

사전을 사용한 주관성 어휘 번역 방법

김준기^o 남상협 이예하 이종혁

포항공과대학교 전자컴퓨터공학부 컴퓨터공학과
{yangpa^o, namsang, sion, jhlee}@postech.ac.kr

Conveying Subjectivity of a Lexicon of One Language into Another Using a Bilingual Dictionary

Jungi Kim^o Sang-Hyob Nam Yeha Lee Jong-Hyeok Lee

Department of Computer Science and Engineering
Division of Electrical and Computer Engineering
Pohang University of Science and Technology

요 약

인터넷 사용의 증가로 인터넷이 사용자의 의견 표출의 장이 되었다. 이에 따라 사용자의 견해나 의견을 자동으로 인식 및 추출하는 방법들이 연구되어 오고 있다. 의견 분석 (opinion analysis)은 한국어에서는 아직 연구가 활발히 되지 않는 분야로 의견 분석에 필요한 자원 및 도구들이 미비하다. 본 논문은 다른 언어권에서 구축된 주관성 어휘를 사전을 이용해 번역하는 방법을 제시하고 문제점 및 개선 방법과 향후 연구방향에 관하여 논의한다.

1. 서 론

웹의 진보로 인해 인터넷 사용자들의 전자출판이 글을 통한 표현욕구의 주요한 부분을 차지하게 되었다. 사용자의 글은 다양한 의견 표출의 장이 되었고 이에 따라 사용자의 견해나 의견을 자동으로 인식 및 추출하는 방법들이 연구되어 오고 있다. 이러한 연구들은 상업과 정치 등 다양한 분야에서 자동으로 의견을 수집하고 분석하고자 하는 욕구를 만족시킬 수 있다.

의견 분석 (opinion analysis) 분야는 주관성 (subjectivity) 분석, 감정 (sentiment 또는 polarity) 분석, 의견의 표출자 (opinion holder) 및 대상 추출, 더 나아가서는 의견의 요약이나 미래 시장 동향 예측 등의 응용분야 등 다양한 세부 주제로 나뉠 수 있다.

이러한 연구들이 영어나 일본어에서는 심도 있게 오랫동안 수행되어 왔고 중국어 등 다른 언어권에서도 활발한 연구가 진행되고 있다. 이에 따라 의견 분석에 필요한 다양한 자원 및 도구들이 개발되어 왔지만 아직 한국어에는 그러한 자원들이 미비한 실정이다.

다른 언어권에서 구축된 의견 분석에 쓰이는 다양한 자원 중 주관성 어휘 (subjectivity lexicon)는 의견분석 연구와 그 시작을 함께 하였다. 주관성 어휘란 주관적인 의미를 표현할 때 쓰이는 단어나 표현의 집합으로, 어휘들의 품사정보나 의미정보 이외에 일반적으로 어휘가 갖는 주관성의 세기 (strength), 긍정 및 부정 같은 주관성의 방향 (polarity)과 같은 정보를 포함한다.

본 논문은 주관성 어휘를 사전만을 이용해서 번역하는 몇 가지의 휴리스틱한 방법을 제시하고 문제점과 앞으로의 방향에 대해서 논의한다.

2. 관련 연구

주관성 어휘를 구축하려는 다양한 방법이 연구되어왔다. [1]은 대등 접속에서 형용사 사용의 언어적 제약과 자동지도학습 방법을 이용하여 형용사들의 주관성 방향을 학습하는 방법을 연구하였다.

[2]는 단어의 공기정보와 방대한 웹 페이지에서 공기정보를 제공하는 웹 정보 검색기를 이용하여 몇 개의 주관성 방향이 알려진 seed 단어들과의 Pointwise Mutual Information (PMI)를 이용해 주관성을 판별하는 실험을 하였다.

[3]은 주관성 방향이 알려진 단어들의 집합을 이용하여 단어의 동의어들이 더 많이 속하는 집합의 주관성 방향을 택하는 방법을 연구하였다.

영어 단어의 의미 조직인 WordNet을 이용하는 연구도 있었다. [4]는 단어들에 대해 “Good”과 “Bad”와의 의미가 얼마나 가까운지를 WordNet을 이용해 계산하여 단어의 주관성을 판별하는 방법을 선보였다. 최근에는 단어들의 사전 설명과 PageRank를 이용해 WordNet synset 간의 유사도를 계산하여 WordNet의 모든 synset에 주관성 속성을 학습하는 방법도 있었다 [5,6,7].

한국어를 대상으로 한 연구에는 말뭉치로부터

자동으로 주관성 후보를 추출하고 [8]의 appraisal theory 에 근거한 형태로 사용자가 입력할 수 있도록 한 [9]가 있다.

[10,11]들은 영어에 구축된 주관성 어휘를 다른 언어의 주관성 분석에 활용하는 방법을 제시하는데 그 방법들은 크게 세가지로 나뉜다. 첫 번째 방식은 주관성 어휘를 사전에 이용하여 대상 언어로 직접 번역하는 것이고, 두 번째는 대상언어의 문서를 주관성 어휘의 언어로 번역 하는 방식, 그리고 세 번째는 병렬 말뭉치와 영어 주관성 문장 분류기를 이용해 대상 언어에서의 학습 말뭉치를 구축하여 기계학습을 통한 분류기를 만드는 방식이다. [11]은 세 번째 방법이 성능이 뛰어나다고 보고하지만 이 방법은 학습 말뭉치나 분류기만 구축할 뿐 주관성 어휘를 생성하지 못하여 이 후의 의견문서 단계에 활용할 수 없게 된다.

본 논문은 [11]이 실험한 두 번째 방식을 위하여 몇 가지 휴리스틱한 번역 방법을 살펴보고 이 방식에 성능을 더 향상시킬 여지가 있는지 살펴어 본다.

3. 주관성 어휘 구축

한 언어에 구축되어있는 주관성 어휘를 언어간 사전과 휴리스틱을 이용해 다른 언어로 번역하는 방법을 고찰한다.

3.1. 주관성 어휘 번역

먼저 우리가 제시하는 주관성 어휘의 번역을 위한 수학적 표현을 제시한다. 출처 언어의 단어 W_s 의 주관성을 가질 확률을 $P(S|W_s)$, 대상언어의 단어 W_f 의 번역 후보를 $TC(W_f)=\{W_{s1}\dots W_{sn}\}$ 라고 표현할 때 W_f 가 주관성을 가질 확률은 다음과 같다.

$$P(S|W_f) = \sum_{W_s \in TC(W_f)} P(S, W_s | W_f)$$

여기서 $P(S, W_s | W_f)$ 를 다시 쓰면서 $P(S|\cdot)$ 에서 S와 독립적이라고 가정하는 W_f 를 생략해주면, 우리는 결국 다음과 같은 식을 얻을 수 있다.

$$P(S|W_f) = \sum_{W_s \in TC(W_f)} P(S|W_s)P(W_s|W_f) \quad (1)$$

수식 (1)은 크게 $P(S|W_s)$ 와 $P(W_s|W_f)$ 두 개의 확률로 이루어진다. $P(S|W_s)$ 는 W_s 가 가지는 주관성의 확률, 그리고 $P(W_s|W_f)$ 는 대상언어의 단어가 출처언어로 번역될 확률이다.

$P(S|W_s)$ 는 대상언어에 이미 구축되어있는 주관성 어휘에서 얻을 수 있다. 보편적인 주관성언어는 주관성의 강도를 “아주 세다”, “더 세다” 등의 불연속 변수나 확률 값과 같이 연속적인 변수일 수도 있다.

$P(W_s|W_f)$ 는 양 언어 사이의 사전을 이용해서 구하도록 한다. 번역 확률은 다음과 같은 휴리스틱들을 이용해 추정해보도록 한다.

First Sense (FS): FS는 가장 naïve한 방법으로 사전의 단어는 그 쓰임이 많은 의미의 순서대로 등재되어있다는 일반적인 양상에 근거한다. 가장 첫 번째에 등재된 의미의 단어를 선택하는 것은 오류를 적게 하는 확실한 방법이다. [12]는 의미 애매성 해소 시스템에서 첫 번째 단어를 선택하는 방법이 75%의 lower bound를 가질 수 있다고 예측했고 [13]은 다의성이 있는 단어만을 대상으로 58%의 성능을 가진다고 보고했다.

FS 는 다음과 같이 수식화할 수 있다.

$$P_{FS}(W_f | W_s) = \begin{cases} 1 & \text{if } SR(W_s, W_f) = 1 \\ 0 & \text{otherwise} \end{cases}$$

여기서 $SR(W_s, W_f)$ 는 W_s 의 후보단어 중 W_f 의 의미 순서를 나타낸다.

Uniform Weight (UW): UW는 첫 번째 의미만을 선택하는 FS와는 달리 모든 의미에 대해서 균등한 확률을 분배하는 방법이다. FS는 출처 언어의 한 단어 당 대상언어의 한 단어에만 대응하는 반면, UW는 가능한 모든 단어와 대응시키는 방법이다. $NS(W_f)$ 를 W_f 의 의미 개수라고 한다면 UW는 다음과 같다.

$$P_{UW}(W_s | W_f) = \frac{1}{NS(W_f)}$$

Sense Rank (SR1 & SR2): FS와 UW이 극단적인 방법이라면 SR은 그 중간을 택하는 방법으로 W_f 의 번역 후보 W_s 의미의 등재 순위를 이용해 점수를 차등화 하는 방법이다.

$$P_{SR1}(W_s | W_f) = \frac{NS(W_f) - SR(W_f, W_s) + 1}{\sum_{W'_s \in TC(W_f)} (NS(W_f) - SR(W_f, W'_s) + 1)}$$

$$P_{SR2}(W_s | W_f) = \frac{(NS(W_f) - SR(W_f, W_s) + 1)^2}{\sum_{W'_s \in TC(W_f)} (NS(W_f) - SR(W_f, W'_s) + 1)^2}$$

3.2. 대상 언어의 주관성 학습

수식 (1)의 방법으로 구축된 초기 어휘는 대상언어 단어의 주관성이 오직 그 단어가 번역될 수 있는 후보들을 이용하므로 후보의 수가 적은 단어의 경우는 그 점수의 신빙성이 떨어진다. 이러한 단점을 보완하기 위해 대상언어 단어의 주관성 확률을 반복형식으로 조절해가는 방법들을 제시한다.

$$S(W_f) = \sum_{W_s \in TC(W_f)} S(W_s) wt(W_f \rightarrow W_s) \quad (2)$$

$$S(W_s) = \sum_{W_f \in TC(W_s)} S(W_f) wt(W_s \rightarrow W_f) \quad (3)$$

PageRank(PR): PageRank는 random-walk model로 directed graph에서 authority있는 node를 찾는 방법이고 웹 문서를 대상으로 한 검색엔진에 많이 사용되고 있는 유명한 방법이다[14]. 우리는 식 (1)에서 주관성 확률을 점수로, 번역확률은 단어들 사이의 간선으로 대체할 수 있다. 모든 단어들이 graph의 node라고 하면 출처언어의 단어들과 대상언어의 단어들 사이에는 간선이 없고 오직 서로의 집합으로만 연결되므로 이 graph는 bipartite graph가 되고, 두 node 집합들의 주관성 확률은 (2), (3)의 식을 반복적으로 계산하면서 구할 수 있다. (rank sink를 위한 smoothing factor를 추가적으로 고려해야 한다. 우리는 이러한 수고를 덜고 신빙성이 덜한 단어들을 없애기 위해 outlink가 없는 node를 제거하였다) 하지만 이 방법은 iteration이 반복되면서 W_f 의 주관성 확률에도 변화를 가져오므로 적은 수의 iteration이 원래의 W_f 의 주관성 확률과 부합하는 적합한 결과를 낳으리라 예상한다.

EdgeWeightAdjustment (EWA): 이 EWA 방법은 PR과 같이 (2)의 식을 이용하지만 W_f 의 주관성 확률을 학습하는 대신 번역 확률인 간선의 가중치를 학습하도록 한다. 그래서 (3)의 식 대신에 아래의 식으로 대신한다.

$$wt(W_s \rightarrow W_f) = \frac{S(W_f)}{\sum_{W_f \in TC(W_s)} S(W_f)}$$

4. 주관성 분류기

주관성 어휘는 그 자체로 자동적인 객관적 분류가 힘들어 [15]에 소개된 주관성 어휘와 몇 가지 휴리스틱 규칙을 사용한 주관성 분류기를 구축하여 제안된 방법으로 구축된 어휘의 질을 판별하고자 한다. [15]에

고무되어 [11]은 그들의 주관성 어휘를 평가하기 위하여 규칙기반의 주관성 분류기를 구축하였다. 분류기의 대략적인 규칙은 다음과 같다.

- 1) 2개 이상의 **강한** (strong) 주관성 단어가 등장하는 문장은 **주관적인** (Subjective) 문장이다.
- 2) **강한** 주관성 단어가 하나도 없고 앞뒤, 현 문장에 **약한** (weak) 주관성 단어가 2개 이하이면 그 문장은 **객관적** (Objective) 문장이다.
- 3) 두 가지 조건을 만족시키지 못하면 그 문장의 주관성 판별을 **보류**한다 (Unknown).

이와 같은 기존 평가방법을 따라 비슷한 분류기를 구축하는데 문장 판별 시 필요한 조건을 약간 변경하여 다음과 같이 한다.

- 1) 2개 이상의 **강한**, 또는 3개 이상의 **약한** 주관성 단어가 등장하는 문장은 **주관적인** 문장이다.
- 2) **강한** 주관성 단어가 하나도 없고 앞뒤, 현 문장에 **약한** 주관성 단어가 3개 이하이면 그 문장은 객관적 문장이다.
- 3) 두 가지 조건을 만족시키지 못하면 그 문장의 주관성 판별을 **보류**한다.

5. 실험 및 토의

본 논문의 실험은 주관성 어휘 구축과 주관성 문장의 분류의 두 가지로 구성되어있다. 실험에서 출처언어는 의견 분석 자원이 풍부한 영어이고 대상언어는 한국어이다.

5.1. 실험 자료

실험에서 사용하는 영어로 구축된 주관성 어휘는 Automatic Opinion Finder System¹ [15]의 것이다. [15]는 기존 연구에서 구축된 자원을 이용해서 주관성 어휘를 수작업으로 수집하였다. 이 자원은 6,856의 영어 단어와 단어들의 품사, 강도 (Strong/Weak), 그리고 극성 (polarity: Positive/Negative)으로 구성되어있다.

표 1. 평가 문서의 511문장에 대한 주관성 일치율.

| 주석자 쌍 | 일치율 | 카파 (κ) |
|-----------------------------------|------|--------|
| An ₁ & An ₂ | 0.74 | 0.47 |
| An ₁ & An ₃ | 0.80 | 0.59 |
| An ₂ & An ₃ | 0.70 | 0.41 |

¹ <http://www.cs.pitt.edu/mpqa/>

표 2. 주관성 문장 분류기의 성능.

| | 정확률 | | | | 재현율 | | | | F-값 | | | |
|---|--------------|--------------|--------------|--------------|--------------|-------|--------------|--------------|--------------|-------|--------------|--------------|
| | FS | UW | SR1 | SR2 | FS | UW | SR1 | SR2 | FS | UW | SR1 | SR2 |
| 번역 휴리스틱만을 이용한 경우 | | | | | | | | | | | | |
| 주관적 | 72.41 | 67.25 | 66.85 | 66.48 | 37.06 | 67.65 | 70.00 | 68.82 | 49.03 | 67.45 | 68.39 | 67.63 |
| 객관적 | 64.49 | 79.07 | 76.74 | 77.27 | 46.62 | 22.97 | 22.30 | 22.97 | 54.12 | 35.60 | 34.55 | 35.42 |
| 모두 | 68.04 | 69.63 | 68.78 | 68.64 | 41.51 | 46.86 | 47.80 | 47.48 | 51.56 | 56.02 | 56.40 | 56.13 |
| 휴리스틱 번역 후 PR방법 적용한 경우 (2 iterations) | | | | | | | | | | | | |
| 주관적 | 75.90 | 66.86 | 67.43 | 67.24 | 37.06 | 67.65 | 69.41 | 68.82 | 49.80 | 67.25 | 68.41 | 68.02 |
| 객관적 | 64.22 | 74.42 | 73.81 | 75.00 | 47.30 | 21.62 | 20.95 | 22.30 | 54.47 | 33.51 | 32.63 | 34.38 |
| 모두 | 69.27 | 68.37 | 68.66 | 68.81 | 41.82 | 46.23 | 46.86 | 47.17 | 52.16 | 55.16 | 55.70 | 55.97 |
| 휴리스틱 번역 후 EWA방법 적용한 경우 (100 iterations) | | | | | | | | | | | | |
| 주관적 | 73.91 | 67.39 | 73.91 | 68.62 | 20.00 | 36.47 | 20.0 | 41.17 | 31.48 | 47.32 | 31.48 | 51.47 |
| 객관적 | 56.54 | 63.10 | 56.54 | 65.59 | 64.18 | 43.91 | 64.18 | 41.21 | 60.12 | 51.79 | 60.12 | 50.62 |
| 모두 | 60.28 | 65.12 | 60.28 | 67.17 | 40.56 | 39.93 | 40.56 | 41.19 | 48.49 | 49.51 | 48.49 | 51.07 |

번역을 위한 사전으로는 네이버 영어 사전²이 쓰였다. 포털사이트들이 제공하는 사전들이 대체로 인터넷에서 무료로 구할 수 있는 영한/한영 사전들에 비해서 등재되어있는 단어의 개수나 품질 면에서 훨씬 월등하다고 판단된다.

성능 검증을 위한 평가 문서로는 2002년 7월부터 12월 사이의 동아일보³ 기사 중 515개의 문장으로 구성된 35개의 문서이다. 이 문장들은 한국어를 모국어로 하는 세 명의 주석자에 의해 주관적/객관적으로 표시되었다. 세 명의 주관성 일치도를 비교하기 위해 annotation study를 해보았다. 일치도는 일치율과 카파(Kappa) 값[16]으로 측정하였다. 표 1에 나온 세 주석자들의 주관성 일치율 중에 두 번째 주석자의 일치율과 카파 값이 특히 매우 낮다. 이는 주관성 분류의 기준이 정하기 어렵고 개인마다 차이가 있음을 보여준다.

최종적으로 사용하는 평가문서는 511개의 문장 중 세 명의 주석자가 모두 동의하는 148개의 객관적, 170개의 주관적, 총 318개의 문장으로 구성하였다.

5.2. 주관성 문장 분류기 성능

분류기의 성능은 주관적 문장 분류성능과 모든 문장 분류성능을 중심으로 분석하는데 이유는 주관적 어휘의 직접적인 성능 향상은 주관적 문장 분류에 반영이 된다고 생각하기 때문이다. 제안된 휴리스틱 방법들만 사용했을 경우에는 FS경우 예상대로 주관적 문장의 경우 정확률은 높은 성능을 보이고 대신 재현율이 현저히 낮았다. 반대로 그 외의 방법들 UW, SR1,

SR2은 정확률이 FS에 10%정도 뒤지지만 반면 재현율은 30% 이상의 향상을 보였다. F-값의 경우 UW, SR1, SR2의 방법들이 FS보다 일관성 있게 높다.

PR이나 EWA 방법을 적용하는 경우, 정확률은 대부분의 경우 향상하였고 주관적 문장 분류에서 특히 FS의 경우 PR을 사용하여 75.90%, SR1은EWA를 사용하여 73.91%의 정확률을 보였다. PR 방법은F-값을 거의 저하 하지 않으면서 주관적 성능을 소폭 향상시키는 반면 EWA 방법은 정확률을 높이는 반면 재현율을 현저히 희생하여 F-값이 전체적으로 많이 낮아졌다. 본 논문에서 제안한 방법이 성능을 거스르지는 않는다고는 판단이 되지만 그 성능의 폭이 현저하여 큰 의미를 보이지 않는다고 판단이 된다.

6. 결론 및 향후 연구

본 논문은 한국어 의견 분석 시스템의 개발을 위한 첫 단계로 다른 언어에서의 자원을 활용할 수 있는 방안의 모색으로 주관성 어휘의 번역방법에 관해 고찰해 보았고, 몇 가지 단순한 휴리스틱과 개선 방법 등을 통해 한국어 어휘를 구축해 보았다. 또한 이 어휘들을 주관성 문장 분류기에 적용해 간접적으로 품질을 평가해 보고 그 결과 여러 방법들의 장단점을 파악하게 되었다.

이 연구에 쓰인 방법을 토대로 안정된 주관성 어휘 번역 방법을 개발하고 만들어진 어휘의 성능평가를 더 정확하게 할 수 있는 방법을 고찰해야 한다.

미래의 연구로는 기존 연구에 언급된 병렬 코퍼스나 웹의 상품평등으로부터 쉽게 수집할 수 있는 주관성 코퍼스를 활용하는 방법을 연구할 필요가 있다.

² <http://endic.naver.com/>

³ <http://www.donga.com/>

감사의 글

본 논문은 2008년도 두뇌한국21사업의 지원을 받았고 지식경제부 및 정보통신진흥연구원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고문헌

[1]Vasileios Hatzivassiloglou and Kathleen R., McKeown, Predicting the semantic orientation of adjectives, ACL'97, 174-181, 1997.

[2]Peter D. Turney and Michael L. Littman, Measuring praise and criticism: Inference of semantic orientation from association, TOIS, 21(4), 315-346, 2003.

[3]soo-Min Kim and Eduard Hovy, Determining the sentiment of opinions, COLING'04, 1367-1373, 2004.

[4]Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke, Using WordNet to measure semantic orientation of adjectives, LREC'04, 1115-1118, 2004.

[5]Andrea Esuli and Fabrizio Sebastiani, Determining the semantic orientation of trens through gloss analysis, CIKM'05, 617-624, 2005.

[6]Andrea Esuli and Fabrizio Sebastiani, SentiWordNet: A publicly available lexical resource for opinion mining, EACL'06, 193-200, 2006.

[7]Andrea Esuli and Fabrizio Sebastiani, PageRanking WordNet synsets: an application to opinion mining, ACL'07, 442-431, 2007.

[8]Casey Whitelaw, Navendu Garg, and Shlomo Argamon, Using appraisal groups for sentiment analysis, CIKM'05, 625-631, 2005.

[9]명재석, 이동주, 이상구, 반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템, 제19회 한글 및 한국어 정보처리 학술대회, 68-75, 2007.

[10]Soo-min Kim and Eduard Hovy, Identifying and analyzing judgment opinions, HLT/NAACL'06, 2006.

[11]Rada Mihalcea, Carmen Banea, Janyce Wiebe, Learning Multilingual Subjective Language via Cross-Lingual Projections, ACL'07, 2007.

[12]William Gale, Kenneth W. Church, David Yarowsky, Estimating upper and lower bounds on the performance of word-sense disambiguation programs, ACL'92, 249-256, 1992.

[13]George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, Robert G. Thomas, Using a semantic

concordance for sense identification, HLTW'94, 240-243, 1994.

[14]Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN systems, 30(1-7), 107-117, 1998.

[15]Janyce Wiebe and Ellen Riloff, Craeting subjective and objective sentence classifiers from unannotated texts, CLCLing'05, 2005.

[16]Jean Carletta, Assessing Agreement on Classification Tasks: The Kappa Statistic, Computational Linguistics, 19(2), 263-311, 1996.