PMI를 이용한 우리말 어휘의 의미 극성 판단

송상일 ○ 이동주 이상구

서울대학교 컴퓨터공학부

{danox, therocks, sglee}@europa.snu.ac.kr

(Identifying Sentiment Polarity of Korean Vocabulary Using PMI)

Sang-il Song○ Dongjoo Lee Sang-goo Lee

School of Computer Science & Engineering, Seoul National University

요 약

웹 2.0시대의 도래에 따라 많은 소비자들은 상품에 대한 다양한 의견을 표현할 수 있게 되었다. 이러한 의견들을 활용하여 상품평 요약 시스템 등이 개발되었다. 어휘의 의미 극성은 이러한 시스템에서 활용될 여지가 많은 요소이다. 영어의 경우 어휘의 의미 극성을 판단하는 연구가 많이 진행되어 어느 정도 결실을 맺었지만, 우리말의 경우 어휘의 의미 극성을 판단하는 연구는 아직 미흡하다. 본 논문에서는 우리말 어휘의 의미 극성을 PMI를 사용하여 판단한다. 또한 PMI를 우리말 어휘에 적용할 때 문제가 되는 이슈를 살펴보고 이에 대한 해결 방법들을 제시한다. 나아가 실제 상품 평에서 많이 쓰이는 형용사에 대하여, 제시한 의미 극성 판단 방법의 성능을 검증해 본다. 제시한 방법은 어휘의 의미 극성을 81%의 정확도로 판단해 주었다.

1. 서 론

웹 2.0과 유비쿼터스 웹의 발전에 따라 더욱 더 많은 사람들이 상품에 대한 다양한 의견을 온라인 쇼핑몰, 블로그, 온라인 커뮤니티 등에 표현하고 있다. 이런 상품평은 사용자의 구매에 영향을 줄 수 있기 때문에 상품평을 이용하여 사용자에게 상품에 대한 정보를 제공하는 상품평 요약 시스템[1] 등이 연구되고 있다. 이러한 상품평 요약 시스템의 주요 기능 중 하나는 상품평이 상품에 대해 긍정적인지 부정적인지 판단하는 것이다.

상품평이 긍정적인지 부정적인지 판단하는 방법은 여러 가지가 있는데, 각 어휘의 의미 극성을 이용해 상품평의 의미 극성을 판단하는 방법도 널리 쓰이는 방법들 중 하나이다. 외국어의 경우 어휘가 가진 의미 극성을 판단하는 연구가 많이 진행되어 상당한 성과를 거두고 있다. 하지만 우리말 어휘가 가진 의미 극성을 판단하는 연구는 아직 미흡하다.

따라서 본 논문에서는 우리말 어휘가 가진 의미 극성을 판단하고자 한다. 이를 위해서 [2]에서 제시된 SO-PMI(Semantic Orientation from Point-wise Mutual Information)를 이용하여 어휘의 의미 극성을 판단하고, 이 때 발생하는 여러 문제점들을 해결하고자 한다. 또한 이를 실제 상품평에서 많이 쓰이는 형용사를 대상으로 적용해 보고 우리가 제시한 여러 가정과 해법의 타당성을 평가해 볼 것이다.

본 논문의 나머지 부분은 다음과 같이 구성되어 있다. 2장에서는 어휘의 극성 판단과 관련된 이전의 연구들에 대한 전반적인 내용을 소개한다. 3장에서는 SO-PMI를 이용하여 어휘의 극성을 판단하고, 이 때 발생하는 문제점과 해결책을 다룬다. 이 때 판단된 의미 극성의 왜곡이 생길 수 있는데, 4장에서는 이를 보정하는 방법을 다룬다. 5장에서는 제시한 방법을 실제 상품평에서 많이 쓰이는 형용사들을 대상으로 적용해보고 이에 대해 상세한 분석을 해 본다. 6장에서는 결론 및 향후 과제에 대해 논의한다.

2. 관련 연구

2.1. 오피니언 마이닝 (Opinion Mining)

오피니언 마이닝은 특정 주제에 대한 글쓴이나 화자의 태도를 찾아내는 것을 말한다. 웹 2.0의 발전으로 사용자들이 더 많은 의견을 표현함에 따라 오피니언 마이닝이 더욱 주목 받고 있다.

[3]은 오피니언 마이닝의 주요 주제를 언어학적 자원을 개발하고 발전 시키는 것[2,4-8], 의견의 의미 극성을 판단하는 등 의견을 요약하는 것[1,9], 텍스트로부터 의견이 표현된 부분을 추출하는 것[10]으로 나눈다.

오피니언 마이닝 초기에는 자연어 처리 기법을 많이

이용하였다. 하지만 이러한 방식은 실제 적용에 있어 많은 한계점들을 보여주었다. 이러한 한계점을 극복하기 위해 최근의 연구들은 통계적 분석을 기존의 자연어 처리 기법과 함께 사용하여 좋은 성과를 거두고 있다[2,4-5].

본 논문에서는 언어학적 자원의 하나인 우리말 어휘가 가지는 의미 극성을 판단하려고 한다. 이는 의견의 의미 극성을 판단하고, 텍스트로부터 의견이 표현된 부분을 추출하는데 도움을 줄 것이다.

2.2. 어휘의 의미 극성 판단

어휘의 의미 극성 정보는 오피니언 마이닝 분야에서 중요한 언어학적 자원 중 하나이다. 어휘의 의미 극성을 전산언어학적인 방법으로 접근하려는 시도는 90년대 말에 시작되었다.

이러한 시도들 중 대표적인 것으로 [4]가 있다. 이는 'and'로 연결된 형용사들은 비슷한 극성을 가질 것이라는 가정을 바탕으로 문제를 해결하고자 하였다. Turney 등은 확률론에 기반한 PMI(Point-wise Mutual Information)를 사용하여 어휘의 의미 극성을 판별하였다[2]. PMI는 비교적 간단함에도 불구하고 좋은 결과를 얻을 수 있어 [5]에서도 사용되었다. PMI에 대한 자세한 소개는 2.3에서 하고자 한다. Kamps 등은 어휘 온톨로지인 WordNet[6]을 이용하여 어휘의 의미 극성을 판단하였다[7]. Esuli 등은 해당 어휘의 용법을 이용하여 어휘의 극성과 객관성을 판단하였고[8], 이를 바탕으로 SentiWordNet[11]을 구축하였다.

이처럼 외국어, 특히 영어의 경우 어휘의 의미 극성을 판단하는 연구가 많이 진행되어 상당히 성숙한 결과를 보여주고 있다. 우리말의 경우, 어휘의 체계를 구축하고자 하는 연구는 많이 진행되어 [12-14]등이 구축되었지만, 우리말 어휘의 의미 극성을 판단하는 연구는 아직 미흡한 실정이다.

우리말 어휘의 극성을 판단하기 위해 고려할 수 있는 방법은 앞에서 제시된 것처럼 다양하다. 최근의 경향은 [7][8]에서 제안된 것과 같이 어휘망을 이용하는 것이다. 하지만 공개 API 등을 제공하여 접근 및 사용이 용이한 우리말 어휘망이 없고, 실제 상품평에서 자주 쓰이는 '이쁘다', '이뿌다'와 같은 비표준어는 어휘망에 포함되지 않아 비표준어의 의미 극성을 판단할 수 없다는 단점이 있다. 따라서 본 논문에서는 [2][5]에서 사용된 PMI에 기반한 방법을 사용한다. PMI에 기반한 방법은 웹 검색 서비스나 시스템을 사용하여 쉽게 적용할 수 있다는 장점이 있고 성능 또한 우수하다.

2.3 Point-wise Mutual Information (PMI)

PMI는 확률론에 기초한 방법으로 두 확률 변수의 특정한 값의 연관성을 표현하는 지표이다. 의미 극성이 비슷한 단어들은 같은 문서 안에 나타날 확률이 높다고 가정하면, PMI를 이용해서 두 어휘의 연관성을 측정할 수 있다. 이는 PMI를 이용한 두 어휘의 연관성은 (1)과 같이 표현된다.

$$PMI(w1,w2) = log\frac{p(w1,w2)}{p(w1)p(w2)} \quad (1)$$

여기서, w1 과 w2 는 연관성을 구하고자 하는 두 단어이다.

PMI는 위의 식처럼 두 단어가 같은 문서 안에 나타날 확률 값과 특정 단어가 문서 내에 나타날 확률 값으로 표현된다. 이 때 두 단어가 나타날 확률이 서로 독립적이라면 PMI값은 0이 될 것이다. 만약 PMI값이 양수이면 두 어휘가 같은 문서 안에 나타날 확률이 높아 비슷한 의미 극성을 가진다는 것을 의미할 것이고, PMI 값이 음수라는 것은 두 단어가 같은 문서에 나타날 확률이 낮아 다른 의미 극성을 가진다는 것을 의미할 것이다.

Web-PMI라 불리는 방법은 검색 시스템을 사용하여, p(w)를 정한다. 여기서 p(w)는 다음과 같이 정의된다.

$$p(w) = \frac{1}{N}hits(w) \quad (2)$$

여기서 hits(w)는 w가 포함된 문서의 개수이고, N은 전체 문서의 개수이다. 이를 (1)에 적용하면 Web-PMI는 (3)과 같이 표현된다.

$$Web-PMI(w1,w2) = log\ \frac{\frac{1}{N}hits(w1\ AND\ w2)}{\frac{1}{N}hits(w1)\ \frac{1}{N}hits(w2)} \quad (3)$$

Web-PMI를 이용하여 어휘 극성을 판단하기 위한 방법으로는 [2]에서 제안된 SO-PMI(Semantic Orientation from Point-wise Mutual Information)이 있다. 이 방법은 전문가가 미리 긍정적 의미를 지닌 어휘들의 집합(긍정 기준 어휘 집합)과 부정적 의미를 지닌 어휘들의 집합(부정 기준 어휘 집합)을 기준 어휘(seed term) 집합으로 정해 놓고, 기준 어휘들과 의미 극성을 알고자 하는 어휘의 Web-PMI값을 이용하여 어휘의 의미 극성 값을 구한다. 이를 이용하면 어휘의 의미 극성 판단을 간단하면서도 좋은 성능으로 할 수 있으므로[2] 본 논문에서는 SO-PMI 방법을 사용하도록 하겠다. 이는 (4)과 같이 표현된다.

$$SO-PMI(w) = \sum_{pw\in PW} PMI(w,pw) - \sum_{nw\in NW} PMI(w,nw) \quad (4)$$

여기서 PW는 긍정 기준 어휘 집합이며, NW는 부정

기준 어휘 집합이다.

어휘가 긍정적일수록 pw와의 PMI값이 크고, nw와의 PMI값이 작아 SO-PMI값이 크게 나올 것이다. 반대로 어휘가 부정적일수록 SO-PMI값이 낮게 나올 것이다. 즉, SO-PMI값을 이용하여 두 어휘의 의미 극성을 상대적으로 비교할 수 있다.

3. 어휘의 의미 극성 판단

이 장에서는 어휘의 의미 극성을 판단하기 위해 앞에서 소개한 SO-PMI를 사용할 때 생길 수 있는 문제점과 개선 방향을 살펴본다.

어휘의 의미 극성을 판단하기 위해서는 어휘의 SO-PMI값을 구해야 한다. 이 때 어휘의 SO-PMI값을 구하기 위해서는 기준 어휘 집합과 문서 집합이 필요하다. 이 때 기준 어휘 집합과 문서 집합에 따라서 SO-PMI값이 달라지므로 올바른 SO-PMI값을 구하기 위해서는 '적절한' 기준 어휘 집합과 '적절한' 문서 집합을 선택해야 되는데, 이는 기준 어휘 집합과 문서 집합의 '적절성' 판단이 필요하다는 것을 의미한다. 이 장의 남은 부분에서 '적절한' 기준 어휘 집합을 선택하는 방법과 '적절한' 문서 집합의 특성을 제시할 것이다.

또, (2)에서 hits(w)를 구하기 위해 우리말의 특성을 고려한다면 더 정확한 SO-PMI를 구할 수 있을 것이다. 한국어의 특성을 고려하여 SO-PMI값을 구하는 방법도 이 장의 남은 부분에서 제시할 것이다.

3.1 기준 어휘(seed term)의 선택

앞에서 언급했듯이 SO-PMI 방법을 사용하기 위해서는 먼저 기준 어휘 집합을 선택해야 한다. SO-PMI값이 기준 어휘 집합에 영향을 받으므로 '적절한' 기준 어휘 집합을 정하는 것이 중요하다. 이는 기준 어휘 집합에 포함될 어휘와 기준 어휘 집합의 크기를 정하는 것인데, 기준 어휘 집합에 포함될 기준 어휘를 선택할 때는 다음과 같은 점을 고려 할 수 있다.

첫 번째로 기준 어휘는 다른 어휘들의 의미 극성을 잘 판별해야 한다. 즉, 긍정 기준 어휘라면 긍정적인 어휘들과의 PMI값이 높고, 부정적인 어휘들과의 PMI값은 낮아야 한다. 이를 판단하기 위해 본 논문에서는 해당 어휘 하나만을 기준 어휘로 사용하여 여러 어휘들의 의미 극성을 판단했을 때의 정확도를 이용한다. 이 정확도가 높을수록 이 어휘는 다른 어휘들의 의미 극성을 잘 판별할 수 있다는 것을 의미한다고 생각할 수 있기 때문이다.

두 번째로 기준 어휘는 널리 쓰이는 어휘이어야 한다. 이는 기준 어휘가 널리 쓰이지 않으면, 그 기준 어휘를 이용하여 구한 PMI값이 표본이 너무 적어 통계적으로 신뢰하기 힘들어 이를 이용해 판단한 어휘의 극성이 왜곡되기 싶기 때문이다.

본 논문에서는 이 두 가지를 요소를 반영하여 특정 어휘의 선호 함수(score function)을 정의하였다.

$$score(w) = accuracy(w) \times log_2 popularity(w) \quad (5)$$

여기서 w는 점수가 계산될 어휘이고, accuracy(w)는 w 하나만을 기준 어휘로 사용 할 때의 극성 판단의 정확도, popularity(w)는 문서 집합에서 w가 포함된 문서의 개수이다.

정의한 선호 함수를 이용하여 기준 어휘를 선정하였을 때, 어휘 극성 판단의 정확도가 어떻게 달라지는지 실험을 통해 알아볼 것이다. 이를 통해 기준 어휘 집합의 선정에 있어 선호 함수의 타당성을 입증할 것이다. 이에 대한 상세한 분석은 5장에서 기술하겠다.

그 다음으로는 기준 어휘 개수도 정해야 한다. 기준 어휘가 적을 경우, 한 두개의 기준 어휘에 의해 어휘의 의미 극성이 왜곡될 확률이 높다. 반대로 기준 어휘가 너무 많을 경우, 앞에서 설명한 '좋은' 기준 어휘가 가져야 할 특성을 올바르게 반영하지 않는 기준 어휘를 선택할 가능성이 높아 오히려 정확도를 떨어뜨릴 가능성이 있다. 5장에서 어휘의 개수를 변화시키면서 의미 극성 판단에 대한 정확도의 변화를 살펴보면서 이러한 주장의 타당성을 증명하고, 최적의 기준 어휘 개수를 찾아낼 것이다.

3.2 문서 집합(document collection)의 선택

SO-PMI값은 문서 집합에 의해서도 영향을 받는다. 따라서 올바른 SO-PMI값을 제공하는 문서 집합을 선택해야 한다. 이것이 가져야 할 특성도 적절한 기준 어휘의 특성과 비슷하다.

먼저 선택된 문서 집합은 해당 어휘의 쓰임을 정확히 반영해주어야 한다. 특정 어휘가 너무 적게 쓰이거나, 너무 많이 쓰이게 되면 SO-PMI값이 왜곡될 가능성이 높다. 또한 어휘의 뜻이 많이 변용되어 쓰이면 안 된다. 즉, 특정 주제와 같이 너무 전문용어가 많이 나타나거나, 너무 변질된 표현이 많이 나타난 문서들이 많이 포함된 문서 집합을 선택해서는 안 된다.

두 번째로는 문서의 수가 충분히 많아야 한다. 문서가 충분히 많지 않으면 앞에서 언급한 바와 같이 검색된 문서의 수가 적어 이 문서 집합을 이용하여 구한 SO-PMI값을 통계적으로 신뢰하기 어렵다

이러한 기준을 바탕으로 본 논문에서는 문서 집합으로 네이버 검색 시스템을 사용하기로 하였다. 네이버는 오픈 API를 제공하기 때문에 사용하기 쉽고, 특정 주제에 치우쳐 있지 않는데다가 충분히 많은 문서를 가지고 있기 때문이다. 단, 네이버 오픈 API는 문서의 총 수를 알 수 없기 때문에 검색된 문서의 수를

이용하여 어휘가 나타날 확률을 계산할 수가 없다. 따라서, 긍정 기준 어휘의 개수와 부정 기준 어휘의 개수가 같다면 (3)에서 PMI항(2)의 문서의 총 개수(N)를 소거해도 SO-PMI값이 같다는 점을 이용하여 이 문제를 해결하였다.

3.3 우리말 어휘의 특징 반영

SO-PMI는 본래 영어 어휘의 의미 극성 판단을 위해 개발된 방법이다. 따라서 이를 우리말에 그대로 적용하기 보다는 본 논문에서는 우리말 어휘의 특징을 반영하도록 SO-PMI를 조금 개선하였다.

먼저, 우리말은 형태소에 다양한 접사가 더해져 단어를 만드는 교착어이다. 예로, 형용사 형태소에 접사 '-다', '-ㄴ/은', '-게' 등을 붙여 서술어, 관형어, 부사어 등으로 사용된다. 이러한 특징을 살려 (2)에서 hits(w)를 어휘의 서술어형('-다'), 관형어형(-'ㄴ/은'), 부사어형('-게')이 검색된 문서의 합으로 정의하였다. 즉, hits('예쁘다')를 '예쁘다', '예쁜' 또는 '예쁘게'가 포함된 문서의 개수로 정하였다. 이렇게 함으로써 SO-PMI값의 정확도를 높이고 같은 형태소에서 파생된 어휘의 의미 극성의 일관성을 얻을 수 있다.

두 번째, 우리말에서는 용언의 부정 표현 방법 중 하나로 '안 -' 있다. 이 '안 -'의 경우에는, Web-PMI의 결과에 부정적인 영향을 미칠 수 있다. 예를 들어 '좋다'라는 어휘의 PMI를 구하고자 할 때 웹 검색 엔진을 이용하면, '안 좋다'가 포함된 문서도 검색되기 때문에 전체 PMI 값이 왜곡된다. 본 논문에서는 이러한 점을 고려하여 검색엔진에 던질 질의(query)에서 '안 ~'을 배제하도록 명시하여 이를 해결한다. 즉 '예쁘다'라는 질의 대신에 '"예쁘다" AND (NOT "안 예쁘다")'라는 질의를 사용한다.

4. 의미 극성 값 보정

위에서 설명한 방법들을 이용하면 해당 어휘의 SO-PMI 값을 구할 수 있다. 하지만 이를 이용해 어휘의 의미 극성으로 바로 사용하기에는 무리가 있다. 긍정 어휘가 부정 어휘보다 많이 포함된 어휘 집합을 대상으로 각 어휘들의 SO-PMI값의 평균을 계산하면 양수가 나와야 하지만, 실제 실험 할 경우 SO-PMI값의 평균은 음수가 나오는 등의 문제가 있었다. 따라서 SO-PMI값을 보정하여 어휘의 올바른 의미 극성을 판단하고자 한다. 본 논문에서는 미리 정해진 학습 데이터를 이용하여 긍정 어휘들의 PMI 평균, 부정 어휘들의 PMI 평균을 구한 뒤, 이 두 값이 평균 값을 중립적인 어휘의 SO-PMI값으로 정하고, 이를 기준으로 하여 이보다 크거나 같을 경우 긍정으로, 작을 경우 부정으로 판단하는 분류기 (classifier)를 사용하였다. 이는 (6)과 같다.

$$classify(w) = \begin{cases} 'P' & if\ PN - PMI(w) \geq c \\ 'N' & o.w \end{cases} \quad (6)$$

$$\text{where } c = \frac{\sum_{pwerE} PNerE\ (pw) + \sum_{nw \in NE} PN \in NE}{2}$$

여기서 PE는 긍정 예제(positive example), NE는 부정 예제(negative example)이다. c 는 긍정 예제와 부정 예제의 SO-PMI값의 평균으로, c 를 중립적인 어휘의 SO-PMI값이라고 생각한다. c보다 큰 SO-PMI값을 가진 어휘는 긍정, 작은 SO-PMI값을 가진 어휘는 부정으로 판단한다.

5. 실험

본 장에서는 앞에서 제시한 선호함수의 타당성, 적절한 기준 어휘 집합의 크기, 문서 집합에 따른 어휘의 의미 극성 판별의 정확성을 살펴볼 것이다. 이 장의 남은 부분은 실험 설계, 실험 결과 및 그에 대한 논의로 구성되어 있다.

5.1 실험 설계

첫 번째 실험으로는 3.1에서 제시한 선호함수의 타당성을 보이기 위해 선호 점수에 따른 정확도의 변화를 살펴 볼 것이다. 두 번째 실험은 기준 어휘의 집합의 크기에 따른 정확도를 살펴보기 위하여 기준 어휘 후보 중 점수가 높은 순서대로 기준 어휘에 포함시키면서 정확도의 변화를 살펴볼 것이다. 마지막으로 네이버 오픈 API에서는 다양한 카테고리를 제공하는데, 그 중에서 블로그와 뉴스를 대상으로 실험할 것이다. 이 때 뉴스에서는 잘 쓰이지 않는 비표준어의 경우 SO-PMI값이 왜곡이 발생한 다는 것을 보여 우리가 가정한 '적절한' 문서 집합의 특성이 타당함을 보일 것이다.

5.2 실험 대상 어휘의 선택

본 논문에서는 실제 상품평에서 많이 쓰이는 어휘들을 대상으로 실험하였다. 이를 위해 많은 사람들이 이용하는 온라인 쇼핑몰 중 하나인 G마켓을 대상으로 상품평을 수집하였다. 2010년 3월 17일부터 3월 28일에 걸쳐 티셔츠, 청바지, 노트북, 모니터, 여성 가방, 휴대폰 등의 카테고리에 걸쳐 상품 리뷰 189643개를 수집했다.

이러한 상품 리뷰를 분석하기 위해서는 형태소 분석기가 필요하다. 본 논문에서는 꼬꼬마 형태소

분석기(KKMA)[14]를 사용했다. 꼬꼬마 형태소 분석기는 2009 공개 소프트웨어 공모 대전에서 은상을 수상하여 형태소 분석의 우수성이 증명되었다. 또한 소스코드가 공개되어 있고, 지속적으로 사전이 관리되고 있다는 장점이 있다.

실험 대상 어휘들은 수집한 상품평에 대해 형태소 분석을 하여, 널리 쓰이는 형용사 형태소들을 추렸다. 이들 형용사 형태소에 대하여 5명의 사람이 1~5로 극성을 판별한 후, 극성이 뚜렷한 형용사 200개를 실험 대상 어휘로 선정하였다. 이 중 긍정으로 판단한 것이 121개, 부정으로 판단한 것이 79개이다.

보정을 위한 학습 데이터는 긍정 형용사 20개, 부정 형용사에서 20개를 무작위로 선택하였다. 그리고 SO-PMI값을 앞에서 언급했던 것과 같이 네이버 오픈 API를 사용하여 구하였다. 앞의 두 실험은 블로그 카테고리로 한정 시켜 실험 하였고 마지막 실험은 블로그와 뉴스에 대해서 실험하여 두 카테고리의 결과를 비교하였다.

5.3 결과 및 논의

5.3.1 기준어휘 선정을 위한 선호 함수에 대한 평가

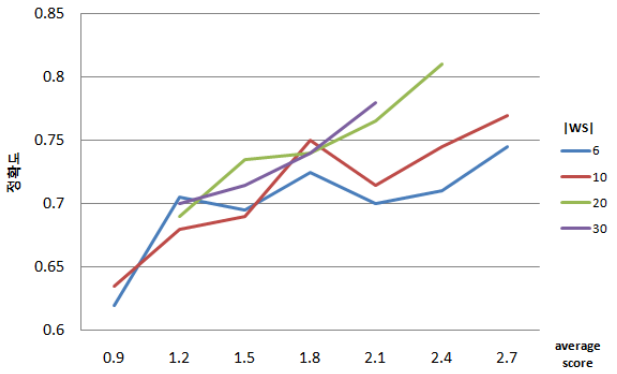

그림 1. score값에 따른 어휘의 정확도 변화

그림 1은 기준 어휘 집합에 속한 기준 어휘들의 선호 점수 평균에 따른 정확도의 변화를 보여준다. x축은 어휘 집합에 포함된 어휘들의 평균 선호 점수를 나타내고, y축은 어휘 극성 판단의 정확도를 나타낸다. 각 그래프는 기준 어휘의 개수를 나타낸다. 전체적으로 기준 어휘의 평균 선호 점수가 높아짐에 따라 정확도도 높아지는 결과를 얻을 수 있었다. 따라서 본 논문에서 제시한 선호 함수가 기준 어휘를 선정할 때 좋은 영향을 미친다고 할 수 있다.

여러 변수 중 기준 어휘 집합의 크기가 20일 때 81%의 정확도로, 가장 좋은 결과를 얻을 수 있었다. 또한 어휘 집합의 크기에 따라 정확도가 차이를 보였는데, 이를 바탕으로 기준 어휘의 개수에 따른 정확도를 비교해 보는 실험을 수행하였다.

5.3.2 기준 어휘 집합의 크기에 따른 정확도 변화

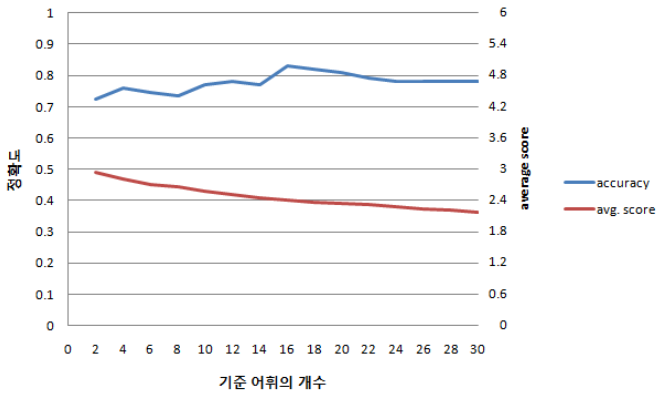

그림2. 어휘집합에 따른 정확도의 변화

그림 2는 기준 어휘의 개수에 따른 정확도의 그래프를 나타낸 것이다. 추가적으로 기준 어휘의 평균 선호 점수도 같이 표시하였다. 이는 어휘들을 선호 점수가 높은 순서대로 나열한 후, 어휘 집합에 하나씩 포함하면서 정확도의 변화를 나타낸 것이다. 어휘 집합을 하나씩 늘려갈수록 점점 증가하는 양상을 나타내다가 일정 수치(N=8)이상 에서는 더 낮아지는 결과를 보여준다. 이는 3.1에서 논하였듯이, 기준 어휘가 많아 질수록 특정한 기준 어휘에 의해 왜곡될 확률이 낮아지지만, 너무 많은 기준 어휘의 경우 오히려 잡음(noise)으로 작용해 정확도를 떨어뜨리기 때문으로 생각된다.

5.3.3 대상 문서 집합 분류에 따른 정확도 변화

마지막으로 문서 집합의 카테고리를 달리하여 어휘들의 SO-PMI값의 양상을 살펴본다. 기준 어휘의 개수를 20으로 하였을 때 블로그와 뉴스 모두 비슷한 정확도(0.81,0.79) 등을 보여주었다. 하지만 실험에 사용된 200개의 형태소 중에 포함된 비표준어 형태소 - “이쁘(예쁘)”, ”이뽀(예쁘)”, ”예뿌(예쁘)”, ”조아하(좋아하)”, ”산듯하(“산뜻하”) 등에서 큰 차이를 보였는데, 그 결과는 표1과 같다.

SO-PMI 값	블로그	뉴스
표준어의 평균	-0.94	-0.62
표준어의 표준편차	2.57	6.43
이쁘	4.71	15.90
이뽀	1.55	-116.74
예뿌	3.14	-76.4886
조아하	-0.07	-116.74
산듯하	2.71	-64.45

표1. 비표준어에 대한 블로그와 뉴스에서의 SO-PMI값

표1은 비표준어에 대해 문서 집합을 블로그와 뉴스로 달리하면서 SO-PMI값을 비교한 것이다. 블로그에서는 비표준어의 SO-PMI값을 비교적 올바르게 구할 수 있었지만, 뉴스에서는 SO-PMI값이 어휘의 의미 극성을 제대로 반영하지 못하였다. 비표준어는 뉴스에서 잘 쓰이지 않으므로, 이 결과는 문서 집합이 특정 어휘의 쓰임새를 잘 반영하지 못하면 SO-PMI값을 올바르게 구하기 힘들다는 것을 보여준다.

또 블로그와 같이 어휘의 쓰임을 잘 반영하는 문서집합을 선택한 경우, '이쁘', '조아하', '산듯하' 등의 비표준어에 대해서도 타당한 SO-PMI값을 구할 수 있었다. 이는 우리가 제시한 방법이 비표준어에 대해서도 올바른 의미 극성을 찾을 수 있다는 것을 보여주는데, 이는 어휘망을 사용한 방법에서는 불가능한 것이다.

6. 결론 및 향후 과제

본 논문에서는 우리말 어휘의 극성을 판별하고자 하였다. 이를 위하여 SO-PMI를 사용하였는데, 이를 실제 우리말에 적용해보는 것은 처음이라는 점에서 의의를 갖는다. 또 이를 적용할 때 생기는 여러 문제점들인 기준 어휘와 문서 집합의 선택, 우리말의 특성을 논하고 이를 해결하는 방법을 제시했다는 점에서도 의의가 있다. 또한 실험을 수행할 때 실제 상품평을 수집하는 수집기를 구현하였고, 수집된 상품평을 형태소 분석기를 이용하여 많이 쓰이는 형용사들에 대하여 실제 적용해보았다. 그 결과 81%의 정확도로 어휘의 의미 극성을 얻어내었다.

향후 과제로서 본 연구에서 제시한 방법을 통해 의미 극성 정보를 담고 있는 대용량 사전을 구축하고 이를 쉽게 사용할 수 있도록 하는 사용자 환경과 오픈 API와 같은 인터페이스를 제공하려고 한다. 또한 실제 상품평을 문서 집합으로 사용하여 본 논문에서 제시한 방법을 적용해보는 연구를 진행 중이다.

참고문헌

[1] 명재석, 이동주, 이상구. "반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템", 정보과학회 논문지: 소프트웨어 및 응용 제 35권 제 6호, pages 392-403, 2008.

[2] P. Turney and M. Littman. "Measuring praise and criticism: Inference of semantic orientation from association", Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pages 417-424, 2002.

[3] Dongjoo Lee, Ok-ran Jeong and Sang-goo Lee "Opinion Mining of Customer Feedback Data on the Web", Proceedings of ICUIMC-08, The Second International Conference on Ubiquitous Information Management and Communication, 247-252, 2008.

[4] V. Hatzivassiloglou and Kathleen R. McHeown. "Predicting the semantic orientation of adjectives", Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174-181, 1997.

[5] Marco Baroni and Stefano Vegnaduzzo. "Identifying Subjective Adjectives through Web-based Mutual Information", Proceedings of KONVENS-04, 7th Konferenz zur Verarbeutung Natürlicher Sparache (German Conference on Natural Language Processing), pages 17-24, 2004.

[6] George A. Miller. "WordNet: A Lexical Database for English", Communications of the ACM (CACM), Volume 38 Number 11, pages 39-41, 1995.

[7] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. "Using WordNet to measure semantic orientation of adjectives", Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV, pages 1115-1118, 2004.

[8] Esuli and Sebastiani. "Determining the semantic orientation of terms through gloss classification", Proceedings of CIKM-05, 14th Conference on Information and Knowledge Management, pages 617-624, 2005.

[9] Jung-Yeon Yang, Jaeseok Myung and Sang-goo Lee. "The method for a summarization of product reviews using the user's opinion", Proceedings of ekNOW-09, international conference on Information, Process, and Knowledge Management, pages 84-89.

[10] Janyce Wiebe and Ellen Riloff. "Creating subjetctive and objective sentence classifiers from unannotated texts", Proceedings of CICLing-05, Conference on intelligent text processing and compuatational linguistics, p.486-497, 2005.

[11] http://sentiwordnet.isti.cnr.it/

[12] Chris Biemann, Sa-Im Shin and Key-Sun Choi. "Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences", Proceedings of COLING-04, 20th international conference on Computational Linguistics, pages 1227-1232, 2004.

[13] 윤애선, 황순희, 이은령, 권혁철. "한국어 어휘의미망 『KorLex 1.5』의 구축", 정보과학회논문지: 소프트웨어 및 응용 제36권 제1호, pages 92-108. 2009.

[14] http://kkma.snu.ac.kr