

패턴을 이용한 상품평 감정 단어 추출 방법

천은혜^o, 심수정, 박혁로
전남대학교, 전남대학교, 전남대학교
mylife0905@naver.com, sjsimox@nate.com, hyukro@chonam.ac.kr

Sentiment words extraction method using pattern

Eun-Hye Chun^o, Su-Jeong Shim, Hyuk-ro Park

요약

최근 오피니언 마이닝 관련 연구 중 감정 분류에 대한 관심이 높아지면서 많은 연구가 진행되고 있다. 기존 영어권 연구에서 제시되어온 방법은 한국어 상품평에 적용하는 것이 쉽지 않다. 영어 시소러스 기반 한국어 감정단어 추출 기술은 한국어와 영어 단어가 일대일로 일치하기가 어렵다는 문제가 있다. 기존 관련 연구 중 k-Structure 기법은 패턴의 길이가 3인 단순한 문장에 속성단어와 감정단어가 포함되었을 경우를 기준으로 한 것이므로 한정적이다. 본 논문에서 제안하는 방법은 상품평에서 의미적인 패턴을 추출하여 감정 단어의 위치를 파악하는 방법이다.

주제어 : 오피니언 마이닝, 속성 단어, 감정 단어, 문장 패턴

1. 서론

본 논문은 오피니언 마이닝 관련 연구 중 한국어 상품평으로부터 패턴을 이용하여 의미적으로 해석하는 방법을 제안하고 있다. 한국어 문서에 대한 연구 중에는 수작업 분석을 통해 감정 단어를 찾는 방법[2]와 한국어 구문 분석기를 기반으로 감정 단어를 찾는 연구도 있다[3].

[5]에서 제안하고 있는 k-Structure기법은 한국어 상품평을 분석했을 때 감정어가 포함되어 있을 확률이 높은 문장의 구조를 정리한 것이다. 각 문장 구조에는 형태소 분석기를 통해 찾아낸 품사 정보가 포함되어 있다. 이 방법은 정확률이 높지만 속성단어와 감정 단어가 패턴 길이 최대 3인 단순한 문장에 포함되어 있을 경우를 전제 하므로 한정적이다. 본 논문에서는 수집한 상품평들을 이용하여 관련된 속성단어를 찾고 그 속성단어에 관련된 의미적인 패턴을 만들어 감정 단어의 위치를 찾으므로서 감정 분류를 하기 위해 필요한 감정 단어를 찾는 방법을 제안하고 있다.

2. 관련연구

수작업 분석을 통해 감정 단어를 찾는 방법[2]는 감정 단어를 추출하는 데 있어서 높은 정확도를 요구할 때 사용할 수 있다. 수작업으로 분석 과정을 거쳐 감정 단어를 찾는 방법으로 정확성은 향상될 수 있지만 감정 단어를 찾는 시간이 오래 걸린다. [3]은 한국어 구문 분석기를 이용한 방법으로 출현 빈도가 낮은 감정 단어를 선정하지 못하는 경우가 있을 수 있다.

이외에도 영어 시소러스를 기반으로 감정 단어를 추출하는 [4]는 한국어와 영어 단어가 일대일로 일치하지 않는데서 오는 정확도 저하의 문제가 올 수 있다. [3,4]는 자동적인 방법으로 감정단어를 추출하지만 [2]보다 정확도가 낮을 수 있다.

오피니언 마이닝 관련 연구에는 크게 문서 단위와 속성 단위로 처리하는 방법이 있다. 감정단어 추출 관련 연구는

주로 속성 단위 오피니언 마이닝 연구에서 이루어지고 있다.

2.1 문서 단위 오피니언 마이닝

[6]은 PMI-IR 기법을 이용하여 문서 전체를 대상으로 긍정 또는 부정으로 분류하였다. PMI-IR은 특정한 문장의 패턴을 만족하는 여러 개 구문들의 Semantic Orientation(SO)을 계산한다. 그런 후에 각 구문에 대한 SO의 총합이 양수 일 경우 긍정으로, 음수 일 경우 부정으로 분류하는 방법이다.

[7-10]는 전통적인 주제 기반 문서 분류의 개념을 오피니언 문서에 적용하여 긍정 또는 부정으로 분류하였다. 이 중 [7]은 Score Function 계산식을 이용하여 확률적으로 긍정 또는 부정을 결정한다. [8-11]에서는 기계학습 알고리즘을 적용하여 문서의 긍정 또는 부정을 결정한다. [11]에서는 자체적으로 정의 한 계산식을 이용하여 Document Sentiment Value(DSV)를 산출한다. 그런 후에 DSV의 값에 따라 긍정 또는 부정으로 분류하였다.

2.2 속성 단위 오피니언 마이닝과 감정단어 추출에 관한 연구

[12]에서는 명사를 속성단어라고 가정하고, 속성단어에 인접한 형용사를 감정단어라고 판단한다. 실제로는 속성단어와 인접한 동사/명사가 감정단어 일 수 있음에도 불구하고 이를 감정단어로 판단하지 못하는 문제를 갖는다. 이에 따라 감정단어 추출의 재현율이 높지 못 할 수 있다.

[13]은 영문 Word Net을 이용하여 감정단어를 추출한다. 초기에 다수의 감어를 Seed Word로 주고, 각 Seed Word에 대한 유의어와 반의어를 확장하여 감정단어 사전을 구축한다. 그러나 Seed Word에 대한 유의어/반의어 확장 시 감정단어에 해당되지 않는 단어가 추출되는 경우가 있다. 이렇게 해서 수집한 감정단어 집합 역시 정확률(precision)이 떨어지는 문제가 있을 수 있다.

[2]에서는 수작업 분석을 통해 감정단어를 찾아낸다. 현재까지 10개 카테고리에 대해서 약 9,000개 정도의

감정단어를 구축하였다. 그러나 감정단어를 찾는데 시간이 많이 걸리므로 모든 분야의 감정단어를 찾아내는 데에는 많은 시간이 소요될 것으로 예상된다.

[3]에서는 한국어 구문 분석기를 통해 상품평을 분석하였다. 구문 분석 결과 서술어(Predicate)가 감정단어가 될 수 있다. 이를 중 문서 전체에서의 출현 빈도수가 임계치 이상이 되는 것을 감정단어로 채택하는 방법을 제안했다. 그러나 실제로는 임의의 단어가 임계치 이하의 값을 가지더라도 감정단어인 경우가 종종 있을 수 있어서 이런 단어는 감정단어로 선정이 되지 않을 수 있다.

3. 방법

본 논문에서는 많은 상품평 중 3개의 카테고리를 정하였고 그 중 ‘카메라’에 대한 예를 [그림 1]에서 보여주고 있다. 상품평에서 반복적으로 등장하는 속성단어들을 추출하고 반복적으로 등장하는 문장의 패턴을 추상화하여 [그림 1]처럼 속성단어가 포함된 문장의 패턴을 만들었다.

Pattern types:
배송 Verb
가격 디자인 Verb
디자인 Verb Verb
Verb 가격 Verb 디자인
Verb 배송
Verb 가격 디자인
Verb Verb 디자인
가격 Verb 디자인 Verb
the sentiment words:
배송(택배) : 빠르다, 느리다, 불친절, 보통, 오래 걸리다
가격 : 고가, 싸다, 착하다, 최저가, 비싸다, 저렴
디자인(모양, 외관) : 멋진, 깔끔하다, 예쁘다, 귀엽다, 아담, 세련, 깔끔하다, 안 예쁘다, 심플하다

[그림 1]. 카메라에 대한 패턴 추출 결과

추출한 속성단어들은 제품, 배송, 기능, 가격, 색, 화질, 사이즈, 디자인이 있으며 그 중 배송, 가격, 디자인에 대한 문장 패턴과 속성단어를 [그림 1]에서 나타내고 있다. 아래의 [그림 1]의 Pattern type에서 Verb는 속성 단어와 관련된 감정단어가 오게 된다. Verb의 위치에 오는 감정단어로는 품사와 상관 없이 동사, 형용사, 명사 모두 해당되며 [그림 1] 안의 sentiment words는 ‘카메라’의 속성단어 중 배송, 가격, 디자인에 대한 감정단어들을 추출한 것이다.

이렇게 정의한 문장의 패턴을 이용하여 Verb의 위치에 오는 단어들을 감정단어들로 추출할 수 있으며 다른 상품평에 적용하게 되면 자동으로 의미적인 감정 단어를 추출하게 된다.

5. 결론

본 논문에서는 기존 연구의 어려움을 극복 하고자 상품평들의 반복되는 문장의 패턴을 찾아 정의하였고 정의한 패턴을 이용하여 자동으로 감정단어를 찾는 방법을 제안하고 있다. 영어권 연구에서 제시하는 방법들은 언어의 구조적 차이로 적용이 쉽지 않다. 수작업 분석을 통해 감정단어를 찾는 연구의 경우, 추출 시간이 너무

오래 걸린다는 문제점이 있고, 자동으로 추출한다 하더라도 정확도의 문제를 해결해야하거나 지극히 단순한 문장에서만 가능하기 때문에 본 논문에서 제안한 방법은 효과적이다. 향후 연구에서는 더 높은 성능을 가지기 위하여 다른 연구와 상호 보완적으로 정확도를 향상 시킬 수 있는 방법을 연구하고 문장 패턴 구조의 재 정의에 대한 연구가 필요하다.

참고문헌

- [1] Suxiang ZHANG, "Entity Relation Extraction to Free Text"
- [2] <http://www.moransoft.com/sentidict.pdf>
- [3] J. Myung, D. Lee, S. 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.347–405, Jun. 2008. (in Korean)
- [4] Jaewon Hwang and Youngjoong Ko, "A Korean Sentence and Document Sentiment Classification System Using Sentiment Features," Journal of Korean Institute of Information Scientists and Engineers (KIISE): Computing Practices and Letters, vol.14, no.3, pp.336–340, May, 2008. (ISSN 1229-6848)
- [5] Hanhoon Kang, Seong Joon Yoo, Dongil Han, "Automatic Extraction of Opinion Words from Korean Product Reviews Using the k-Structure"
- [6] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," In Proceedings of the Meeting of the Association for Computational Linguistics(ACL'02), pp.417–424 (2002).
- [7] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.79–86, 2002.
- [8] K.Dave, S. Lawrence, and D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," In Proceedings of the 12th Intl. World Wide Web Conference (WWW '03), pp. 512–528, 2003.
- [9] Qiang Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destination by supervised machine learning approaches," Expert Systems with Applications, Elsevier, pp.1–9, 2008.
- [10] Hanhoon Kang, Seong Joon Yoo, Dongil Han, "Accessing Positive and Negative Online Opinions," In Proceedings of the 13th International Conference on Human–Computer Interaction, HCII 2009, LNCS 5616, pp.359–368.
- [11] Youngho Kim, Yuchul Jung, and Sung-Hyon Myaeng, "An Opinion Analysis System Using Domain-Specific Lexical Knowledge," In Proceedings of the 4th Asia Information Retrieval Symposium, AIRS 2008, LNCS 4993, pp. 466–471.
- [12] M Hu and B. Liu, "Mining and Summarizing Customer Reviews," In Proceedings of ACM SIGKDD Intl. Conf on Knowledge Discovery and Data Mining(KDD '04), pp.168–177, 2004.
- [13] Soo-Min Kim, Eduard Hovy, "Determining the Sentiment of Opinions," Proceedings of the COLING conference, pp.1–8, 2004.