이다. 영어권에서는 SentiWordNet과 같은 공개된 의견어 사전이 있는 반면, 활용 가능한 한국어 의견어 사전은 찾기 힘들다. 이를 극복하기 위해 SentiWordNet 사전을 한국어로 자동 번역하여 이용하였다. 우선 대표 영어 의견어 어휘를 선정하여 시소러스사전으로부터 대역어를 검색하였다. 검색 결과에서 빈도수가 높은 한국어 대역어를 중심으로 긍정어휘 980개 부정어휘 970개를 구축하였다.

2.2 의견어휘 중심 의존트리패턴 추출

의존구문분석기를 통해 생성된 의존트리 패턴을 루트노드의 극성에 맞게 먼저 긍정(positive), 부정(negative), 혼합(mix)별로 분류한다. (그림1)과 같이 혼합된 경우를 없애기 위해 전산학을 전공한 2 명의 평가자가 직접 긍정, 부정으로 분류하였다. 긍정 의존트리패턴 1,280개, 부정 의존트리패턴 1,320개를 생성하였다.



(그림 1)혼합으로 분류된 의존트리패턴

2.3 PF-IDF 가중치

본 논문에서 제안한 패턴 자질의 가중치는 기존의 TF-IDF 기법을 그대로 사용하기가 적합하지 않으므로 아래와 같이 PF-IDF 가중치 기법으로 변형하였다. 식(2)와 같이 문서 내에서 트리패턴 p가 나타난 패턴 빈도수(pf: pattern frequency) pf_p와 역문서 빈도수 (idf: inverse document frequency)식(1)을 곱으로 나타낸다.

$$idf_p = \log_2 \frac{N}{df_p} \qquad (1)$$

$$weight_p = pf_p \times idf_p \qquad (2)$$

3. 실험 구성 및 결과분석

3.1 실험 구성 및 결과

본 논문에서는 학습을 위하여 전산학 전공자 2명이 참여하여 긍정문서 670개, 부정문서 530개의 블로그문서를 구축하였다. 의견극성분류의 성능을 비교하기 위한 기준시스템으로 의견어자질을 사용한 분류시스템과 비교실험 하였고 지지벡터기계, 다항시행접근 단순베이지안 분류알고리즘으로 학습하였다. <표4>와 <표5>는 PF-IDF가중치를 의존트리패턴에 적용하여 학습한 뒤에 10-fold cross validation 방식으로 실험하여 나온 오류율(ErrorRate), 정확률(precision), 재현율(recall), $F_1 - Measure$, 정확도(accuracy)이다.

<표 4> 지지벡터기계(SVM) 알고리즘 성능측정

평가방법	의견어기반 (TF-IDF)		의존트리패턴기반 (PF-IDF)	
	긍정	부정	긍정	부정
Error Rate	0.390	0.168	0.376	0.111
Precision	0.759	0.705	0.798	0.771
Recall	0.832	0.602	0.889	0.624
$F_1 - Measure$	0.794	0.649	0.841	0.689
Accuracy	0.74		0.79	

<표 5> Naive Bayes Multinomial 알고리즘 성능측정

평가방법	의견어기반 (TF-IDF)		의존트리패턴기반 (PF-IDF)	
	긍정	부정	긍정	부정
Error Rate	0.398	0.195	0.238	0.131
Precision	0.753	0.673	0.86	0.777
Recall	0.805	0.602	0.869	0.762
$F_1 - Measure$	0.778	0.635	0.864	0.769
Accuracy	0.72		0.829	

PF-IDF가중치가 적용된 의존트리패턴자질기반의 분류기는 TF-IDF가중치를 적용한 의견어자질기반의 분류기보다 정확도가 지지벡터기계 모형에서는 5%, 다항시행접근 단순베이지안 모형에서는 8.9% 향상되었다.

3.2 의존트리패턴 자질의 유효성

의존트리패턴자질은 모든 척도에서 의견어자질기반의 분류 시스템의 성능보다 향상되었다. 이는 본 논문에서 제안한 의존문법기반의 구문분석이 주변어와의 관계를 통해 생성된 의존트리가 의견이 명확한 패턴을 찾아주고, 새로운 의견표현을 찾아주었다는 것으로 분석된다. 즉, 의견의 대상을 찾게 됨으로써 극성이 명확해지고, 부정보조용언('XXX을 좋아하지 않는'), 이중부정('XXX을 좋아하지 않는 사람은 없다'), 강조부사('정말로 좋은') 등과 같이 주변어에 의해 의견의 극성이 변화된 현상을 고려할 할 수 있게 되어 전체적으로 보다 정확한 분류를 수행한 것으로 분석된다.

3.3 오류율 분석

제안한 방법을 사용하였을 경우가 제일 낮은 오류율을 보여주며, 다수의 블로그문서들이 의존트리패턴을 통하여 발생된 의견표현들을 포함하고 있기 때문이라고 분석된다. 오류율이 의존트리패턴자질을 사용함으로써 감소한 이유 중 하나는 단일 의견어휘만으로는 모호한 극성의 표현으로 인해 긍정과 부정을 뚜렷하게 나누기 힘들었던 반면, 의존트리패턴은 주변어와의 관계를 통하여 모호했던 극성을 보다 명확하게 하고, 새로운 의견표현을 찾을 수 있었기 때문으로 분석된다.

4. 결론 및 향후 과제

본 논문에서 제안한 방법은 의존트리정보와 PF-IDF 가중치를 활용하여 기존에 연구되었던 의견어자질을 선택한 방법보다 우수한 성능을 보였다.

향후 과제로는 의존구문분석의 성능을 개선하고 의존트리패턴을 통하여 극성의 강도를 측정한 뒤에 의견표현이 강한 문장을 추출하여 가중치를 적용할 수 있는 새로운 방식의 가중치기법을 연구해야 할 것이다.

참고문헌

[1] Turney,Peter D., Measuring praise and criticism:Inference of semantic orientation from association, ACM Transactions on Information System Vol.21, No.4, pages 315-346, 2003

[2] 황재원, 고영중, 감정자질을 이용한 한국어 문장 및 문서감정 분류시스템, 정보과학회, 제14권 제3호, 2008.5

[3] Pang, Bo and Lillian Lee, A Sentimental Education:Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, 42nd ACL pages 271~278, 2004