상품평 분류 성능 향상을 위한 긍정/부정 사전 자동 구축

송종석○ 백종범 이수원
숭실대학교 대학원 컴퓨터학과
ysong77@mining.ssu.ac.kr, jbb100@ssu.ac.kr, swlee@ssu.ac.kr

Automatic Construction of Positive/Negative Dictionary to Improve Performance of Product Review Classification

Jongseok Song○ Jongbuk Baik Soowon Lee
Dept. of Computing, Graduate School, Soongsil University

1. 서 론

오피니언 마이닝에서 상품평을 분류하기 위해서는 긍정/부정 사전과 같은 어휘 사전을 사용한다. 어휘 사전은 도메인마다 수동으로 구축하여 사용할 수도 있다. 하지만 관리자가 여러 도메인마다 긍정/부정 사전을 수동으로 구축하는 것은 구축비용, 시간적 비용, 유지보수 등의 관점에서 비효율적이라고 할 수 있다. 또한 긍정/부정 사전을 여러 도메인에 공통으로 구축하여 사용할 때에는 도메인마다 다르게 사용될 수 있는 서술어의 의미 방향을 반영하지 못하는 문제점이 존재한다. 예를 들어 '크다'라는 서술어는 의류 도메인에서 "사이즈가 크다"와 같이 부정적인 의미 방향으로 사용되지만, 전자제품 도메인에서는 "화면이 크다"와 같이 긍정적인 의미방향으로 사용되어 도메인별로 다른 의미 방향을 갖는다. 이와 같은 문제를 해결하기 위하여, 본 연구에서는 도메인별 긍정/부정 사전을 자동으로 구축하는 것을 목표로 하며, 한국어 문장에 존재하는 접속 부사, 연결어미 정보를 활용하여 서술어들의 의미방향을 자동으로 분류하여 도메인별 긍정/부정 사전을 자동으로 구축하는 방법을 제안한다.

2. 긍정/부정 사전 구축 시스템

본 연구에서 제안하는 긍정/부정 사전 구축 방법은 평점이 포함된 상품평을 분석하여 평점 긍정/부정 사전을 구축하고, 구축된 평점 긍정/부정 사전을 이용하여 여러 도메인에서 공통적으로 활용되는 공통 긍정/부정 사전을 구축한다. 구축된 공통 긍정/부정 사전의 서술어와의 접속정보를 활용하여 도메인별로 서술어의 의미방향을 긍정/부정으로 분류함으로써 도메인 긍정/부정 사전을 자동으로 구축한다.

2.1 평점 긍정/부정 사전 구축 방법

수집된 상품평은 평점에 따라 긍정/부정 상품평으로 분류된다. 긍정적인 상품평에서 출현한 서술어의 출현 빈도 정보와 부정적인 상품평에서 출현한 서술어의 출현 빈도 정보의 차이로써 서술어의 긍정/부정을 분류할 수 있다. 하지만 온라인 쇼핑몰은 긍정적인 상품평이 부정적인 상품평 보다 상대적으로 많은 특징을 가지고 있다. 이러한 상품평의 특징을 고려하지 않고 서술어의 단순 빈도 차이를 이용하여 서술어의 의미방향을 분류하게 될 경우 분류한 서술어가 긍정적인 성향으로 치우치는 문제가 존재하게 된다. 이 문제를 해결하기 위해 본 연구에서는 서술어의 출현 비율을 고려하여 서술어의 극성을 계산한다. 서술어의 극성을 계산하는 방법은 [식 1]과 같다. 서술어의 의미방향은 $wordPolairty^d(w)$ 가 0보다 크면 긍정, 0보다 작으면 부정으로 분류된다.

$$wordPolarity^d(w) = \frac{f_p^d(w)}{reviewCnt_p^d} - \frac{f_n^d(w)}{reviewCnt_n^d}$$

$wordPolairty^d(w)$: 도메인 d에서 서술어 w의 긍정/부정 극성
$f_p^d(w)$: 도메인 d에서 서술어 w가 긍정적 상품평에 출현한 수
$f_n^d(w)$: 도메인 d에서 서술어 w가 부정적 상품평에 출현한 수
$reviewCnt_p^d$: 도메인 d에서의 긍정적 상품평 수
$reviewCnt_n^d$: 도메인 d에서의 부정적 상품평 수

[식 1] 평점을 정보를 활용한 서술어의 극성 계산 식

2.2 공통 긍정/부정 사전 구축 방법

2.1절에서 구축한 평점 긍정/부정 사전은 공통 긍정/부정 사전을 구축하기 위해 활용한다. 공통 긍정/부정 사전은 평점 긍정/부정 사전의 교집합 부분 즉, 여러 도메인에서 서술어가 공통으로 사용되고 그 의미방향까지 동일하게

사용되는 서술어의 집합이다. 예를 들어 각 도메인에서 '좋다'라는 서술어가 사용되고 의미방향도 긍정적으로 동일하게 사용되는 경우에 '좋다'라는 서술어는 공통 긍정/부정 사전으로 추출된다. 추출된 공통 서술어는 공통 긍정/부정 사전에 삽입되어 도메인 긍정/부정 사전을 구축하기 위한 초기 서술어로 사용된다.

2.3 도메인 긍정/부정 사전 구축 방법

공통 긍정/부정 사전에 구축된 공통 서술어는 도메인별로 도메인 긍정/부정 사전에 초기화 되어 Seed Word로 활용된다. 초기화된 Seed Word와 서술어사이의 접속부사 및 연결어미 정보를 이용하여 도메인별로 서술어의 의미방향을 다시 분류하게 된다. Seed Word와 서술어사이의 접속정보가 역접 관계일 때 서술어는 Seed Word의 의미방향 {1,-1}과 반대 의미방향(×-1)이 부여된다. 새롭게 분류된 서술어들은 도메인 긍정/부정 사전에 확장한다. 도메인 긍정/부정 사전에 확장된 서술어들은 더 많은 상품평을 분석하는 과정을 반복하여 도메인 긍정/부정 사전에 확장/구축된다. 모든 상품평을 분석하여 서술어가 긍정 또는 부정으로 예측된 수를 이용하여 도메인에서 서술어가 일반적으로 사용되는 의미방향으로 분류한다. 서술어들을 긍정 또는 부정으로 분류해 주기 위해 [식 2]를 사용하여 서술어의 의미방향을 결정한다. 서술어의 의미방향은 $wordpolarityScore^d(w)$가 0보다 크면 긍정, 0보다 작으면 부정으로 분류된다.

$$wordPolarityScore^d(w) = \frac{f^d(w_p) - f^d(w_n)}{f^d(w_p) + f^d(w_n)}$$

$$-1 \le wordPolarityScore^d(w) \le 1$$

$wordpolarityScore^d(w)$:도메인 d에서 서술어 w의 긍정/부정 극성

$f^d(w_p)$:도메인 d에서 서술어 w가 긍정적으로 예측된 수

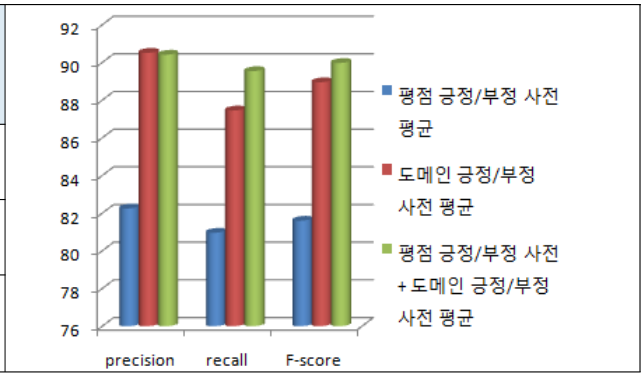
$f^d(w_n)$:도메인 d에서 서술어 w가 부정적으로 예측된 수

[식 2] 접속정보를 활용한 서술어의 극성 계산 식

3. 실험 및 평가

평가 데이터 구축을 위해 긍정/부정 사전을 구축할 때 사용한 상품평 중에서 각 도메인별로 200개의 상품평을 무작위로 추출하고 상품평에 존재하는 서술어들의 의미방향을 수동으로 태깅하였다. 평가를 위해 평가 데이터에 태깅된 서술어의 의미방향과 긍정/부정 사전에 구축된 의미방향과 일치하는지를 판단하였다. 평가척도로는 Precision, Recall, F-Score를 사용하였다. 실험을 위해 각 도메인에서 단계별로 구축된 평점 긍정/부정 사전과 도메인 긍정/부정 사전의 평균 성능을 비교하고, 두 사전을 같이 사용하였을 때의 성능을 분석하였다.

평가척도	평점 긍정/부정 사전 평균	도메인 긍정/부정 사전 평균	평점 긍정/부정 사전 + 도메인 긍정/부정 사전 평균
precision	82.25	90.51	90.41
recall	80.96	87.45	89.54
F-score	81.6	88.95	89.97



[표 1] 긍정/부정 사전 성능 평가

실험 결과, 평점을 활용하여 구축한 평점 긍정/부정 사전보다 접속정보를 활용하여 구축한 도메인 긍정/부정 사전이 높은 성능을 보였다. 또한, 평점 긍정/부정 사전과 도메인 긍정/부정 사전을 모두 이용한 경우가 하나의 긍정/부정 사전만을 사용한 것보다 F-Score가 향상된 것을 확인하였다.

Acknowledge

본 연구는 '서울시 산학연 협력사업(10581)'의 지원으로 수행되었다

참고문헌

[1] Peter D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp.417-424, July 2002.

[2] Xiaowen Ding, Bing Liu and Philip Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", Department of Computer Science, University of Illinois at Chicago, WSDM 2008.

[3] Jaeseok Myung, Dongjoo Lee, Sang-goo Lee, "A Korean Product Review Analysis System Using a Semi-Automatically Constructed Semantic Dictionary", KIISE: Software and Applications, VOL.35, NUM.6 pp.392-403, 2008.