독일언어문학 제65집(2014.9). 27-46

바이그램을 활용한 텍스트 논조자동분석

홍 문 표(성균관대)

I. 서론

우리는 어떤 주제에 대해 언어적 수단으로 감정을 표출할 때 (1)의 예문과 같이 가치평가와 관련된 어휘를 선택하거나 (2)의 예문에서처럼 감정과 관련된 어휘나 비속어를 사용하기도 한다. 직접적인 감정표출이 불편할 경우에는 비유나 비교 등을 통해 간접적으로 감정을 표현하기도 하며 (3)과 같이 부정어 Negation를 사용하여 완곡하게 감정을 표현한다.

(1) Ich verbrachte diesmal vier Nächte hier und es war schrecklich.
(2) Es tut mir leid sagen zu mussen dass ich enttäuscht war und diesen Ort nicht empfehlen kann.[1)]
(3) Man kann hier nicht in Ruhe schlafen.

어떤 주제에 대한 언어표현의 논조 Sentiment는 언어표현을 구성하는 단어의 외연 Denotation 및 공시의 Konnotation 이외에도 단어의 통사적 결합관계 Syntaktische Relation 등에 의해 결정된다. 문장에 드러난 논조나 감정을 자동으로 분석 및 분류하는 연구분야를 논조분석 Sentimentanalyse 또는 의견마이닝 Opinion Mining이라고 부른다.

논조분석을 위한 기존의 연구는 대부분 단어의 외연 및 공시의 정보가 기술된 감성사전 Sentiment-Wörterbuch의 엔트리를 기계학습 Maschinelles Lernen의 자질 Merkmal로 활용한 방법론이 주류를 이루었다. 독일어 논조분석의 경우도 'Senti-

1) 위 예문은 실제 독일 인터넷 게시판 등에서 수집한 예문으로 문장부호 사용 등의 오류가 있으나, 원문을 그대로 옮겼음을 밝힘.

Wortschatz'[2]와 'German Polarity Clues'[3] 등과 같은 감성사전의 엔트리를 기계학습의 자질로 활용한 방법이 연구의 주류를 이루었다.

감성사전 및 학습코퍼스 Lernkorpus의 단어들을 기계학습 자질로 활용하는 방법은 단순히 긍/부정 단어의 개수만을 파악하여 문장의 논조를 결정하는 방법론보다는 월등히 높은 성능을 보이나, 단어와 단어들 간의 관계가 고려되지 못하는 문제를 가지고 있다. 이러한 문제를 해결하기 위하여 문장의 통사구조를 기계학습에 반영하는 방법론도 최근 활발히 연구되고 있다. Pang/Lee (2004), Wilson et al. (2005), Moilanen/Pulman (2007), Choi/Cardie (2008), 홍문표 (2013a), 홍문표 (2013b) 등의 연구는 통사관계를 고려한 논조분류 방법론을 제안하고 있다. 이들의 연구에서는 공통적으로 입력문장에 대해 통사분석을 수행한 후, 통사분석결과를 기계학습 내지는 의미계산을 위한 입력으로 활용하였다.

그러나 입력텍스트의 종류가 통사구조분석에 적합하지 못한 경우 이러한 방법론은 그 한계를 드러낸다. 예를 들어 트위터 Twitter 텍스트나 인터넷 게시판 텍스트들과 같이 철자오류가 빈번하고 문법규칙에도 위배되는 문장이 많은 경우에는 통사구조분석에 기반한 방법론은 그 한계를 드러내기 쉽다. 현존하는 어떠한 구조분석기도 철자오류와 문법오류가 포함된 문장을 작성자의 의도를 파악하여 정확하게 구조분석을 해낼 수는 없다.[4] 잘못된 문장구조 정보를 기계학습에 반영하여 논조분석을 수행하면 단순히 문장에 출현하는 단어정보만을 활용한 분류방법보다 성능이 좋다는 보장이 없다. 따라서 본 연구에서는 통사구조분석을 하지 않고서도 문장의 통사구조를 어느 정도 고려할 수 있는 방법론을 제안하고자 한다.

이 방법론은 자연언어처리분야에서 언어모델링 Sprachmodellierung을 위해 활발히 사용되고 있는 바이그램 Bigramm 기반의 방법론이다. 본 연구에서는 독일어 문장의 논조분석을 위해 바이그램 정보가 성공적으로 활용될 수 있음을 보일 것이다. 그리고 바이그램 Bigramm, 트라이그램 Trigramm 중 논조분석에 보다 성공적으로 활용될 수 있는 자질이 무엇인지 살펴보게 될 것이다. 또한 이 방법론이 기존의 유니그램 Unigramm기반의 방법론과 비교하여 어느 정도의 성능 향상

2) Remus et al. (2010)
3) Waltinger (2010)
4) 슈투트가르트 대학의 의존문법 파서를 사용하여 본 연구의 실험코퍼스를 대상으로 문장구조분석을 시도한 결과 47.5%의 파싱정확률만을 나타내었다.

에 기여할 수 있는지 실험을 통해 알아볼 것이다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 논조분석시 통사구조 등과 같은 문맥정보를 반영하기 위한 기존 연구를 소개한다. 여기서는 바이그램을 활용한 통사구조 반영 방식과 통사분석에 직접 기반한 논조분석 방식을 소개한다. 3장에서는 언어모델링을 위해 바이그램이 어떻게 활용될 수 있는지에 대해 논의한다. 4장에서는 본 연구에서 제안하는 방법론의 성능을 파악하기 위한 실험을 수행한다. 또한 이어서 실험결과를 분석할 것이다. 끝으로 5장에서는 본 연구의 가능성과 한계에 대해 논의하며 향후 필요한 연구에 대해 언급하며 논문을 끝맺을 것이다.

II. 관련연구

II.1. n-그램 기반 논조분석

Pang et al. (2002)의 연구에서는 기존의 주제별 텍스트분류 방법론을 논조분석에도 적용하여 성공적인 결과를 얻을 수 있음을 보였다. 이들이 주제별 텍스트분류를 위해 활용한 방법론은 학습코퍼스에 등장하는 유니그램을 기계학습을 위한 자질로 사용하는 방법이다. 이들의 실험결과는 <표 1>에 나타나 있다.[5] 그러나 논조분석은 이들이 스스로 밝힌 바와 같이 주제별 텍스트분류와는 다른 어려운 점들이 많이 존재한다. 그래서 분석 정확도도 주제별 텍스트분류와 비교하여 논조분석이 약 10% 정도 낮다.

이들이 지적한 논조분석의 어려운 점 중의 하나는 문장에 존재하는 부정어의 처리이다. 영어의 경우 'not', 'never' 등과 같은 부정부사어들은 피수식어의 긍/부정 정보를 뒤바뀌게 하는 역할을 하는 단어들이다. 이들의 방법론은 구조분석 정보를 활용하지 않으므로, 문장내에 'not', 'never' 등과 같은 부정부사어가 등장하면 그 다음 단어부터 문장의 끝까지 모든 단어들의 긍/부정 정보를 뒤바꾼다. 즉, '긍정' 논조의 단어는 '부정'으로 계산되고, '부정' 논조의 단어는 '긍정'으로 계산된다.

---

5) <표 1> 실험결과의 정확률은 문서단위의 분석정확률로서, 문장단위의 분류를 시도한다면 그 정확도는 훨씬 낮아질 것으로 예상된다.

<표 1> : Pang et al. (2002)의 영어문서 논조분석 정확률

	자질의 수	빈도/ 출현여부	NB[6]	ME[7]	SVM[8]
유니그램	16,165	빈도	78.7	알려지지 않음	72.8
유니그램	16,165	출현	81.0	80.4	82.9
유니그램+ 바이그램	32,330	출현	80.6	80.8	82.7
바이그램	16,165	출현	77.3	77.4	77.1
유니그램+ 품사정보	16,695	출현	81.5	80.4	81.9

Pang et al. (2002)의 실험결과에서 한 가지 주목할만한 점은 유니그램만을 사용하여 기계학습을 시도한 경우가 유니그램과 바이그램을 모두 고려한 경우보다 'ME' 알고리즘을 사용했을 때를 제외하고서는 모두 분석 정확률이 높았다는 점이다. 이들의 실험에서 가장 높은 분석 정확률을 기록한 것은 유니그램만을 자질로 사용하여 'SVM' 알고리즘을 통한 기계학습을 수행하였을 때이다.

이 실험결과의 이유를 정확하게 알아내기는 불가능하지만 유니그램과 바이그램의 자질수를 지나치게 많이 사용했기 때문일 것으로 추정된다. 즉, 유니그램만을 활용하였을 때는 16,165개의 자질을 사용하였고, 유니그램과 바이그램을 모두 활용한 경우에는 정확하게 두 배인 32,330개의 자질을 사용하였다. 이렇게 기계학습에서 지나치게 많은 자질을 사용할 경우, 'overfitting'의 문제가 발생하여 더 적은 수의 자질만을 사용할 때보다 오히려 성능이 떨어지는 현상은 매우 자주 발생한다.

6) Naive Bayesian
7) Maximum Entropy
8) Support Vector Machine

II.2. 의미분석기반 논조분석

문장의 논조를 계산하는 또 다른 하나의 방법은 전통적인 합성적 의미도출 compositional semantic derivation 방법이다. 몬테규 의미론의 의미계산 방식과 거의 흡사한 이 방법은 Moilanen/Pulman (2007)의 연구에서 제안되었다. 이 방법은 논조분석사전을 기반으로 하여, 통사규칙에 의미해석규칙을 부착하여 문장의 논조를 계산한다.

(4) OUTα ij → SPRα i + SUBα j
Mod:Adj + Head:N

의미해석규칙은 다시쓰기규칙 Rewriting Rule의 형식으로 작성된 통사규칙에 일대일로 부착되어 있다. (4)의 의미해석규칙은 형용사와 명사가 결합하여 NP내지는 N'의 구조를 만드는 통사규칙에 부착되어 있다. 의미해석규칙도 통사규칙과 마찬가지로 다시쓰기규칙의 형식을 띄고 있으며 의미결합의 핵심어 Kopf는 밑줄로 표시가 되어 있다. 즉, (4)의 예시 규칙에서 수식어의 역할을 하는 형용사는 의미결합관계의 핵심어 역할을 하고 논조정보가 어미 노드로 전승되어 최종적으로 만들어진 구조의 논조는 핵심어인 형용사의 논조와 같게 된다.

이와 같은 방식으로 문장 전체의 논조를 문장의 구조분석에 적용되는 통사규칙 및 통사규칙에 부착된 의미해석 규칙에 의거해 계산해낼 수 있다.

이 논조분석 방식은 문장의 통사구조분석이 정확하게 이루어졌다는 가정하에 성공적으로 적용될 수 있는 방식이다. 만약 통사구조의 분석과정에서 오류가 발생했다면 논조분석규칙도 잘못 적용될 것이기 때문에 통사구조 분석의 정확률이 떨어지는 경우에는 이 방식을 적용하기는 어렵다. 트위터 문장과 같이 비정형 문장이 많이 등장하는 경우에는 통사구조분석의 정확률이 50% 미만이므로 의미분석기반의 논조분석방식을 적용하는 것이 성공적이지 못할 것이다.

III. n-그램 기반 언어모델링

III.1. n-그램

유니그램을 자질로 하는 논조분석 방법론은 'bag of words' 방식이라는 별칭이 말해주듯이 단어들 간의 관계 또는 문맥을 전혀 고려하지 않는 단점이 있다. 기존연구에서 확인한 바와 같이 문장의 극성을 결정해주는 단어들이 어떤 문맥에서 사용되었느냐에 따라 문장 전체의 극성이 달라질 수 있다. 따라서 기존의 연구들은 문장논조의 결정시 문맥을 반영하기 위해 구조분석 등과 같은 방법을 사용하였으나 문장의 자동분석 정확률이 보장되지 못하는 환경에서는 그 한계가 있다.

본 연구에서는 이 문제를 극복하기 위한 하나의 방법으로 음성인식, 정보검색, 통계기반 기계번역 등에서 언어모델링 Sprachmodellierung을 위해 널리 사용되는 방법인 n-그램 방법론을 제안한다. 논조분석을 위해 널리 사용되었던 기존의 유니그램도 사실상 n=1인 경우의 n-그램이므로 이 방법이 새로운 방법은 아니지만, 본 연구에서 제안하는 방법은 n=2 또는 n=3이 되는 바이그램 내지는 트라이그램 기반의 방법론이다.

기존의 유니그램 기반의 방법론은 학습코퍼스를 최종적으로는 'bag of words'의 형태로 변환하여 학습코퍼스에 출현하는 유니그램과 그것의 출현여부 내지는 빈도간의 벡터형태로 활용한다. 따라서 어떤 문장이 출현할 확률은 그 문장을 구성하는 모든 단어의 출현확률의 곱이 된다. 그러나 사실상 어떤 언어에서건 문장을 구성하는 단어의 출현확률은 앞선 단어연쇄의 확률에 의존적이다. 예를 들어 독일어에서 형용사 'schönes'가 등장할 수 있는 환경은 'ein', 'sehr', 'kein' 등과 같은 단어의 다음으로 한정되지, 'das', 'der', 'dem', 'den', 'die' 등의 다음에는 나올 수 없다. 이렇듯 한 문장에서 어떤 단어의 출현확률은 해당 문장에서 그 단어의 앞에 나온 단어 연쇄의 확률을 조건으로 하여 해당 단어가 출현하는 조건확률로 이해될 수 있다.

w1에서 wn까지 n개의 단어로 이루어진 문장의 확률은 맨 처음 단어가 출현할 확률과 그 단어가 출현한 조건하에 그 다음 단어가 출현하고, 마찬가지로 두 단어가 출현한 조건하에 세 번째 단어가 출현하여 최종적으로 첫 번째 단어부터

n-1 번째 단어가 연쇄적으로 출현한 조건하에 n번째 단어가 출현할 확률이다. 이를 수식으로 나타내면 (5)와 같다.

(5) $P(w_1^n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1^2) \ldots \cdot P(w_n|w_1^{n-1})$

그러나 위의 수식에서 현실적으로 아무리 큰 코퍼스라 할지라도 $P(w1^{n-1})$의 확률을 구하는 것은 큰 의미가 없다. 왜냐하면 이러한 단어연쇄의 빈도는 0 내지는 매우 낮을 것이기 때문이다.

위의 수식보다 현실적인 문장의 확률을 구하는 수식은 마르코프 가정 Markov Assumption에 의해 아래의 수식 (6)으로 재작성 될 수 있다.

(6) $P(w_n|w_1 w_2 \ldots w_{n-1}) \approx P(w_n|w_{n-1})$

마르코프 가정이란 한 상태의 확률은 단지 그 이전 상태에만 의존한다는 것으로 이를 문장의 확률계산에 적용하면 어떤 단어가 나올 확률은 단지 그 직전 단어가 출현할 확률에만 의존한다는 것을 의미한다.[9)]

바이그램은 결국 두 단어의 연속출현 확률에 기반하는 것이고 트라이그램은 세 단어의 연속출현 확률에 기반하여 문장의 출현확률을 계산하는 방법이라고 볼 수 있다. 바이그램을 기반으로 학습코퍼스를 모델링할 경우 독일어의 특성상 부정어 Negationswort가 피수식어와 인접하여 나오는 구조적 특징이 그대로 반영되어 통사적 문맥에 따른 문장 극성의 변화현상이 유니그램 기반의 모델보다는 훨씬 잘 반영될 것으로 기대된다. 이러한 방법론이 유니그램 기반의 모델과 비교하여 실제로 어느 정도의 분석성능의 향상을 가져오는지는 4장의 성능평가에서 다루기로 한다.

III.2. 학습코퍼스

n-그램을 자질로 활용한 기계학습 기반의 논조분석기를 개발하기 위해 본 연

9) 신효필 (2009:183)

구에서는 1천 문장 규모의 학습코퍼스를 구축하였다. 학습코퍼스의 도메인은 호텔예약 분야였으며, 코퍼스의 출처는 독일 호텔예약 사이트인 트립어드바이져(tripadvisor.de) 사이트였다. 이 사이트에는 세계의 주요 관광지에 대한 호텔예약 정보 및 사용자들의 평가가 실시간으로 게시된다. 본 연구에서는 '태국여행'이라는 주제에 대한 독일 사용자들의 호텔이용후기 텍스트를 코퍼스 구축의 소스로 택하였다.

코퍼스는 총 1천 문장으로 구성되어 있으며 500개의 긍정논조 문장과 500개의 부정논조 문장이 포함되어 있다. 트립어드바이져 사이트에서 사용자들은 본인이 투숙했던 호텔에 대해 '만족', '불만족' 등에 대한 평점을 부여한 후 구체적으로 만족스러웠던 내용이나 불만족스러웠던 점 등을 작성한다. 평점이 낮은 호텔의 사용후기에는 주로 불만족스러웠던 내용이 등장하고, 평점이 높은 호텔의 사용후기에는 주로 만족스러웠던 내용이 등장하므로 이를 문장 논조의 분류를 위해 참조하였다.

코퍼스 구축을 위한 개별 문장 논조의 긍부정 분류는 저자가 직접 수행하였다. 문장 논조의 긍부정 분류 기준을 명확하게 제시하기가 쉽지 않고 이에 대한 연구도 영어에 대해서는 일부 수행된 바 있으나, 본 연구의 목적은 n-그램의 활용방안을 알아보는 것이므로 본 연구에서는 학습코퍼스 구축을 위한 문장논조의 분류방안은 따로 다루지 않는다.[10)]

학습코퍼스는 문장과 문장의 논조로 구성되었다. 문장의 논조는 긍정의 논조일 경우 'p'태그가 사용되었고, 부정의 논조 경우에는 'n' 태그가 사용되었다.[11)] 학습코퍼스의 일부를 소개하면 다음과 같다 (7~16).

(7) Als ich mich über das Essen fragte die Rezeptionistin nur mit den Schultern gezuckt und ignorierte mich., n
(8) Als wir das nicht einsehen wollten wurde er laut und beinahe handgreiflich!, n
(9) Bett macht störende Geräusche wenn man sich bewegt, n

10) 논조분석을 위한 학습코퍼스 구축에 대한 대표적인 연구로는 Wiebe et al. (2005)을 들 수 있다.
11) 본 연구에서는 한 문장 내에 긍정 논조와 부정 논조가 혼합되어 있는 경우나 논조 정보가 없이 객관적인 사실만을 전달하는 경우는 배제하였다.

(10) Bin von A-Z entauscht von diesem Hotel, n
(11) Das Personal war nicht besonders freundlich oder hilfsbereit., n
(12) Er zeigte uns unser Appartement und wir waren sehr beeindruckt., p
(13) Habe den perfekten Schlafkomfort nach dem langen Flug genossen., p
(14) In der Nahe des Hotels findet man leckere Kleinkuchen und auch ein kleines Shopping Center mit klassischen Restaurants., p
(15) obwohl ich eigentlich weniger mit dieses Mittelklasse Hotelketten habe hat mich doch dieses sehr positive Ueberrrascht., p
(16) Vielen Dank an alle für diesen unvergesslichen Urlaub sehr angenehm und wir werden im nächsten Jahr wiederkommen., p

개별 문장은 논조태그 (p 또는 n)와 컴마를 구분자로 하여 분리되어있다. 컴마를 구분자로 사용하는 것은 웨카를 사용하여 기계학습을 수행하기 위해 알맞은 포맷으로 변환하기 위해서이다.

IV. 성능평가

IV.1 실험

본 논문에서 제안하는 바이그램 기반 논조분석 모델의 성능을 파악하기 위해 유니그램 기반의 논조분석 모델과의 비교평가를 실시하였다. 또한 트라이그램 기반의 논조분석 모델과도 비교평가를 수행하였다. 이를 위해 앞서 소개한 1천 문장 규모의 학습코퍼스를 각 100문장씩 총 10개의 세트로 구분하여, 9개의 세트로 기계학습을 수행하여 논조분석 분류기를 만들고, 이 분류기로 나머지 1개의 세트를 논조별로 분류한 후 정확성을 평가하며, 이를 총 10회 반복하는 '10-fold' 방식의 실험을 수행하였다.

먼저 본 연구결과의 비교대상이 되는 베이스라인 시스템으로 Pang et al. (2002) 방식의 유니그램만을 자질로 하는 논조분류기 Sentiment Classifier를 웨카시스템을 통해 구현하였다. 유니그램 기반의 논조분류기를 구현하기 위해서는 우선 학

습코퍼스를 웨카시스템이 인식할 수 있는 ‘ARFF’ 파일 형식으로 변환해야하는데 이는 웨카시스템의 빌트인 built-in 함수인 ‘StringToWordVector’를 활용하였다.

‘StringToWordVector’ 함수를 적용할 때는 몇 개의 파라메터를 조정해주어야 하는데 대표적으로 ‘tf (term frequency)’, ‘idf (inverse document frequency)’의 적용 여부, 대소문자 정규화 여부 등이 이에 해당한다. 또한 입력 문장을 단어 단위로 쪼개는 ‘Word Tokenizer’를 사용하였다. 이 함수를 적용하게 되면 학습코퍼스는 자질-값 행렬의 형식으로 변환된다. 예를 들어 학습코퍼스에 (7), (8)번의 문장이 아래와 같은 형태로 저장되어 있다고 하면, 이 문장들은 ‘StringToWordVector’ 함수의 적용 후 <표 2>와 같이 변환된다.

(7) Als ich mich über das Essen fragte die Rezeptionistin nur mit den Schultern gezuckt und ignorierte mich. , n

(8) Als wir das nicht einsehen wollten wurde er laut und beinahe handgreiflich! , n

<표 2> : 자질-값 행렬 형식의 의미표상

id	class	als	beinahe	das	den	die	er	essen	fragte	...
1	n	1	0	1	1	1	0	1	1	...
2	n	1	1	1	0	0	1	0	0	...
3	...	...	...	...	...	...	...	...	...	...

위와 같이 학습코퍼스를 자질-값 행렬의 형식으로 변환한 결과 총 3,327개의 자질이 추출되었다. 이 중 논조분석과 직접적인 연관이 없을 것으로 판단되는 아라비아 숫자 및 하이픈과 같은 단순 기호는 제거하였다. 이 과정에서 숫자가 포함되어 있는 복합명사 (예를 들어 ‘2-Zimmerappartment’, ‘4 Sternehotel’) 등은 제거하지 않았으며, 숫자와 기호만으로 이루어져 있더라도 그것이 고유명사를 나타내는 경우 (예를 들어 ‘7-11’)에도 역시 제거하지 않았다.

이 결과 총 3,287개의 유니그램 자질이 추출되었고, 이 자질들을 활용하여 기계학습 알고리즘을 적용하였다. 기계학습 알고리즘은 기존연구들에서 논조분류를 위해 가장 높은 성능을 발휘하는 것으로 알려진 ‘SVM (Support Vector

Machine)'과 'MNB (Multinomial Naive Bayesian)'을 적용하였다.[12)]

바이그램에 기반한 모델의 구현을 위한 전처리 과정은 유니그램 기반의 모델과 동일하지만, 'StringToWordVector' 함수의 적용시 'Word Tokenizer' 대신에 'N-gram Tokenizer'를 사용한다는 차이가 있다. 이 때 'n'의 값을 2로 정해주면 바이그램 모델이 준비되는 것이고, 3으로 정해주면 트라이그램 모델을 위한 전처리 과정에 들어가게 된다.

'N-gram Tokenizer'를 적용하여 바이그램 자질을 추출한 결과 총 1,988개의 자질이 추출되었다. 이 중 앞서 본 유니그램의 경우와 마찬가지로 아라비아 숫자 자질과 기호 자질은 제거해서 최종적으로 사용한 바이그램 자질의 수는 1,971개였다. 트라이그램 자질은 총 2,228개가 추출되었다. 이 중 아리비아 숫자 및 기호를 제거한 후 최종적으로 2,213개가 실험을 위해 사용되었다.

실험에서는 유니그램 모델, 바이그램 모델, 트라이그램 모델을 'SVM'과 'MNB' 알고리즘을 각각 적용하여 구현하였으며 그 성능을 비교평가 하였다. 각 모델을 구현할 때는 웨카가 제공하는 자질 선택 Merkmalsauswahl 기능을 사용하지 않고 추출된 자질을 모두 활용하였다.

Ⅳ.2 실험결과

본 연구에서 제안하는 바이그램 기반 논조분석 모델의 성능은 정확률의 측면에서 비교평가 되었다. 정확률은 논조분석 모델에 의해 정확하게 분류된 문장의 비율이다. 예를 들어 전체 1천개의 문장에 대해 700개 문장의 논조를 정확하게 분류했다면 이 모델의 정확도는 70%이다. 논조분석과 같은 데이터 마이닝 분야에서 어떤 모델의 성능을 평가할 때 가장 많이 사용하는 개념이 정확률이다.

먼저 <표 3>은 유니그램만을 사용한 모델, 유니그램과 바이그램을 사용한 모델, 유니그램, 바이그램, 트라이그램을 모두 사용한 세 개의 논조분석 모델의 정확률을 나타낸다. 각각의 모델은 'SVM'과 'MNB'의 알고리즘을 사용해 학습되었다.

---

12) Pang et al. (2002), Wilson et al. (2005) 등의 연구결과에 따르면 논조분석을 비롯한 텍스트 마이닝 분야에서 가장 높은 성능을 발휘하는 것으로 알려진 기계학습 알고리즘은 'SVM'과 'MNB'이다.

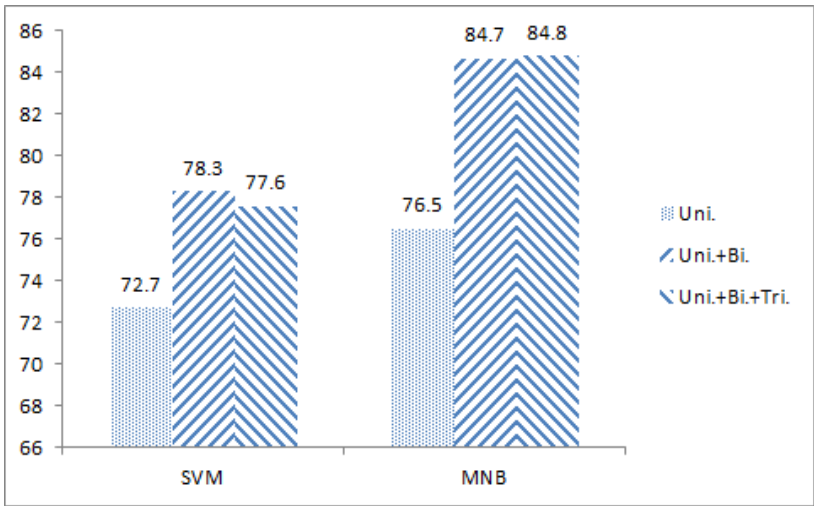
<표 3> : n-그램을 활용한 논조분석 정확률

	자질의 수	SVM	MNB
Unigram	3,245	72.7	76.5
Unigram+Bigram[13]	1,971	78.3	84.7
Unigram+Bigram+Trigram	2,213	77.6	84.8

베이스라인 방법인 유니그램기반의 논조분석 모델은 'SVM' 알고리즘에서는 72.7%의 정확률, 'MNB' 알고리즘에서는 76.5%의 정확률을 나타내었다. 본 논문에서 제안하는 바이그램 기반의 방법론은 'SVM' 알고리즘에서는 78.3%의 정확률을 'MNB' 알고리즘에서는 84.7%의 정확률을 나타내었다. 이는 베이스라인 대비 각각 5.6%, 8.2%의 성능향상을 보인 것이다. 트라이그램 기반의 방법론은 77.6%, 84.8%의 정확률을 나타내었는데 각각 베이스라인 모델보다는 4.9%, 8.3%의 성능향상을 나타내었으나, 바이그램 모델과 비교해서는 'SVM' 알고리즘의 경우에는 오히려 0.7%의 성능하락, 'MNB' 알고리즘의 경우는 0.1%의 성능향상을 보였다.

<그림 1>에서 시각적으로 좀 더 뚜렷이 볼 수 있듯이 본 논문이 제안하는 바이그램 기반의 논조분석 모델은 'SVM'과 'MNB' 알고리즘을 활용한 기계학습에서 모두 베이스라인보다 월등히 더 높은 성능을 나타내었다. 문장의 구조정보를 전혀 반영하지 못하는 유니그램 모델과 비교하여 언어모델링을 통해 문장의 구조정보를 간접적으로 반영하는 바이그램 모델은 'SVM'과 'MNB' 알고리즘의 경우 각각 5.6%, 8.2%의 정확률 향상을 보였다.

13) 유니그램과 바이그램을 모두 합친 개수가 유니그램만의 개수보다 작은 이유는 문장의 길이들을 고려하여 특정 임계치 이하의 저빈도 유니그램과 바이그램은 고려하지 않기 때문이다. 이는 트라이그램의 경우도 마찬가지이다.



<그림 1> 논조분석 정확률 평가결과

5.6%와 8.2%의 정확률 향상은 상당히 긍정적으로 해석될 수 있다. 이와 직접적인 비교대상은 아니지만 일반 문장에 대해 구조분석을 수행하고 구조분석 정보를 기계학습에 적용하여 문장의 논조분석을 시도한 홍문표 (2013b)의 연구에 따르면 구조분석 기반의 논조분석은 유니그램 기반 논조분석 모델과 비교하여 2.1%의 성능향상을 보이는 것으로 보고되었다. 따라서 구조분석을 수행하지 않고서도 바이그램을 통해 학습코퍼스의 언어모델링을 수행한 후 이를 기계학습에 적용한 방법이 성능향상에 기여하는 정도는 매우 높다고 볼 수 있겠다.

그러나 트라이그램 기반의 모델은 기계학습 알고리즘에 따라 상이한 결과를 나타내었다. ‘SVM’ 알고리즘의 경우에는 바이그램 모델대비 오히려 0.7%의 정확률이 하락하였고, ‘MNB’ 알고리즘의 경우에는 바이그램 모델대비 불과 0.1%의 정확률만 상승하여 거의 차이가 없다고 볼 수도 있다. 트라이그램이 바이그램과 거의 성능의 차이가 없거나 오히려 성능이 떨어지는 이유는 학습코퍼스의 규모에 기인한 것으로 보인다. 그러나 현실적으로 대부분의 학습코퍼스의 크기에 한계가 있으므로 트라이그램을 활용한 언어모델링은 바이그램 기반 모델에 비해 현실성이 떨어지는 것으로 보인다.

기계학습 알고리즘의 관점에서 보면 모든 경우에 ‘MNB’ 알고리즘이 ‘SVM’ 알고리즘보다 더 높은 정확률을 나타내었다. 또한 실험결과에는 반영되어 있지 않지만 실행속도면에서도 월등히 빠른 시간을 나타내어 대용량의 데이터 처리에 있어서는 특히 다른 기계학습 알고리즘보다도 효과적인 학습알고리즘이 될 것으로 예상된다.

IV.3 오류분석

이번 절에서는 최고 84.7%의 정확률을 보인 바이그램기반 논조분석기가 잘못 분류한 문장들을 유형별로 살펴보고 그 이유를 분석해보기로 한다. 다음의 예문들은 논조분석기에서 잘못 분류된 문장들이다. 먼저 (17)~(19)의 문장은 긍정논조의 문장임에도 부정논조로 잘못 분류된 경우이다. 그리고 (20)~(22)의 문장은 부정논조가 긍정으로 잘못 분류된 경우이다.

(17) Würde jederzeit wieder ins Adelphi.
(18) Wir waren im 3. Stock obwohl direkt an der Hauptstrasse gelegen war es sehr ruhig kein Problem zum schlafen.
(19) Wir haben uns für das Buffett entschieden und waren trotz der relativ kleinen Auswahl sehr zufrieden.
(20) Zusätzlich das "Business Center" war einfach ein Schreibtisch mit einem Computer in der Lobby.
(21) Zimmer nicht sauber Klimaanlage laut Dimm Beleuchtung schalteten Lichter vor allen Gästen sogar vor Sonnenaufgang.
(22) Wir hatten ein nicht Raucher Zimmer gebucht und ein stinkendes Zimmer im 3ten Stock (nicht Raucher Etage) erhalten.

(17)번 문장은 사용자가 다시 여행을 떠난다면 ‘Adelphi’ 호텔로 기꺼이 다시가겠다는 내용을 담은 긍정적인 논조의 문장이다. 이 문장은 그럼에도 불구하고 긍정논조의 단어가 전혀 포함되어 있지 않고, 문장 자체도 단 5개의 단어로 구성된 매우 짧은 문장이다. Klenner et al. (2009)에서 지적한 바와 같이 기계학습 기반의

논조분석 방법론은 입력문이 짧을 경우에는 특히 정확률이 떨어진다.

(18)번 문장은 바이그램 기반의 모델에서 충분히 정확하게 분류될 수 있는 문장이나 아마도 학습코퍼스의 데이터 부족 문제 때문에 발생한 것으로 추정된다. 입력문장에 존재하는 'kein Problem'은 대부분 긍정논조의 데이터에 등장하는 바이그램이나 이 문장은 부정으로 분류되었다.

(19)번 문장은 방법론 자체의 한계에 기인한 것으로 보인다. 먼저 'kleinen Auswahl'은 주로 부정논조의 데이터에 등장하는 바이그램이다. 그러나 이것은 'trotz'라고 하는 전형적인 소위 '부정극성 쉬프터 negative polarity shifter'에 의해 긍정논조로 해석되어야 한다. 바이그램 모델에서는 연속하는 두 단어 이상의 표현을 고려하지 못하기 때문에 이 표현의 부정극성값이 이어 나오는 'zufrieden'보다 높은 가중치를 갖게 되어 문장 전체가 부정으로 분류된 것으로 추정된다.

(20)번 문장은 문장 자체만 본다면 긍,부정을 논하기가 어려울 수도 있다. 그러나 이 문장이 등장한 텍스트상의 문맥이 호텔내 비즈니스센터의 취약함을 비판하는 내용이다. 따라서 이 문장도 자연스럽게 부정의 논조로 분류되는 경우이나, 본 연구에서 제안하는 방법론에서 이러한 문맥정보를 파악하는 것은 불가능하다.

(21)번 문장은 (18)번 문장의 오류와 비슷한 이유로 부정으로 분류되어야할 문장이 긍정으로 분류되었다. 바이그램 'nicht sauber'가 등장함에도 학습코퍼스의 부족으로 부정으로 분류되지 못한 것으로 보인다.

(22)번 문장도 (21)번 문장과 유사하게 'stinkendes Zimmer'라는 바이그램이 학습코퍼스에 등장하지 않는 이유로 발생한 오류로 분석된다. 위의 오류들을 종합해보면 (17), (19), (20)의 오류는 본 연구에서 제안하는 방법론의 한계에 기인하여 발생한 것들이고 그 외의 오류들은 대용량의 학습코퍼스에서 모델을 학습한다면 해결할 수 있는 오류들로 판단된다.

본 연구에서는 추가적으로 웨카 시스템의 자질선택 Merkmalsauswahl 기능을 활용해 호텔예약분야 코퍼스에서 긍정, 부정 논조를 분류하는데 가장 큰 역할을 하는 유니그램과 바이그램, 트라이그램을 알아보았다. 이를 통해 사용자들이 호텔예약시 어떠한 키워드에 관심이 많으며 무엇을 통해 긍정적인 감정과 부정적인 감정이 형성되는지 간접적으로 알아볼 수 있다.

이 실험을 위해서 웨카 시스템에서 제공하는 자질선택 알고리즘 중 'Info Gain'

과 ‘ ChiSquared’ 알고리즘을 선택하여, 각 알고리즘별로 긍정, 부정을 분류하는데 가장 큰 역할을 하는 유니그램, 바이그램, 트라이그램을 알아보았다. <표 4>는 실험의 결과이다. 여기서는 각 알고리즘별로 최상위 20개의 자질들이 제시되었다. 각 알고리즘별로 1위부터 3위까지의 자질은 공통적으로 ‘gut’, ‘sehr’, ‘pool’이었다. ‘gut’이 긍정, 부정을 분류하는 대표적인 어휘이므로 어느정도 예상된 결과라고 할 수 있으나, ‘sehr’가 논조를 분류하는데 매우 큰 역할을 한다는 것은 예상하지 못한 결과라고 할 수 있다. 일반적으로 부사 ‘sehr’는 긍정과 부정 논조에 고루 나타날 것으로 예상되나, 이 분석의 결과는 그렇지 않다는 것을 의미한다. 이에 대한 연구는 또 하나의 흥미로운 연구가 될 것이므로 본 연구에서 본격적으로 다루지는 않고 향후 연구로 미루고자 한다.

<표 4> 호텔예약분야텍스트 논조분류에 가장 효과적인 자질 리스트

InfoGain		ChiSquared	
uni+bi	uni+bi+tri	uni+bi	uni+bi+tri
gut	gut	gut	gut
sehr	sehr	sehr	sehr
pool	pool	pool	pool
sehr gut	alt	sie	sie
alt	freundlich	freundlich	freundlich
freundlich	sie	sauber	lage
sie	flecken	bangkok	alt
bangkok	lage	alt	sauber
flecken	sehr gut	sehr gut	bangkok
sauber	bangkok	lage	sehr gut
super	sauber	flecken	flecken
freundlich und	super	mr	mr
lage	freundlich und	freundlich und	freundlich und
mir	mir	sauber	sauber
war nicht	nur	kam	nur
sauber und	sauber und	eingerichtet	sauber und
über	eingerichtet	sauber und	eingerichtet
eingerichtet	über	über	über
nur	war nicht	nur	war nicht
kam	kam	war nicht	kam

V. 결론

본 연구에서는 문법오류, 철자오류 등과 같은 비문을 다수 포함한 인터넷 게시글의 자동 논조분석 방법을 다루었다. 문장 의미의 긍정성, 부정성을 분류하는 논조분석 분야는 단어가 가지고 있는 긍정성, 부정성 정보에 기반한 방법론이 분석 방법의 대다수를 이루었다. 그러나 문장의 논조가 단지 단어가 본래 가지고 있는 공시의에만 의존하는 것이 아니라 단어와 단어의 결합관계와 같은 통사구조에 의해서도 흔히 결정된다는 것도 최근의 연구에서 밝혀졌다. 따라서 문장의 구조분석 결과를 기계학습에 반영하는 방법론이 논조분석 분야에서 가장 최근의 연구주제였다. 그러나 비문이 빈번히 등장하는 인터넷 게시물의 경우에는 문장의 정확한 구조분석이 매우 어렵다는 점에서 기존의 방법론은 그 한계를 보인다.

이러한 문제를 해결하고자 본 연구에서는 바이그램 정보가 통계적 타당성을 가지고 문장구조를 반영하는 점에 착안해 바이그램을 이용한 논조분석 방법을 제안하였다. 본 연구는 기존의 연구결과와는 달리 바이그램이 유니그램과 함께 사용될 경우 구조분석을 실제 수행한 것 이상의 성능을 나타낼 수 있음을 실험을 통해 보였다. 특히 'MNB' 알고리즘을 사용하여 기계학습을 수행한 경우 유니그램만을 사용한 베이스라인 대비 최대 8.2%의 성능향상이 있음을 보았다. 추가적으로 본 연구의 발견은 기존 논조분석 연구에서 가장 성능이 높다고 알려진 'SVM' 알고리즘보다 'MNB' 알고리즘이 분류의 정확도가 6% 이상 높으며 처리속도 또한 월등히 짧다는 것이다.

그러나 본 연구의 한계점은 오류분석과정에서 드러났듯이 5단어 이하의 짧은 문장의 논조분석에 어려움이 있다는 점이다. 이에 대한 연구는 향후 연구로 남겨놓고자 한다.

참고 문헌

신효필 (2009). *언어학과 통계모델*, 서울대학교 출판문화원
홍문표 (2013a). 기계학습에 기반한 문장의미의 주관성/객관성 자동분류

방법 – 독일어 트위터 문장을 중심으로, *독일언어문학 61집*, 47-67

홍문표 (2013b). 독일어텍스트 논조자동분석, *독어학 28집*, 333-355

Choi, Y. & Cardie, C. (2008): Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 793-801

Esuli, A. & F. Sebastiani (2006): SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC-06, 5th Conference of Language Resources and Evaluation*, 417-422

Klenner, M., Fahrni, A. & S. Petrakis (2009): PolArt: A robust tool for sentiment analysis. *Proceedings of the 17th Nordic Conference of Computational Linguistics. Vol. 4,* 235-238

Moilanen, K. & Pulman, S. (2007): Sentiment composition. *Procedings. of RANLP-2007*, 378-382

Ohana, B & Tierney, B. (2009): Sentiment classification of reviews using SentiWordNet. *Proceedings of 9th. IT&T Conference,* 13-20

Pang, B., L. Lee & Vaithyanathan, S. (2002): Thumbs up? Sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Volumn 10,* 79-86

Pang, B. & L. Lee (2004): A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, 271-278.

Remus, R., U. Quasthoff & G. Heyer (2010): SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*, 1168-1171

Strapparava, C. & A. Valitutti (2004): WordNet-Affect: an affective extension

of WordNet, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1083-1086

Waltinger, U. (2010): German Polarity Clues : A Lexical Resource for German Sentiment Analysis. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 1638-1642

Wiebe, J., Wilson, T. & C. Cardie (2005): Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation, 39(2/3)*, 164-210

Wilson, T., Wiebe, J. & P. Hoffmann (2005): Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, *Proceedings of HLT/EMNLP*, 347-354

Zusammenfassung

Bigramm-basierte Sentimentanalyse

Hong, Munpyo(Sungkyunkwan Univ.)

In der vorliegenden Arbeit wird ein alternativer Ansatz zur Sentimentanalyse vorgeschlagen. Unter dem Begriff ‘Sentiment’ wird in dieser Arbeit die als positiv oder negativ geäußerte Haltung des Sprechers verstanden. Beispielsweise zeigt ein Satz wie “*Das Hotel ist sehr preiswert und Ich kann es jedem empfehlen*” ein positives Sentiment. Dagegen kann z.B. in einem Satz wie “*Die Angestellten sind sehr unfreundlich und ich würde nie wieder kommen*” eine negative Stimmung konstatiert werden.

Eines der größten Probleme der bisherigen Ansätze mit Unigramm als ein Hauptmerkmal für die Sentimentanalyse ist dass kein Kontext berücksichtigt werden kann. Das Sentiment eines Satzes ist nicht nur von der Polarität

der Wörter im Satz, sondern auch von der syntaktischen Struktur des Satzes abhängig.

Moilanen/Pulman (2007) schlugen einen Ansatz vor, der syntaktische Repräsentationen für die Sentimentanalyse direkt heranzieht. Ihr Ansatz kann für kurze Sätze erfolgreich angewendet werden, da normalerweise ein Parser für kurze Sätze eine hohe Präzision bei der Analyse aufweist. Wenn ein Satz allerdings viele grammatische Fehler oder Tippfehler enthält, was häufig der Fall bei SNS Texten ist, kann der Ansatz schnell scheitern.

Um diese Problematik zu lösen, wird in dieser Arbeit ein alternativer Ansatz vorgeschlagen, der das Bigramm als ein Hauptmerkmal für maschinelles Lernen verwendet. In diesem Ansatz kann der lokale Kontext eines Wortes mitberücksichtigt werden, was dazu führt, dass lokale Negationen erfolgreich bei der Sentimentanalyse behandelt werden.

Das Experiment zeigte, dass unser Ansatz im Vergleich zum herkömmlichen Unigramm-basierten Ansatz die Genauigkeit der Sentimentanalyse von 76.5% auf 84.7% um 8.2% erhöhen kann.

핵심어 : 논조분석 Sentimentanalyse, 의견마이닝 Opinion Mining,
바이그램 Bigramm, 기계학습 maschinelles Lernen,
자질선택 Merkmalsauswahl

필자 E-mail : skkhmp@skku.edu

논문투고일 : 2014. 7. 15 / 심사일 : 2014. 8. 11 / 게재확정일 : 2014. 9. 5