한글 감성 분류를 위한 감성 사전 구축에 관한 실험적 연구

An Experimental Study on Construction of a Sentiment Dictionary for Sentiment Classification of Korean Texts

저자 (Authors)
김수연, 정유경, 송민
Su Yeon Kim, Yoo Kyung Jeong, Min Song

출처 (Source)
한국도서관정보학회 동계 학술발표회 , 2015.11, 143-150(8 pages)

발행처 (Publisher)
한국도서관정보학회
Korean Library And Information Science Society

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06559131

APA Style
김수연, 정유경, 송민 (2015). 한글 감성 분류를 위한 감성 사전 구축에 관한 실험적 연구. 한국도서관정보학회 동계 학술발표회, 143-150

4

An Experimental Study on Construction of a Sentiment Dictionary for Sentiment Classification of Korean Texts

Suyeon Kim, Yookyung Jeong, Min Song
(Yonsei Univ.)

An Experimental Study on Construction of a Sentiment Dictionary for Sentiment Classification of Korean Texts

Su Yeon Kim, Yoo Kyung Jeong, Min Song
Department of Library and Information Science, Yonsei University

Slang and Urban variations of Korean are pervasive in the online community diverse, and this pauses a daunting challenge to process Korean natural language text. Even if existed sentiment dictionaries are available to classify sentiment, new terms are constantly coined in a short-term period. This hinders to keep updating the sentiment dictionary manually. To tackle this problem, this study suggests a method to extend the sentiment dictionary automatically to classify Korean sentiment.

To this end, NAVER movie comments were collected. Total 1,219 movies were collected, and 2,560 thousand comments of each movie were collected. The sentiment dictionary and classification module were implemented in Java. To generate a sentiment classification model, 30 thousand movie comments of which movie rating was from 1 (negative) to 10 (positive) were extracted and used as train data. The experiment was conducted in the following order. First, preprocessing on Korean was done through sentence split, Korean morphological analysis, stopword removal, and lemmatization. Next, the bag of words-based Korean sentiment classification model and dictionary-based sentiment classification model were generated. Third, with the Maximum Entropy Classifier, sentiment polarity was classified. Once new Korean text is entered, the trained classifier operates to classify sentiment polarity, and words are added to the sentiment dictionary.

According to the results of evaluating performance with 10-fold cross validation, existing dictionary-based sentiment classification achieved the performance of 0.5122 +/- 0.0058, and the extended sentiment dictionary obtained the performance of 0.6677 +/- 0.0218.

In the future, we plan to incorporate other features such as part of speech tagging to filter the movie comments not including sentiment words. In addition, we will employ different approaches to calculate weight on new sentiment words added to the sentiment dictionary.

한글 감성 분류를 위한 감성 사전 구축에 관한 실험적 연구*

김수연, 정유경, 송민
연세대학교 문헌정보학과

〈목 차〉

Ⅰ. 서 론

오피니언 마이닝(Opinion mining)은 긍정·부정 혹은 기쁨·슬픔·행복·분노 등과 같이 개인의 의견이나 감성을 극성(Polarity)으로 정의하고, 마이닝 기법을 통해 텍스트가 어떤 극성을 나타내는지, 출현빈도 및 언어학적 강도(Strength)에 따라 그 세기가 어느 정도인지를 분석하는 것이다. 사전 기반 오피니언 마이닝은 문장을 구성하는 단어들의 극성 정보를 색인한 감성 사전을 이용하며, 사전의 품질이 감성 분류 성능에 큰 영향을 끼치므로 사전을 어떻게 구축할 것인지가 중요한 이슈가 된다. 영어 데이터일 경우 명사, 동사, 형용사, 부사로 분류된 WordNet 어휘에 긍정, 부정, 중립 값(sentiment score)을 부여한 SentiWordNet (Baccianella, Esuli, & Sebastiani 2010)을 이용하여 감성 극성을 판별하는 연구들이 이루어지고 있다 (Dang, Zhang & Chen 2010; Thet, Na & Khoo 2010; Singh 등 2013). Taboada 등(2011)도 감성사전을 구축하여 상품평, 영화평, 도서평 등 다양한 의견의 감성 분석을 시도하였다.

한글은 영어에 비해 문장 구조 분석, 형태소 분석 등이 복잡하고, 특히 온라인 상에서 사용되는 한글은 변형이 굉장히 다양해서 전처리에 많은 노력과 시간이 소요된다. 감성 분류를 위해서 감성 사전을 구축하다고 하더라도 새로운 용어가 생성되는 주기가 짧기 때문에 수작업으로 감성 사전을 계속 업데이트 하는 것이 쉽지 않다. 때문에, 기존에 구축된 한글 감성 사전을 기반으로 감성 분류 성능을 높일 수 있는 확장된 감성 사전 구축의 필요성이 커지고 있다.

본 연구는 용어 벡터기반 감성 분류 모델과 사전기반 감성 분류 모델을 사용하여 자동으로 감성 사전을 확장하고자 한다. 실험적 연구를 통해 가능성과 동시에 개선점을 찾아내고자 한다.

* 이 연구는 2012년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012S1A3A2033291)

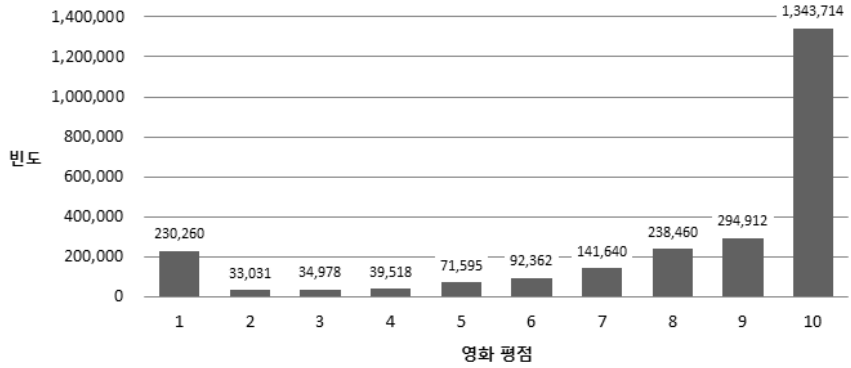
Ⅱ. 연구 방법

1. 실험설계

1.1 데이터 수집

본 연구에서는 감성 사전을 자동 확장하기 위한 실험 데이터로 감성어가 상대적으로 많이 포함된 네이버 영화평을 수집하여 사용하였다. 총 1219건의 영화에 대한 ID를 수집한 후, 각 영화에 대한 평을 수집하여 총 256만개의 리뷰 데이터를 수집하였으며, 데이터가 생성된 시기는 2003년11월 21일부터 2015년10월16일까지로 구성되어있다.

수집된 리뷰의 영화 평점 분포는 <그림 1>과 같이 1점과 10점인 것이 많았고, 특히 10점을 받은 영화평이 많은 비율을 차지하고 있다. 이러한 현상은 영화 평점이 가진 일반적인 현상으로, 영화 평점 결과는 정규 분포와 유사하지 않은 형태를 가지고 있다(김경민 외 2014).



〈그림 1〉 영화 평점 분포

수집된 데이터의 메타데이터는 리뷰ID, 평점, 영화ID, 영화 명, 이용자ID, 리뷰, 작성날짜로 구성되어있다. <그림 2>는 영화 “드래곤 길들이기 2”에 대한 수집된 리뷰 데이터 중 일부로, 한글 긍정어휘인 “재미있다”의 다양한 어휘변형 행태를 파악 할 수 있다. 표준어휘를 기반으로 한 감성 사전인 경우에는, 감성어의 다양한 변형 형태를 반영할 수 없기 때문에 이런 행태를 반영할 수 있는 감성 사전 확장 방법이 필요하다.

리뷰ID	평점	영화ID	영화 명	이용자ID	리뷰	작성날짜
9090741	10	76020	드래곤 길들이기 2	jiju****	넘잼났어요~~짱짱짱짱	14.08.15
9089846	10	76020	드래곤 길들이기 2	kss6****	아이들과 같이봤는데제가봐도재밌네요재밌게봤습니다	14.08.15
9089389	9	76020	드래곤 길들이기 2	yoon****	재미재미잼잼이네ㅋㅋ	14.08.15
9089451	10	76020	드래곤 길들이기 2	cksg****	진짜 한번만 보셈 개꿀잼	14.08.15
9072929	8	76020	드래곤 길들이기 2	jjun****	초등4학년 아이들이 보았는데..잼났다고 하네요.	14.08.12
9072981	10	76020	드래곤 길들이기 2	cjh9****	완전 꿀잼이였어요 두번봐도 재밌 ㅋㅋ	14.08.12
9072948	10	76020	드래곤 길들이기 2	eunb****	지루하지않고 재밌었음!	14.08.12
9073389	9	76020	드래곤 길들이기 2	choi****	투슬리스~!개개인이 보는 시각이 다르겠지만 저는 재밌게봤습니당	14.08.12
9072162	9	76020	드래곤 길들이기 2	poop****	하늘을 날고싶다는생각이들었ㅆ다 재밌고 머찌더라	14.08.12
9071476	10	76020	드래곤 길들이기 2	sod_****	완전 재미짐 퀄리티가 남다름	14.08.12
9071368	10	76020	드래곤 길들이기 2	yocc****	내용도 알차고 볼거리도 많고 5살 아들도 잼나했지만 제가 더 잼나게 봤네요~^^	14.08.12

<그림 2> “재미있다”의 어휘 변형 예시

1.2 실험설계

한글 감성 분류를 위한 감성 사전의 자동 확장 구축을 위해 <그림 3>과 같이 실험설계를 하였다. 수집된 데이터 중 긍정 리뷰(평점 10점) 1만 5천개와 부정 리뷰(평점 1점) 1만 5천개를 학습데이터로 사용하였다. 감성 분류 모델을 구축하기 위해 문장 분리, 형태소 분석, 불용어 제거 및 원형 복원 등의 한글 데이터 전처리 과정을 거쳤다. 한글 형태소 분석을 위해서 Komoran 형태소 분석기(http://shineware.co.kr)를 사용하였으며, 감성 사전 자동 확장 구축을 위해 자바 기반의 파이프라인 시스템을 구축하여 실험하였다.

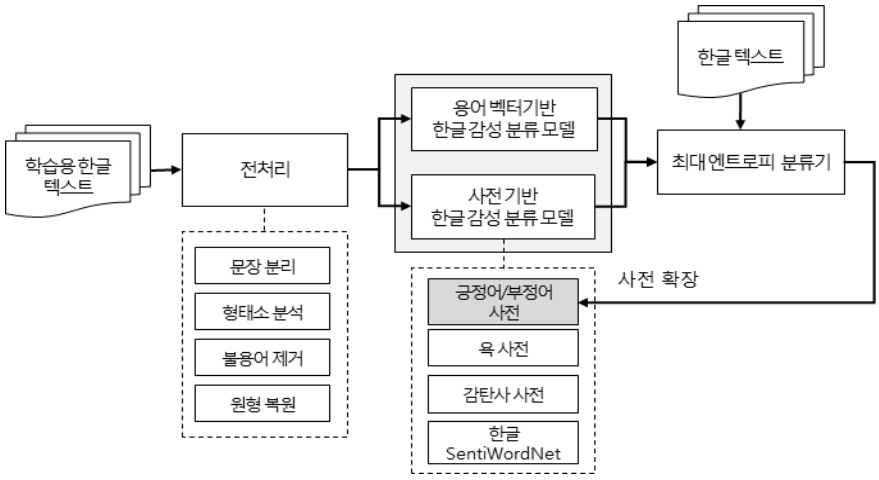
다음 2가지 방식으로 한글 감성 분류 모델을 생성하였다.

① 감성 단어의 출현빈도를 기반으로 한 용어 벡터(Bag of Word)기반 모델

② 감성 사전 기반 모델: 감성 사전은 긍정어와 부정어로 구성된 감성어 사전 뿐 아니라, 욕 사전, 감탄사 사전, 한글 SentiWordNet으로 구성하였다. 각 사전에서 문장(s)에 대한 감성 점수(S)를 계산하여 4개 사전의 감성 점수를 합산하는 방식으로 감성 점수를 계산하여 극성을 판별하였다.

$$S(s) = (D_{욕}(s) + D_{감탄사}(s) + D_{KSW}(s) + D_{P/N}(s))$$

감성 분류 모델을 생성 한 후, 최대 엔트로피 분류기를 이용하여 한글 텍스트의 감성 극성을 분류하였다. 최대 엔트로피 분류기는 사전에 알려진 정보를 기반으로 엔트로피가 최대가 되는 확률분포를 최적으로 선정하는 원리를 기반으로 확률모델을 만들어 데이터를 분류하는 분류기이다.



〈그림 3〉 감성 사전 자동 확장 실험 설계

감성 사전의 자동 확장은 새로운 한글 텍스트를 대상으로 감성 극성을 판별한 후, 감성 사전에 등록되어 있지 않은 긍정어와 부정어를 사전에 자동 추가하는 방식으로 진행하였다. 이러한 방법을 통해 새롭게 생성되었거나, 변형된 한글 감성어 텍스트가 사전에 추가될 수 있다.

확장된 사전을 반영하여 이미 생성되어 있는 사전기반의 분류 모델을 보강하는 과정을 거쳤으며, 새로운 데이터가 계속 입력될 경우 반복적으로 사전을 보강할 수 있는 구조로 이루어져 있다. 김승우, 김남규 (2014)도 영화 리뷰 데이터를 이용하여 감성사전을 자동으로 구축하려는 시도를 했으나 이들은 영어 데이터를 대상으로 했고, 한글 데이터를 대상으로 한 연구는 없다.

Ⅲ. 연구 결과

실험 성능 검증을 위해 10-fold 교차검증을 실시하였으며, 성능평가를 위해서 전체 분류 중 옳은 분류의 비율을 나타내는 정확도(Accuracy) 척도를 이용하였다. 정확도를 구하는 공식은 아래와 같다.

〈표 1〉 성능평가를 위한 2*2 테이블

방법	옳은 범주	틀린 범주
범주에 분류	a	b
범주에 비분류	c	d

정확도 (Accuracy) = *(a+d) / (a+b+c+d)*

베이스라인 실험은 기존에 구축된 4가지 감성 사전을 활용한 사전 기반 감성 분류 모델을 이용한 것이며, 자동 확장 감성어 사전 기반 분류 모델 이외에 감성 사전 중 한글 SentiWordNet 사전을 제외한 실험 성능도 평가하였다.

〈표 2〉 성능

방법	정확도
사전 기반 (4가지 사전)	0.5122 +/- 0.0058
감성어 사전 기반	0.5067 +/- 0.0049
자동 확장 감성어 사전 기반	0.6677 +/- 0.0218

한글 SentiWordNet 사전을 제외하고 실험한 결과 성능이 조금 떨어지긴 했지만, 감성 분류 시간이 훨씬 단축되었다. 중립 단어를 많이 포함하고 있는 SentiWordNet은 극성 감성 분류 성과에 큰 영향을 주지 않는 것으로 분석하였다. 조상현, 강행봉(2011)의 연구에서는 영화 분야 댓글 감정 분류 성능이 다른 분야(여행, 인물, 음식)에 비해 낮다고 보고하면서, 이유로 다른 분야에 비해 우회적으로 감정을 표현하는 것이 상대적으로 많기 때문이라고 분석하였다. 이런 평가결과 및 선행연구의 분석을 바탕으로 본 연구에서는 단어의 감성 극성을 효과적으로 학습하기 위해 영화 평점 1점과 10점인 리뷰만을 학습 데이터로 이용하였다.

Ⅳ. 결론 및 제언

사전 기반 감성 분류 성능에 비해 자동으로 확장한 감성사전의 성능은 정확도에서 약 16% 향상을 보였다. 하지만 성능 자체로 보면 이전 연구(조상현, 강행봉 2011; 이철성 외 2013)보다 낮은데 그 이유를 살펴보면 선행 연구보다 더 많은 데이터를 대상으로 실험했기 때문이다. 기존의 소셜 데이터를 사용한 감성 분류의 정확도 성능 또한 50%~60%대의 성능을 보이는 것으로 나타났다(홍소라 등 2014). 소셜 빅데이터를 처리하기 위한 기법들의 경우, 대량의 학습데이터 확보가 어렵기 때문에 한정된 데이터를 대상으로 실험된 기존의 모델보다 소셜 빅데이터를 대상으로 한 모델의 정확도가 상대적으로 떨어지게 나타난다.

감성 분류 성능을 더 높이기 위해서 사전 확장을 위해 자동으로 추가되는 단어들의 가중치를 계산할 필요가 있다. 또한 영화평은 일반 소셜미디어 데이터보다는 감성 어휘가 많이 포함되어 있지만 영화에 대한 긍정·부정 리뷰는 배우, 감독, 영화 내용 자체, 상영시간, 캐릭터에 대한 평가 등을 포함하고 있으므로 감성어휘가 포함되지 않은 “악법도 법인가”와 같은 영화평은 평점이 10점이라 하더라도 감성 사전에 단어를 추가하지 않는 방법을 고려해야 한다.

본 연구에서 제안한 방법은 소셜 데이터, 웹에서의 한글 데이터 처리를 간단하게 하면서 사전을 자동적으로 확장할 수 있는 방법으로서의 의의를 가지며, 온라인 상에서 사용되는 언어 행태를 반영하기 쉽다는 장점이 있고, 향후 악의적 용어 판별을 위한 기초가 되는 부정어 사전으로 활용될 수 있다. 추후 긍정·부정뿐만 아니라 좀 더 다양한 감성 범주를 대상으로 한 연구를 진행할 것이다.

참고문헌

Baccianella, S., Esuli, A., & Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

Dang, Y., Zhang, Y., & Chen, H. 2010. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *Intelligent Systems, IEEE*, 25(4): 46-53.

Singh, V. K., Piryani, R., Uddin, A., & Waila, P. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In Automation, Computing, Communication, Control and Compressed Sensing(iMac4s), 2013 International Multi-Conference on (pp. 712-717). IEEE.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267-307.

Thet, T. T., Na, J. C., & Khoo, C. S.. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6): 823-848.

김경민, 안무혁, 이윤호. 2014. 영화 평점에서 악의적 평점제공자 판별 및 실평점 유추.『정보과학회논문지: 컴퓨팅의 실제 및 레터』, 20(4): 213-218.

김승우, 김남규. 2014. 오피니언 분류의 감성사전 활용효과에 대한 연구.『지능정보연구』, 20(1): 133-148.

이철성, 최동희, 김성순, 강재우. 2013. 한글 마이크로 블로그 텍스트의 감정 분류 및 분석.『정보과학회논문지:데이타베이스』, 40(3): 159-167.

조상현, 강행봉. 2011. 형식적 및 비형식적 어휘 정보를 반영한 문장 감정 분류.『정보처리학회논문지 B』, 18(5): 325-332.

홍소라, 정연오, 이지형. 2014. 대용량 소셜 미디어 감성분석을 위한 반감독 학습 기법.『한국지능시스템학회 논문지』, 24(5): 482-488.