SentiWordNet을 활용한 기계학습 기반 한글 감성사전 구축

Construction of Korean sentiment dictionary based on machine learning and SentiWordNet

저자 (Authors)
박성홍, 이동기, 신현정
Sunghong Park, Dong-gi Lee, Hyunjung Shin

출처 (Source)
한국정보과학회 학술발표논문집 , 2018.6, 634-636(3 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07503100

APA Style
박성홍, 이동기, 신현정 (2018). SentiWordNet을 활용한 기계학습 기반 한글 감성사전 구축. 한국정보과학회 학술발표논문집, 634-636

SentiWordNet을 활용한 기계학습 기반 한글 감성사전 구축

박성홍O, 이동기, 신현정*
아주대학교 산업공학과
{pshong513, ldg1226, shin}@ajou.ac.kr

Construction of Korean sentiment dictionary based on machine learning and SentiWordNet

Sunghong ParkO, Dong-gi Lee, Hyunjung Shin*
Department of Industrial Engineering, Ajou University

요 약

감성분석에서 가장 중요한 부분은 감성사전의 구축이다. 감성사전에 포함될 단어를 결정하는 방법으로는, 단어의 의미에 기반하여 전문가가 직접 감성을 부여하는 정성적 방법이 있다. 이 접근방법은 주어진 조건에 적합하게 단어의 감성을 파악할 수 있지만 상당한 시간과 비용을 소요한다. 한편, 단어의 동시 출현 빈도에 기반한 정량적 방법이 있다. 이 방법은 효율적이기는 하나 이미 정해진 감성단어에만 국한되어 환경에 따라 단어를 확장하지 못하는 한계점이 있다. 본 연구에서는 주어진 도메인 및 환경에 따라 확장이 가능한 감성사전 구축방법을 제안한다. 감성단어가 풍부한 영어사전인 SentiWordNet을 번역하여 한글 감성단어 네트워크를 구축한다. 그 다음, 기계학습 알고리즘을 적용하여 단어의 극성(긍정/부정)을 보다 많은 단어로 확장시킨다. 제안하는 감성사전을 자동차 관련 웹 블로그 감성분류에 적용했을 때, 0.941 AUC의 우수한 성과를 얻을 수 있었다.

1. 서 론

감성분석은 사람들의 의견이나 감정과 같은 주관성을 분석하는 연구분야로써, 분석 기준이 되는 감성사전 구축이 가장 중요시된다[1]. 감성사전은 단어의 감성을 극성으로 나타낸 어휘 집단을 의미하며 단어의 극성을 긍정과 부정으로 이분화하는 방법이 기본이 된다.

대표적인 감성사전 연구로써 영어를 기반으로 구축된 SentiWordNet(SWN)[2]을 들 수 있다. 약 12만개의 단어들 사이의 유사도를 네트워크로 나타낸 WordNet를 활용했으며, 의미를 기반으로 동의어·반의어나 접두사·접미사 등을 활용해 단어의 극성을 확장했다. SWN은 다양한 감성 분석 연구의 기반이 되어 300개 이상의 연구 그룹에서 활용되고 있다. 이외에도 LIWC[3] 등 다양한 영어 감성사전들이 존재한다. 한편, 한국어를 기반으로 감성사전을 구축한 연구도 있다. 대표적 연구로는 한국어 감성분석 코퍼스 Korean Sentiment Analysis Corpus (KOSAC)[4]를 들 수 있다. KOSAC은 3명의 연구자가 7,774개의 신문 기사 문장을 선정하여 17,582개의 감정 표현을 주석한 한국어 감성사전이다. 이 방법은 언어학 전문가의 판단과 검증을 통해 감성사전을 구축하므로 단어의 극성에 대한 정확성은 높을 수 있으나, 필터링과 반복적으로 이루어져야 하는 수작업에 상당한 시간과 비용이 소모된다.

이러한 정성적 방법을 보완하는 방법으로서, 텍스트로부터 얻을 수 있는 정량적 특성을 활용하는 연구가 있다. 대체로 단어의 동시 출현 빈도나 유사도, 상관관계 등을 활용하며, 딥러닝 기반 word2vec이나 doc2vec 등과 같은 기계학습 알고리즘을 활용한다. 이러한 연구들은 단어의 극성이 표출된 형태, 즉 점수화된 평점 리뷰의 형태를 띤다. 주로 영화평 분석이나 상품평 분석 도메인에 편중되어 있다. 이러한 연구들에서는 단어의 극성을 평점에서 쉽게 얻을 수 있으므로 감성분석에서는 다소 용이한 분석이라 할 수 있다. 하지만 평점이 주어지지 않는 대부분의 도메인에서는 정량적 기법을 활용한다 하더라도 감성사전의 생성이 쉽지만은 않다.

본 연구에서는 정량적 방법을 활용하여 보다 많은 감성단어가 수록된 감성사전을 구축하고자 한다. 이를 위하여 기존 연구인 KOSAC 보다 단어수가 풍부한 SWN을 기반으로 단어 네트워크를 구축한다. 한편, 단어 네트워크는 도메인에서만 알 수 있는 감성단어의 특수성을 반영할 수 있어야 한다. 도메인에 따라, 일반적 감성단어가 아닌 단어가 감성을 표현하는 데 쓰여질 수 있다. 예를 들어, '잘 터진다'라는 표현은 통신 분야에서는 긍정 감성표현이 된다. 심지어는 단어의 극성이 바뀌기도 한다. 종종 '죽여 준다'는 '매우 좋다'는 의미로 긍정 감성의 반어적 표현이다. 이러한 사항들을 감성분석에 반영하기 위해서는 감성단어 네트워크를 도메인에 맞게 적응시키는 과정이 고려되어야 한다. 또한 감성분석이 잘 이루어지기 위해서는 감성을 갖는 단어를 증가시키는 것이 좋다. 이는 기존의 감성단어의 극성을 단어간 유사성으로 표현된 네트워크를 통해 전파시켜서 확장시키는 방향으로 진행되면 좋을 것이다. 이를 위하여 본 연구에서는 그래프 기반 준지도학습을 적용한다. 제안하는 방법은 전문가의 개입을 필수로 하지 않으므로 효율적이다.

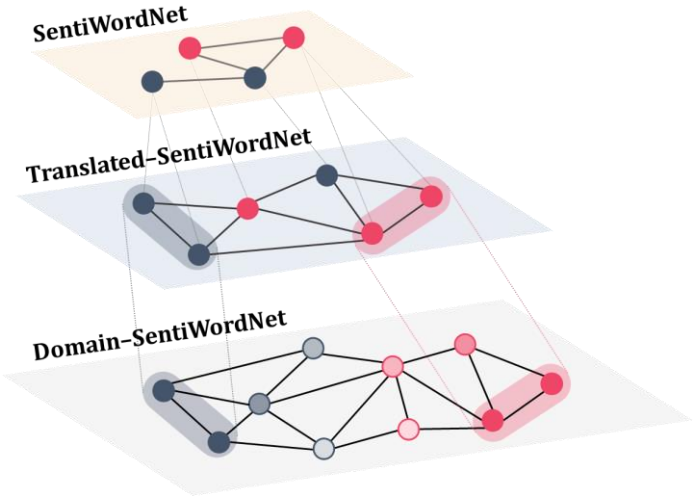


그림 1. 감성단어 네트워크의 확장

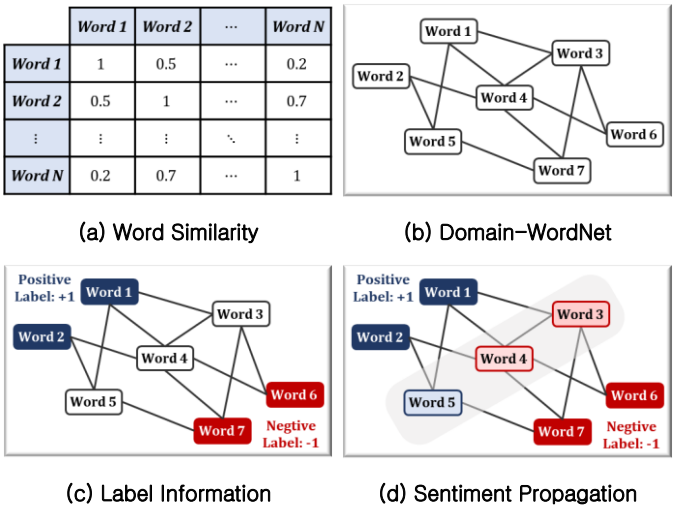


그림 2. 도메인 별 특수화와 유연화 과정

또한 도메인에 맞는 확장성과 유연성이 있다. 따라서 평점 등과 같은 확실한 정보가 텍스트로부터 주어지지 않는 경우라 할 지라도 감성사전을 쉽게 구축할 수 있다.

다음의 각 절에서는 제안 방법을 구체적으로 설명한다. 실험에서는 각 단계별 결과 및 감성분류 정확도를 보인다.

2. 제안 방법론

제안하는 방법론의 구성은 다음과 같다. 첫째, SWN을 한글로 번역하여 Translated-SentiWordNet(T-SWN)을 만든다. 둘째, 도메인에 특화된 Domain-SentiWordNet(D-SWN)을 구축한다: 이 과정은 단어의 확장과정과 극성의 전파과정으로 나뉜다. <그림 1>은 제안 방법론의 개요도를 나타낸다.

2.1. 영어사전을 번역한 감성단어 네트워크: T-SWN

SWN은 영어 단어를 노드로, 단어 의미간 유사도를 엣지로 구성한 그래프로 표현된다. T-SWN은 Cambridge 영한사전[5]을 활용한 SWN의 한글 번역 내용으로 구성한다. 따라서, T-SWN의 노드는 번역한 한글 단어들이 된다. 엣지 구성은 SWN과 유사한 방식으로 번역 시 사전 상에서 각 한글 단어 의미에 대한 영어 표현을 활용한다. WordNet을 활용하여 T-SWN 노드들의 영어 표현 간의 유사도를 도출하여 엣지로 구성한다. T-SWN을 구성한 후, 그래프 기반 준지도학습 알고리즘을 활용하여 감성을 확장한다. 레이블 정보는 SWN 데이터를 활용하며 +1/−1은 각각 긍정/부정을 나타낸다.

$$\min_f (f-y)^T(f-y)+\mu f^T L f \quad (1)$$

$$y=(y_1,\cdots,y_l,0,\cdots,0)^T,\ y_l\in\{-1,+1\}$$

$$L=D-W, D=diag(d_i), d_i=\sum_j w_{ij}$$

$$f=(I+\mu L)^{-1}y \quad (2)$$

2.2. 도메인 특수성을 반영한 감성단어 네트워크: D-SWN

T-SWN을 구성한 후, Domain-SentiWordNet (D-SWN)을 구성하고 감성을 확장한다. 웹 블로그의 텍스트에 대한 단어 그래프를 구성하고 T-SWN을 활용하여 감성 레이블을 확장한다. D-SWN 구성은 4단계로 진행되며 <그림 2>는 각 단계를 나타낸다. 첫째, 단어 사이의 유사도를 도출한다. word2vec 알고리즘[6]을 활용하여 수집한 웹 블로그 텍스트를 단어 벡터로 정량화 한다. <그림 3>은 텍스트 정량화 과정을 나타낸다. 이후, 식(3)의 Gaussian 함수를 이용하여 <그림 2(a)>와 같은 단어간 유사도를 도출한다.

$$w_{ij}=\begin{cases} exp(-\|x_i-x_j\|^2/\sigma^2) & if\ i\sim j \\ 0 & otherwise \end{cases} \quad (3)$$

둘째, 도메인 단어 네트워크(Domain-WordNet)을 구성한다. 이 때, 노드는 도메인에서 사용되는 단어들로, 엣지는 식(3)과 k-Nearest Neighbor (k-NN) 방법에 의한 단어간 유사도로 구성된다. 셋째, T-SWN의 단어 감성 레이블 정보를 적용한다. T-SWN 상의 단어가 도메인 단어 네트워크에 존재한다면 해당 단어에 레이블을 부여한다. 최종적으로 식(1)과 식(2)를 통해 도메인 단어 네트워크 전체로 감성분류를 수행한다.

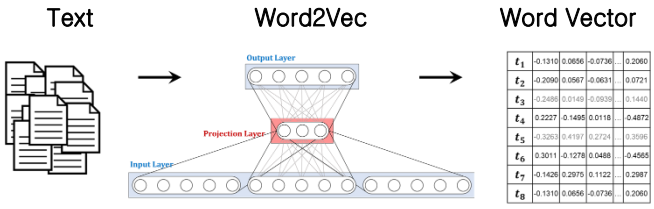


그림 3. Word2Vec을 활용한 텍스트 정량화

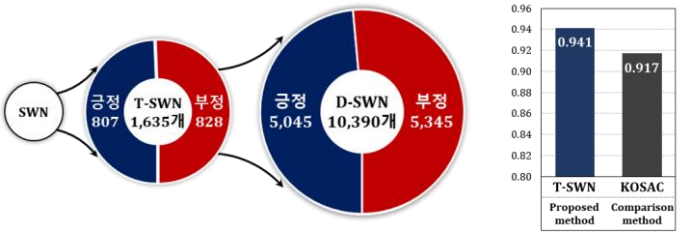
3. 실험

3.1. 데이터

T-SWN 구성을 위해, SWN 데이터를 수집하였다. 전체 117,659개로 구성된 SWN의 단어 중 감성이 명확한 단어를 추출하여 전체의 3.22%에 해당하는 3,787개의 단어를 실험에 활용하였다. 웹 블로그 텍스트의 도메인은 자동차 리뷰로 선정하였다. 국내 대표 자동차 회사의 판매 차종 5개에 대해 차종별로 네이버 대표 카페 2개씩, 총 10개의 카페에 대한 웹크롤링을 통해 311,550건의 텍스트를 수집하였다.

3.2. 영어사전으로부터의 감성단어 확장 결과: T-SWN

T-SWN을 구성하기 위해, Cambridge 영한사전을 이용한 번역을 수행하였다. SWN으로부터 추출된 단어 3,787개를 번역하고 중복된 의미를 제거하여 총 1,635개의 단어로 T-SWN의 노드를 구성하였다. 엣지는 식(4)을 통해 영어 단어간 유사도를 한글단어 유사도 계산에도 반영하였다. 이후, 식(1)과 식(2)를 통해 SWN에서의 단어극성 레이블을 확장함으로써 T-SWN 노드들의 감성을 분류하였다. 그 결과, 전체 1,635개의 단어로부터 긍정 감성 단어 807개(49%)와 부정 감성 단어 828개(51%)를 도출하였다.



(a) 감성단어 네트워크 확장 결과 (b) 감성분류 정확도

그림 4. 감성분류 실험 결과

3.2. 도메인 특수성을 반영한 감성단어 확장: D-SWN

자동차 도메인에서 수집된 웹 블로그 텍스트는 word2vec을 통해 62,486개의 단어를 벡터로 정량화되었다. 이 중 출현빈도가 100회 이상인 단어 10,390개를 선별하고 유사도를 계산하여 k-NN 방법 (k=20)으로 0.38% 밀도의 도메인 단어 네트워크를 구성하였다. 구성된 네트워크 상의 단어들 중에서, T-SWN에 존재하는 단어는 총 1,110개로, 이를 감성 레이블(긍정: 539개, 부정: 571개)로 활용하였다.

제안하는 방법의 성능을 평가하기 위해, 도메인 단어 네트워크와 감성 레이블을 준지도학습 알고리즘에 적용하였다. 준지도학습의 파라미터인 μ는 100으로, 검증 방법으로는 10-fold cross validation으로 100회 반복 수행하였다. 그 결과, 단어 감성분류 정확도는 0.941 AUC를 보였다. 최종적으로 전체 도메인 단어 10,390개에 알고리즘을 적용한 결과 5,045개를 긍정으로, 5,345개를 부정으로 분류하였다. <그림 4(a)>는 SWN으로부터 T-SWN을 거쳐 D-SWN으로 감성단어 네트워크가 확장된 결과를 나타낸다.

또한 제안하는 한글 감성사전 구축 방법과 기존 감성사전인 KOSAC과의 성능 비교를 수행하였다. KOSAC에 수록되어 있는 전체 17,582개의 단어 중 1,570개를 도메인 단어 그래프에 대한 감성 레이블로 사용하였다. 실험 설정은 제안 방법론과 동일하게 수행하였으며, 그 결과 KOSAC의 단어 감성분류 정확도는 0.917 AUC를 보였다 (<그림 4(b)> 참조). 제안한 네트워크에서 사용한 단어수가 KOSAC 단어수의 약 1/5 배임을 감안하면 상당히 우수한 결과라 할 수 있다. 제안하는 방법인 T-SWN이 KOSAC보다 더 높은 정확성을 보이는 것은 두 가지 요인에 기인한 것으로 볼 수 있다. 첫째, 감성사전 구축 단계에서 단어 간의 의미 관계성 반영 여부에 기인한다. T-SWN은 네트워크 구조를 활용함으로써 감성단어들이 서로 연결되어 의미론적 유사도가 도출되는 구조이다. 이에 반해 KOSAC은 개개 단어의 감성 의미로만 이루어졌다. 둘째, 감성단어의 확장성에 기인한다. T-SWN는 감성단어가 풍부하고 도메인 용어까지 감성단어에 포함시킬 수 있는 유연한 구조이다. 이는 KOSAC에 비해 제안한 방법이 갖는 장점이라 할 수 있다.

4. 결 론

본 연구에서는 SWN의 번역과 기계학습을 활용한 한글 감성사전 구축 방법을 제안한다. SWN을 기반으로 하여 기존 한글 감성사전보다 단어수가 훨씬 풍부한 한글 감성단어 네트워크인 T-SWN을 구축했다. 또한 웹 블로그 텍스트에 적용하여 도메인 감성단어 네트워크인 D-SWN을 구축했고, 그래프 기반 준지도학습을 활용하여 도메인 특성에 맞게 단어의 극성이 전파되도록 했다. 제안하는 방법은 단어 감성분류에 있어 전문가의 개입을 필수로 하지 않으므로 효율적이다. 또한 도메인 특성에 맞게 감성단어에 대한 확장성이 좋기 때문에 도메인별 감성사전을 쉽게 구축할 수 있다. 단어 감성분류 정확도 역시 기존 감성사전 KOSAC보다 제안 방법인 T-SWN이 더 우수했다.

본 연구는 실험환경의 제한상, 전체 단어가 아닌 일부 단어들만을 사용한 프로토타입 네트워크로 이루어졌다. 추후 전체 단어를 포괄하는 네트워크로 확장되면 현 연구에서 보다 월등히 우수한 감성 분류 성능을 얻을 수 있으리라 기대한다.

참고문헌

[1] Chen, H. & D. Zimbra,. (2010). AI and opinion mining. IEEE Intelligent Systems, Vol.25, No.3, 74-80.

[2] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the 7th Conference on International Language Resources and Evaluation(LREC '10), 2200-2204.

[3] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29, 24-54.

[4] Jang, H., Kim, M. & Shin, H. (2013). KOSAC: A Full-fledged Korean Sentiment Analysis Corpus. In: Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation, 366 – 373.

[5] Cambridge Dictionary, https://dictionary.cambridge.org/

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).