형용사만을 사용한 의견어 사전의 한계점 분석1)

유원희*, 지혜성*, 양영욱*, 임희석*
*고려대학교 컴퓨터교육학과
e-mail:gala@korea.ac.kr

Analysis of limitations using only adjectives sentiment word dictionary

WonHui Yu*, HeuiSeok Lim*
*Dept of Computer Education, Korea University

요 약

최근 많은 연구가 되고 있는 오피니언 마이닝은 의견어 사전의 구축이 가장 기본적으로 선행되어야 하는 연구이다. 오피니언 마이닝의 의견어 사전 구축 연구는 영어를 중심으로 많은 연구가 진행 되었다. 하지만 형용사 위주의 의견어 사전 구축으로 많은 부분의 문제들이 해결되는 영어에 비해서 한국어는 여러 가지 품사와 문장구조를 고려하여 의견어 사전을 구축해야한다. 이것을 실험으로 밝히기 위하여 형용사로만 구성되어진 의견어 사전을 구축하고 영화평에 적용하여 분석해 봄으로써 형용사로만 구성되어진 의견어 사전의 한계점을 확인한다. 실험은 세종계획 말뭉치에서 나타나는 형용사로 구성된 의견어 사전을 구축하고 네이버 랩에서 제공하는 영화평을 형용사로 구성된 의견어 사전으로 의견 분석하여 시행하였다. 분석 결과 재현율 약 50%, 정확률 약 60%정도의 성능을 보였다.

1. 서론

SNS와 같이 온라인에서 자신의 의견을 표현하기가 쉬워진 상태에서 회사는 자사의 상품, 브랜드, 회사이미지들에 대해서 고객들이 느끼는 긍정 부정 감성을 온라인에서 쉽게 획득 할 수 있게 되었다. 특히 리뷰데이터와 같은 대량의 데이터 안에서 유용한 정보를 찾아내는 오피니언 마이닝 분야가 최근 활발하게 연구되고 있으며, 오피니언 마이닝 기술을 가지고 서비스 하는 회사들도 늘어나고 있다.

하지만 오피니언 마이닝을 하고 있는 회사에서도 아직까지 많은 부분 단순 형용사 의견어 사전에 의존되어 있는 것 같은 결과를 보인다. 이와 같은 현상은 영어권 오피니언 마이닝 기술을 한국어 특성을 고려하지 않은 상태에서 그대로 적용하거나, 한국어 자연어처리 기술의 부재로 나타날 수 있을 것이다.

본 논문에서는 형용사들로 구성된 의견어 사전을 사용하여 영화평을 분석하고 단순하게 형용사로 구성된 용언으로 분석한 분석 결과에서 나타나는 한계점들을 확인하여 본다.

1) “이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0018014).”

2. 관련연구

Hatzivassiloglou and McKeown(1997)는 접속사에 의해 연결된 형용사 짝을 예측하는 연구를 진행하였다[1]. 예를 들어 AND는 동일 방향성을 나타내고 BUT는 반대 방향성을 나타낸다는 것이다. 접속사에를 기반으로 하여 단어가 연결된 형태를 그래프로 생성하고 긍정 클러스터와 부정 클러스터로 나누어 보았다. Turney and Littman(2003)은 두 가지의 작은 단어 모음을 시드로 하여 단어 목록을 증가시켜가는 방식으로 의견 단어를 분류하였다[2]. poining mutual information(PMI)로 단어들을 계산하여 분류하였다. Kamps et al.(2004)은 WordNet의 유사어 사전을 이용하여 분류하고자 하는 형용사의 긍정, 부정 방향성을 결정하였다[3].

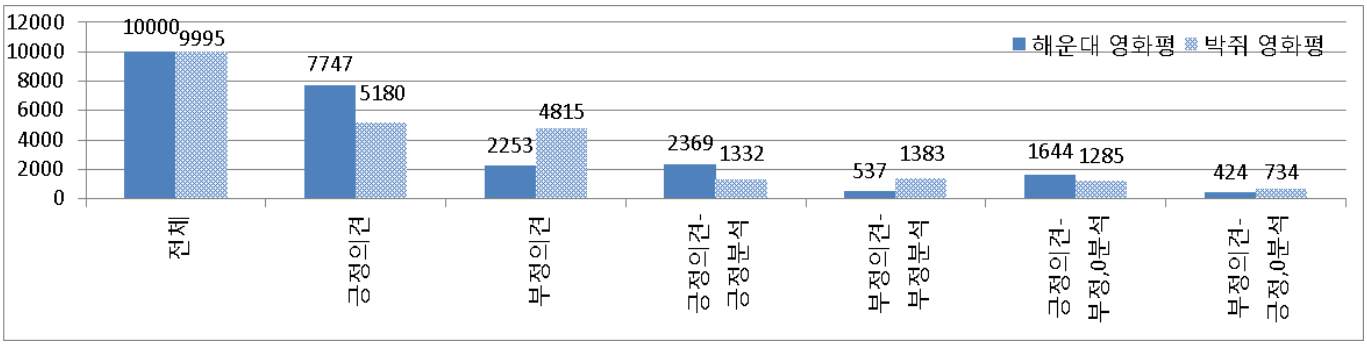
이와 같이 의견어 사전 구축에 관한 논문들은 영어를 중심으로 많은 연구가 진행 되었다. 영어는 고립어적 특성을 가지고 교착어적 특성을 많이 가지지 않기 때문에 단어 자체가 가지고 있는 의견만을 고려하면 의견어를 구분하는 문제의 많은 부분이 쉽게 해결되는 특징을 가지고 있다.

명재석 외(2008)는 단일 용언들로 구성된 의견어 사전으로 한국어 상품평을 분석하였다[4]. 양정연 외(2009)는 PMI를 사용하고, 문맥정보를 추가로 고려하여 의견어 사전을 구성하였다[5]. 강한훈 외(2010)는 한국어의 문맥구

조에서 고빈도로 나타나는 문장 구조를 이용하여 의견어 사전을 구성하였다[6]. 영문 분석에 사용되는 의견어 사전과는 다르게 한국어문 분석에 사용되는 의견어 사전에는 추가적으로 구문 구조 정보들을 사용한다는 차이점을 보이고 있다.

람들이 평가해놓은 10점 척도와 비교하여 분석하였다.

먼저 "해운대"영화평의 결과이다.
전체 영화평의 개수는 10,000개이다. 영화평은 10점 척도로 일반인들이 평가해 놓았으며 평가할 때 자신의 의견도 함께 작성해 놓은 형태이다. 본 논문에서 제시한 형태소로



(그림 1) 영화평 분석 통계 데이터

3. 형용사 의견어 사전 제작

일반적으로 오피니언 마이닝 연구는 "의견어 사전 구축"과 "긍정적인 의견과 부정적인 의견 분류"가 가장 기초적인 연구 분야로 해당 연구가 어느 정도 선행 된 상태에서 "의견 속성 추출", "토픽 관련 문서 검색", "의견 추출", "의견요약", "의견 질의/응답"등의 확장된 연구를 진행 해 볼 수 있다[7].

본 논문에서는 의견어 사전 구축에서 주로 사용되는 형용사들을 의견어 사전으로 구축하고, 구축된 의견어 사전을 사용하여 의견을 담고 있는 영화평을 긍정과 부정으로 분류해 본다.
실험에 쓰일 형용사들로 이루어진 의견어 사전은 세종계획 코퍼스[8]에서 출현 빈도 5이상의 형용사 678개를 대상으로 전문가들이 수동으로 PNO(긍정/부정/중립)태깅 하였다.

여러 전문가들이 각각 678개의 형용사들을 직접 태깅하였으며 가장 고빈도로 태깅된 PNO속성을 해당 형용사의 속성으로 선택하였다. 태깅 결과 긍정 형용사 129개, 부정 형용사 252개 그리고 긍정과 부정이 아닌 형용사 297개로 구성된 의견어 사전을 제작하였다.

4. 실험

형용사로 구성된 의견어 사전의 한계점을 알아보기 위하여 제작한 의견어 사전으로 영화평을 분석해 보았다. 실험에 사용한 데이터는 네이버 랩에서 오피니언 마이닝 실험 데이터로 제공하는 영화평 "해운대"와 "박쥐"를 사용하였다[9]. 제공된 영화평을 형태소 분석을 하고 형태소가 분석되어 나타난 형용사들을 의견어 사전에서 나타나는 빈도로 점수화 하였다. 점수화된 영화평 분석결과를 실제 사

구성된 의견어 사전으로 분석한 결과는 (그림 1)에서 보는 것과 같이 긍정의견 7747개, 부정의견 2253개로 구성되어 있는 영화평 중에서 긍정의견을 긍정으로 분석한 개수는 2369, 부정의견을 부정으로 분석한 537개가 올바르게 분석되었다. 긍정의견을 부정의견이나 중립의견으로 분석한 경우가 1644개, 부정의견을 긍정의견으로 분석하거나 중립의견으로 분석한 경우가 424개이다.

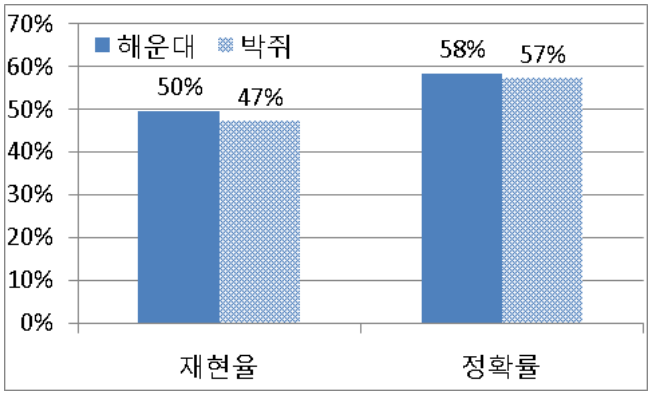
```
<data>
 <opinion>
 <rating>10</rating>
 <comment>정말 극장에서 눈물흘린 적은 없었는데 ㅋ 정말 감동적이고 재미있는 영화였습니다ㅎ
</comment>
 </opinion>
 <opinion>
 <rating>5</rating>
 <comment>스토리가 평범</comment>
 </opinion>
        :
 <opinion>
 <rating>1</rating>
 <comment>관객수 천만이 넘었다는 해운대.. 나에겐 아쉽기만 한 영화.</comment>
 </opinion>
 </data>
```

(표 2) 실제 XML 형태 제공 데이터

두 번째로 "박쥐"영화평의 결과이다.
전체영화평은 오류문자가 포함된 5개의 영화평을 제외한 9995개로 이루어져있다.
전체 영화평의 개수는 9995개이다. 영화평은 10점 척도로

일반인들이 평가해 놓았으며 자신의 의견도 함께 제시해 놓은 형태이다. 긍정의견은 5180개, 부정의견은 4815개로 구성되어 있으며 긍정의견을 긍정으로 분석한 경우가 1332 부정의견을 부정으로 분석한 경우가 1383개 이다. 또한 긍정의견을 부정이나 중립의견으로 분석한 경우가 1205, 부정의견을 긍정이나 중립의견으로 분석한 경우가 743건이었다.

두 가지 영화평의 분석결과를 재현율과 정확률로 나타내면 (그림 2)와 같다. 두 경우 모두 50%와 47%의 재현율로 전체데이터의 절반 정도의 분석력을 보여주었다. 또한 정확률을 58%와 57%로 분석된 데이터에 대해서 절반 정도의 정확한 결과를 보였다.



(그림 2) 분석된 영화평의 재현율과 정확률

한국어의 문장구조가 반영되지 않고 단순 형용사로 구성된 의견어 사전으로 영화평을 분석하는 것은 전체 데이터 중에서 약 50% 정도의 분석과 분석한 내용의 약 60%정도의 분석 정확률을 보였다. 이것은 전체 데이터의 약 30%를 정확히 분석 한다는 것을 뜻한다. 형용사만 포함된 의견어 사전을 도메인별로 생성한다고 하더라도 정확히 분석하는 데이터는 재현율의 한계치인 50%정도에 불과할 것으로 확인된다. 이것은 의견어 사전을 구성할 때 형용사 품사뿐만 아니라, 체언을 이루는 명사류와 용언의 동사, 수식언의 관형사와 부사를 고려하여 의견어 사전을 구성해야한다.

5. 결론

본 논문에서는 형용사로 이루어진 의견어 사전을 구축하고 구축된 의견어 사전을 이용하여 영화평 두 가지를 분석해 보았다. 분석결과 약 50%의 재현율과 약 60%의 정확률을 확인하였다. 이것은 의견어 사전을 구축할 때 형용사 뿐문아니라 명사류, 동사, 관형사, 부사 등의 품사를 가지고 있는 형태소들이 추가적으로 필요하다는 것을 말해준다. 또한 한국어의 교착어적인 특성을 반영하기 위해서는 추가적으로 한국어의 문장구조를 고려한 규칙들이 함께 고려되어야 한다. 차후 연구로 여러 품사를 포함하고, 한국어의 문장 구조 규칙을 학습한 의견어 사전을 구축하고, 구축된 의견어 사전으로 다양한 사용평들을 비교 분석해 보려 한다.

참고문헌

[1] Vasileios Hatzivassiloglou and Kathleen McKeown. "Predicting the semantic orientation of adjectives". 1997. In Proc. of the 35th ACL/8th EACL, pages 174-181.

[2] Peter D. Turney and Michael L. Littman. "Measuring praise and criticism: Inference of semantic orientation from association". 2003. ACM Transactions on Information Systems, 21(4), pages 315 - 346.

[3] Jaap Kamps, Maarten Marx, R. ort. Mokken, and Maarten de Rijke. "Using WordNet to measure semantic orientation of adjectives". 2004. In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV.

[4] 명재석, 이동주, 이상구 "반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템", 2009, 정보과학회 논문지 : 소프트웨어 및 응용 제 35권, 제6호 , Volume , Page 392-403

[5] 양정연, 명재석, 이상구 "상품 리뷰 요약에서의 문맥 정보를 이용한 의견 분류 방법", 2009 ,정보과학회논문지 : 데이타베이스 , Volume 제36권 제4호 , Page 254-262

[6] 강한훈, 유성준, 한동일, "k-Structure를 이용한 한국어 상품평 단어 자동 추출 방법 ", 2010, 정보과학회논문지 : 소프트웨어 및 응용, pages 470-479

[7] Wiki, http://en.wikipedia.org/wiki/Sentiment_analysis

[8] 21세기 세종계획, 국립국어원, www.sejong.or.kr

[9] 네이버랩, http://lab.naver.com/research