한국컴퓨터정보학회 동계학술대회 논문집 제23권 제1호 (2015. 1)

도메인 별 감성분석을 위한 도메인 맞춤형 감성사전 구축 기법

김다해°, 조태민*, 이지형*
°단국대학교 소프트웨어학과
*성균관대학교 전자전기컴퓨터공학과
e-mail:kimdh35@dankook.ac.kr°, {tmchojo, john}@skku.edu*

A Domain Adaptive Sentiment Dictionary Construction Method for Domain Sentiment Analysis

Dahae Kim°, Taemin Cho*, Jee-Hyong Lee*†
°Dept. of Computer Software, Dankook University
*Dept. of Electrical and Computer Engineering, SungKyunKawn University

● 요 약 ●

SNS의 확산으로 대중들은 제품, 서비스, 사회적 이슈 등 다양한 도메인에 대하여 자신의 기분이나 의견을 적극적으로 표현하고 있다. 이에 따라 SNS를 분석하여 제품의 수요, TV 시청률, 주가 등의 다양한 현상을 예측하는 데 있어 감성분석을 활용하는 연구가 활발히 진행되고 있다. 감성분석은 각 어휘에 대한 품사, 극성, 감성지수를 규정하고 있는 감성사전을 기반으로 이루어진다. 하지만 동일한 단어라도 도메인에 따라 중요도가 달라지기 때문에 도메인의 특성을 고려한 감성사전을 사용해야 할 필요성이 있다. 따라서 본 연구에서는 다양한 도메인에 대하여 각각의 특성에 맞게 더욱 정확한 감성분석을 할 수 있도록 도메인 맞춤형 감성사전을 구축하는 기법을 제안한다. 도메인 별로 긍/부정 평가에 있어 중요한 척도가 되는 단어들을 도메인 감성어휘로 선별하여 목록을 구축하고, 각 감성어휘의 중요도에 따라 도메인 감성지수를 새롭게 정의하였다. 실험 결과, 평가 도메인에 적합한 감성사전이 다른 도메인의 감성사전 및 범용 감성사전보다 우수한 성능을 보였다. 이를 통해 도메인 맞춤형 감성사전 구축 기법의 효용성을 확인하였다.

키워드: 도메인 감성분석(domain sentiment analysis), 감성어휘(sentiment word), 감성지수(sentiment score), 감성사전(sentiment dictionary)

I. 서 론

최근 인터넷은 소셜 네트워크 서비스(Social Network Service, 이하 SNS)의 확산으로 인하여 자신의 기분이나 의견을 적극적으로 표현하는 공간이 되었다. 대중들은 제품, 서비스, 각종 사회적 이슈 등의 다양한 도메인에 대하여 서로의 생각을 공유하고 있다. 이러한 SNS의 분석을 통해 제품의 수요[3], TV 시청률[4], 주가 상승[5]과 같은 다양한 현상의 예측을 가능하게 하는 감성분석과 관련된 많은 연구들이 활발히 수행되고 있다[1-6].

감성분석은 각 어휘에 대한 품사, 극성, 감성지수를 규정하고 있는 감성사전을 기반으로 이루어진다. 하지만 동일한 단어라도 도메인에 따라 중요도가 달라지기 때문에 모든 도메인에 같은 감성사전을 사용하게 되면 다음과 같은 문제점이 발생한다. 예를 들어, '그 음식점은 맛있다.'라는 문장을 음식 도메인에서 분석할 경우, '맛있다'는 해당 음식점을 맛집으로 분류하는 핵심적인 정보로 활용되어야 한다. 반면, '영화관의 팝콘이 맛있다.'라는 문장을 영화 도메인에서 분석할 경우, '맛있다'는 영화에 대한 직접적인 평가가 아니므로 음식 도메인에서와 같이 중요한 정보로 활용되어선 안 된다. 따라서 하나의 감성사전을 범용적으로 사용하기보다는 도메인의 특성을 고려한 감성사전을 재구축하여 사용해야 할 필요성이 있다.

본 논문에서는 도메인 별 맞춤형 감성분석을 위해 각각의 도메인의 특성과 분석의 목적에 맞게 사용할 수 있도록 도메인 맞춤형 감성사전 구축 기법을 제안한다. 이를 위해 도메인 별로 긍/부정 판별에 있어 중요한 척도가 되는 감성어휘를 선별하고, 감성지수에 가중치를 부여하여 중요성을 강조시킴으로써 도메인 별로 감성어휘의 의미를 새롭게 정의하고자 한다. 이를 통해 감성분석의 정확성이 향상된다면, 관련 기관들은 더욱 효과적으로 분석결과를 활용할 수 있을 것이며 궁극적으로는 매출 향상에 기여할 수 있다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3장에서는 제안하는 도메인 맞춤형 감성사전 구축 기법에 대해 기술하고, 4장에서는 실험 및 결과를 기술한다. 마지막으로 5장에서는 결론

및 향후 연구에 대해서 논의한다.

II. 관련 연구

최근 도메인의 특징을 고려하여 감성사전을 구축하는 연구들이 활발히 수행되고 있다. 이 연구들은 동일한 어휘라 하더라도 도메인에 따라 중요도가 달라지는 감성분석의 특징을 고려하여 감성사전을 구축했다는 공통점이 있다.

Myung [3]은 소비자들의 상품평을 분석하기 위하여 쇼핑몰 시장을 도메인으로 선정하였으며, 상품의 특징을 표현하는 단어와 각 단어들의 극성 정보로 정의된 감성사전을 구축하였다. 구문 분석기를 통해 단어들의 관계를 수식 거리를 이용하여 계산함으로써, 상품의 속성을 나타내는 단어와 상품의 속성을 설명 하는 단어로 구분하였다. 그리고 실제 상품평을 수집하여 제안하는 상품 점수 계산 모델을 통해 상품에 대한 순위를 추론할 수 있음을 보였다.

Yu [4]는 주가의 등락을 예측하기 위해서 주식 시장을 도메인으로 선정하였다. 주식 관련 뉴스 기사에 등장한 명사들을 추출하여 해당 어휘들의 빈도수를 바탕으로 긍/부정 지수를 재계산함으로써 주식 시장에 특화된 감성사전을 구축하였다. 이를 통해 주가등락 예측의 정확성을 향상시킬 수 있음을 보였다.

Young [5]은 정치 분야를 도메인으로 선정하였다. 정치 관련 뉴스인 범죄, 정치, 분야의 기사에 나타난 단어들을 추출하여 PMI기법을 사용함으로써 단어의 극성을 재 정의하였다. 그리고 정치 도메인에 특화된 감성사전은 다른 범용 감성사전들보다 정치 분야 말뭉치의 긍/부정 판별에 있어 더욱 정확성이 높음을 보였다.

하지만 위 연구들의 도메인 맞춤형 감성사전 구축 과정은 자동으로 이루어지지 않았기 때문에 사전 구축 기법의 재사용성이 낮다. 또한 모두 한 가지 도메인에 초점을 맞추고 있어 다른 도메인 분석에는 제안하는 기법을 적용할 수 없는 한계점이 존재한다.

III. 도메인 감성어휘를 활용한 도메인 맞춤형 감성사전 구축

본 장에서는 특정 도메인의 감성분석에 적합한 도메인 맞춤형 감성사전을 구축하는 방법에 대해서 기술 한다. 그림 1은 그 과정을 나타낸다. 도메인 감성어휘는 도메인의 특징을 잘 나타내는 단어들로 선별하고, 추출된 감성어휘들에 대해서 감성지수를 재계산 한다. 말뭉치의 긍/부정 판별은 위와 같은 방법으로 구축한 도메인 맞춤형 감성사전을 이용한다. 각 감성어휘에 대한 품사, 극성, 감성지수는 연세대학교 정보대학원 디지털서비스 연구실에서 구축한 범용 감성사전인 오픈한글[7] 서비스를 통해 제공받는다.

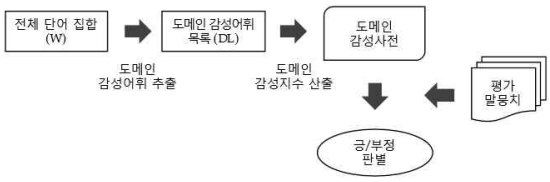

그림 6. 도메인 맞춤형 감성사전 구축 과정
Fig. 6. Building Process of Domain Sentiment Dictionary

1. 도메인 감성어휘 목록 구축

도메인 감성어휘 목록 DL은 범용 감성사전에 존재하는 전체 단어 집합 W중에서 해당 도메인의 특징을 나타내지 않는 단어들을 제외하여 구축한다. 그 구체 적인 제외 기준 및 과정은 다음과 같다. 식 1은 특정 도메인에 대하여 중요도가 낮은 단어들 FW를 나타낸 다. 단어 w가 해당 도메인에서 출현한 빈도수 count(w)가 α 미만이면 도메인 내에서 중요하지 않은 것으로 가정한다. 식 2는 특정 도메인과 상관없이 일반적으로 자주 쓰이는 단어들 CW를 나타낸다. 모든 도메인에서 공통적으로 나타난 빈도수 count(w)가 β 이상이면 일반적인 감성어휘인 것으로 가정한다.

$$FW = \{w | count(w) < \alpha\} \quad (1)$$

$$CW = \{w | \text{모든 도메인에서 } count(w) \geq \beta\} \quad (2)$$

$$DL = W - (FW \cup CW) \quad (3)$$

식 3은 전체 단어 집합W으로부터 식 1, 2의 기준을 충족하는 단어들 FW, CW를 제외하여 구축한 최종 감성어휘 목록 DL을 나타낸다.

2. 도메인 감성지수 산출

도메인 감성지수는 범용 감성사전에서 제공하는 감성지수를 기반으로 가중치를 부여하며, 구체적인 계산 방법은 다음과 같다. 식 4는 도메인에서 출현한 감성어휘들 중 특정 감성어휘 s가 차지하는 비율을 나타내는 감성 빈도수 sentiment frequency(SF)다. 이는 감성어휘가 도메인 말뭉치 집합 D에서 나타난 빈도수 f(s) 대비 도메인 내에서 가장 많이 나타난 감성 어휘의 빈도수 max(f(s))의 비율로 정의한다. 이를 통해 도메인 내에서 출현한 단어들 중에서 감성어휘 s 의 상대적인 중요도를 알 수 있다. 식 5는 감성어휘 s와 도메인 간의 관련 정도를 나타내는 도메인 연관성 domain relation(DR)이다. 이는 감성어휘 s가 전체 말뭉치 T에서 나타난 빈도수 f(s,T) 대비 해당 도메인 말뭉치 D에서 나타난 빈도수 f(s,D)의 비율로 정의한 다. 도메인 내에서의 비교가 아닌 다른 도메인들과의 비교를 통해 감성어휘 s의 도메인 연관성을 더욱 정확하게 표현할 수 있다.

$$\text{sentimen frequency(SF)} = 1 + \frac{f(s)}{\max_{s \in D}(f(s))} \quad (4)$$

$$\text{domain relation(DR)} = 1 + \frac{f(s,D)}{f(s,T)} \quad (5)$$

$$\text{domain score} = \text{general score} * SF * DR \quad (6)$$

범용 감성지수(general score)에 대한 가중치는 식 4와 식 5의 곱을 통해 계산한다. 식 6은 감성어휘 s의 최종적인 도메인 감성지수(domain score)다.

IV. 실험 및 결과

1. 실험 데이터

본 실험에서는 국립국어원[8]에서 제공하는 21 세기 세종 말뭉치 중 일부를 사용하였다. 말뭉치는 '영화', '인물', '스포츠'도메인에 대하여 각 130개씩 총 390 개를 구축하였다. 각 도메인은 100개의 말뭉치를 학습 말뭉치로 사용하였고, 30개의 말뭉치를 평가 말뭉치로 사용하였다. 긍정과 부정 말뭉치의 개수는 동일하며, 전문가 3명에 의해서 말뭉치의 극성을 부여하였다. 말뭉치의 형태소 분해 및 품사 태깅을 위한 형태소 분석기로는 KOMORAN[9]을 사용하였다.

2. 실험 방법

본 장 1절의 도메인 별 감성어휘 목록 구축을 위해서, 식 1의 α는 5 로, 식 2의 β는 50으로 설정하였다. 평가 도메인으로는 '영화'를 선정하였고, '영화' 도메인 감성사전을 '인물', '스포츠' 도메인 감성사전 및 범용 감성사전과 비교하여 실험을 진행하였다. 각 말뭉치에 대한 감성지수는 말뭉치에서 나타난 모든 단어들의 감성지수의 합으로 계산하며, 도메인 감성사전에 없는 단어는 도메인과 관계없는 단어라고 가정하므로 범용 감성사전의 감성지수를 이용하여 계산한다.

3. 실험 평가

계산된 말뭉치의 감성지수가 0보다 큰 경우에는 긍정 말뭉치로, 작은 경우에는 부정 말뭉치로 평가한다. 말뭉치의 긍/부정을 예측한 결과를 평가하기 위한 지표로는 정확률(precision), 재현율(recall) 그리고 F-척도(F-measure)를 이용하였다.

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (7)$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (8)$$

$$\text{F-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

4. 실험 결과

실험에 사용한 모든 감성사전들은 도메인에 나타난 감성어휘들의 중요도에 따라 감성지수가 다르게 산출되어 있다. 그림 2는 동일한 말뭉치에 대해서 영화, 인물, 스포츠, 범용 감성사전을 사용하여 긍/부정 평가의 성능을 비교한 그래프이다. 긍정과 부정 간의 F-척도의 편차를 보면, 영화 도메인 감성사전을 적용했을 경우가 다른 도메인 감성사전을 적용했을 경우보다 적은 것을 알 수 있다. 이는 도메인 맞춤형 감성사전을 사용했을 경우 감성분석이 더욱 정확하게 이루어짐을 보여준다. 또한 전체적인 그래프의 모습을 보았을 때, 영화 도메인 감성사전(83.4%)이 인물(59.2%), 스포츠(77.4%), 범용(80.0%) 감성사전보다 성능이 우수함을 알 수 있다.

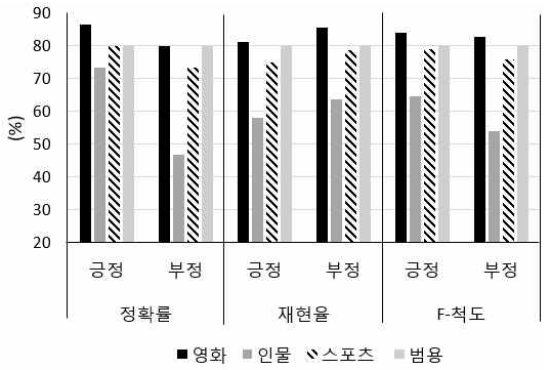

그림 7. 감성사전 성능 비교
Fig. 7. Performance Comparison of Sentiment Dictionaries

실험을 통해 도메인의 주요 감성어휘를 추출하여 중요성을 강조시키면 해당 분야의 감성분석에 있어 정확성을 향상시킬 수 있음을 보였다. 이를 통해 도메인 맞춤형 감성사전 구축 기법의 효용성을 확인하였다.

IV. 결론 및 향후 연구

본 논문에서는 정확한 감성분석을 위하여 각 도메인의 특성에 맞는 도메인 맞춤형 감성사전을 구축하는 기법을 제시하였다. 이를 위한 제안 방법은 도메인과 관련이 높은 감성 어휘들을 선별하고, 도메인 별로 중요도에 따라 감성어휘들의 감성지수를 다르게 정의하였다. 실험을 통해 본 논문에서 제안하는 방법으로 구축한 도메인 맞춤형 감성사전이 해당 도메인의 감성분석에 가장 적합함을 확인하였다. 이를 통해 제안하는 방법으로 다양한 도메인에 대해서 도메인 맞춤형 감성사전 구축 기법의 효용성을 확인하였다.

한편, 몇몇의 특정 어휘들은 도메인에 따라 극성이 전환되는 경우가 존재한다. 향후 감성사전을 구축하는 데 있어 도메인에 따라 극성이 바뀌는 것을 고려한다면, 더욱 양질의 감성사전이 될 것이다.

Acknowledgement

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단-차세대정보 컴퓨팅기술개발사업의 지원(No. NRF-2014M3C4A7030503)과 정부(미래창조과학부) 및 한국산업기술평가관리원의 SW컴퓨팅산업 융합원천기술개발사업의 일환으로 수행된 연구임(2014-044-024-002).

참고문헌

[1] Hana Cho, Yeounoh Chung, Jaedong Lee, and Jee-Hyong Lee, “Sentiment Analysis Using News Comments for Public Opinion Mining,” Korean Institute of Intelligence Systems, Vol. 23, No. 1, pp. 149-150, 2013.

[2] TaeMin Cho, Hana Cho, Jaedong Lee, and Jee-Hyong Lee, “TV Drama Rating Prediction based on Sentiment Analysis of Viewers’ Comments,” KIIS Spring Conference 2013, Vol. 24, No. 1, pp. 83-84, 2014.

[3] Jaeseok Myung, Dongjoo Lee, and Sang-goo Lee, “A Korean Product Review Analysis System Using a Semi-Automatically Constructed Semantic Dictionary,” Journal of KIISE :Software and Applications, Vol. 35, No. 6, pp. 392-403, 2008.

[4] Eunji Yu, Yosin Kim, Namgyu Kim, and SeungRyul Jeong, “Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary,” Journal of Intelligence and Information System, Vol. 19, No. 1, pp. 95-110, 2013.

[5] Bo Pang, and Lillian Lee, “Opinion Mining and Sentiment Analysis,” Foundations and Trends in Information Retrieval, Vol. 2, No. 1-2, pp. 1-135, 2008.

[6] Lori Young, and Stuart Soroka, “Affective News: The Automated Coding of Sentiment in Political Texts,” Political Communication, Vol. 29, No. 2, pp. 205-231, 2012.

[7] OpenHangul API. http://openhangul.com/senti_text

[8] The National Institute of the Korean Language. http://ithub.korean.go.kr

[9] Shineware KOMORAN. http://shineware.tistory.com/entry/KOMORAN-ver-23