독일언어문학 제61집(2013.9). 47-67

기계학습에 기반한 문장의미의 주관성/객관성 자동분류 방법
– 독일어 트위터 문장을 중심으로*

홍 문 표(성균관대)

I. 서론

최근 자연언어처리 분야 NLP에서 의미처리 Semantische Verarbeitung 연구는 문장의 진리조건적 Wahrheitsbedingung 의미를 분석하는 연구로부터 문장의 작성자 또는 발화자가 주제에 대해 갖는 호감도 또는 심리상태 등을 분석하는 연구 등으로 확장되고 있다. 문장의 진리조건적 의미를 분석하는 연구가 문장으로부터 사실관계 등을 추출하는데 응용되었다면, 문장의 주제에 대한 저자의 호감도 등에 대한 연구는 특정 제품이나 인물 또는 사건 등에 대한 사람들의 긍정 positiv, 부정적 negativ 태도를 자동으로 분석하는데 응용되고 있다. 논조분석 Sentimentanalyse이라 불리는 이 기술은 최근 정보처리 분야에서 가장 큰 화두가 되고 있는 빅 데이터 Big data의 처리에서 가장 중요한 모듈로 자리 잡고 있다.

주제에 대한 저자의 긍정적, 부정적 태도를 자동으로 분석하기 위해서 선행되어야 하는 것은 텍스트를 구성하는 문장들을 객관적 의미 objektive Bedeutung 만을 전달하는 문장과 긍정적, 부정적 태도 등과 같은 저자의 견해가 포함된 주관적 의미 subjektive Bedeutung를 전달하는 문장으로 분류하는 것이다. 이 분야는 자연언어처리 분야의 영어권 연구에서는 주관성 분류 Subjektivität Klassifizierung라는 주제로 연구되고 있으나, 독일어권에서는 아직 이 분야에 대한 연구가 활발히 진행되고 있지는 못하다.

본 연구에서는 주관적 의미의 문장과 객관적 의미의 문장을 자동으로 분류하기 위해 기계학습 maschinelles Lernen 기법을 사용한다. 기계학습은 인공지능 분야에서 패턴인식 Patternerkennung을 위해 최초로 소개된 기법이며, 최근 정보처리 분

* 이 논문은 성균관대학교의 2010학년도 삼성학술연구비에 의하여 연구되었음.

야에서 텍스트 마이닝, 바이오 인포매틱스 등과 같이 다양한 목적을 위해 성공적으로 사용되고 있다. 본 연구는 문장 의미의 주관성과 객관성의 분류를 위해 기계학습 기법이 얼마나 성공적으로 활용될 수 있으며, 그 한계는 무엇인지를 밝히는 것을 목적으로 한다.

이하 본 논문의 구성은 다음과 같다. 2장에서는 문장의미의 주관성에 대한 논의를 진행한다. 여기서는 특히 영어권 연구에서 주관성과 객관성 분류의 대표적 연구인 Wiebe et al. (2005)의 연구를 기반으로 하여, 독일어 문장의미의 주관성과 객관성 분류 기준을 논의한다. 또한 독일어 어휘에 긍정성, 부정성 등과 같은 정보가 부착된 'Senti-Wortschatz'에 대해 알아볼 것이다. 실험부분에서도 언급이 되겠지만, 'Senti-Wortschatz'는 문장의미의 주관성과 객관성을 분류하는데 결정적인 역할을 하게 되는 자질 Merkmal로서 활용된다.

3장에서는 본 연구에서 제안하는 기계학습 기반 분류방법에 대해 소개할 것이다. 기계학습 방법론은 크게 학습을 위한 데이터 Lerndaten가 필요한 지도학습방법론 überwachtes Lernen과 비지도학습방법론 unüberwachtes Lernen으로 나눌 수 있다. 본 연구에서는 지도학습방법론의 대표적 알고리즘인 'Support Vector Machine (이하 SVM으로 표기)'을 택하며, 기계학습에서 가장 중요한 자질선택 Merkmalsauswahl의 문제를 다룰 것이다.

4장에서는 3장에서 제안한 방법론에 대한 실험결과를 논한다. 3장에서 제안한 방법론은 '웨카 WEKA'라고 불리는 기계학습 실험용 오픈소스 프로그램을 활용하여 테스트되었다. 여기서는 본 연구에서 제안한 방법론을 실제 구현하는 방법과 그 실험결과를 소개할 것이다. 마지막으로 5장에서는 본 연구의 주장을 정리하며, 본 방법론의 가능성과 한계를 논하게 될 것이다.

II. 관련연구

II.1. 문장 의미의 주관성

문장이나 단어의 의미를 주관적 의미와 객관적 의미로 분류한 연구는 언어학

분야보다는 실용적인 목적 때문에 자연언어처리 분야에서 더 활발하게 진행되었다. 이 중 가장 대표적인 연구인 Wiebe et al. (2005)는 심리상태 표현의 세 가지 유형을 다음과 같이 구분하였다.

- 심리상태에 대한 직접적인 언급 explicit mentions of private states
- 심리상태를 표현하는 화행 speech events expressing private states
- 심리상태의 주관적 표현 expressive subjective elements

심리상태에 대한 직접적인 언급을 통해 감정이 표현되어 문장의미의 주관성이 나타난 예문은 다음과 같다.

(1) Auserdem hasse ich Apple mindestens so wie du

위 예문에서는 'hassen'이라는 동사를 통해 글쓴이의 감정이 표현된다. 이어서 (2)는 화행표현을 통해 감정을 직접 전달하는 예문이다.

(2) "Das ist doch nicht meine Tasche", weinte Anna

위 예문에서 동사 'weinen'의 선행절 'Das ist doch nicht meine Tasche'가 주어의 감정을 나타내는 문장이라고 볼 수는 없다. 그러나 저자는 'sagen'이나 'behaupten' 등과 같은 중립적인 의미의 동사를 사용하는 대신에 'weinen'의 사용을 통해 주어의 심리상태를 나타내고 있다.

마지막으로 일반적으로 코퍼스 분석시 가장 많이 등장하는 유형은 심리상태를 주관적으로 표현하는 어휘들이다.

(3) Das schlimmste Foltergerät ist mein Handy, wenn es nicht klingelt.

(3)에서 저자는 'schlimm'이라는 형용사를 통하여 주제에 대한 자신의 감정을 이미 나타내고 있지만, 가치중립적인 어휘인 'Gerät' 대신에 부정논조의 어휘

‘Foltergerät’를 사용하여 부정적인 논조를 추가하였다.

Wiebe et al. (2005)은 ‘MPQA (Multi-Perspective Question Answering)’라고 불리는 영어 논조분석 코퍼스를 구축하기 위해 위와 같은 기준을 적용하였으며, 본 연구에서도 독일어 문장에 대해 위와 동일한 기준을 적용하여 실험을 위한 코퍼스를 구축하였다.

II.2. Senti-Wortschatz

‘Senti-Wortschatz’ (이하 ‘ SW’로 표기함)는 라이프치히 대학에서 개발한 논조분석을 위한 독일어 사전이다.[1] 이 사전에는 1,650개의 긍정단어와 1,818개의 부정의미 단어가 수록되어 있으며, 단어의 활용형태로 보면 15,649개의 긍정 어휘목록과 15,632개의 부정 어휘목록으로 구성되어 있다. 품사별로는 감정을 나타내는 형용사, 부사 뿐만 아니라 명사와 동사도 수록되어 있다. 이 사전에 수록되어 있는 모든 어휘는 긍정의미와 부정의미 중 어느 쪽에 더 가까운가에 따라 -1에서 1의 값을 갖는다. -1에 가까울수록 부정적 의미의 어휘이고, 1에 가까울수록 긍정적 의미의 어휘이다.

‘SW’는 영어권의 ‘SentiWordnet’[2]과 유사한 포맷으로 구성되어 있으나, 사전구축과정은 다르다. 이 사전은 크게 세 개의 소스를 통해 구성되었으며, 첫 번째 소스는 영어권에서 많이 사용되는 ‘General Inquirer’사전이다. 이 사전은 수록 어휘에 대해 긍정과 부정 정보가 부착되어 있는데, 이 사전의 엔트리를 구글 Google 독-영 번역기를 사용하여 번역한 후, 수작업으로 수정하는 과정을 거쳤다. 두 번째 소스는 어휘 공기정보 분석 Kookkurenzanalyse을 통한 것이다. 상품평에 대해 수작업으로 ‘긍정’과 ‘부정’의 태그를 붙인 후, 각 논조별로 특히 자주 등장하는 어휘를 ‘log-likelihood" 측정을 통해 골라내었다. 이렇게 자동으로 골라진 어휘들은 수작업자들에 의해 최종적으로 선택 또는 삭제되었다.

마지막 소스는 ‘German Collocation Dictionary’를 활용한 것이다. 이 사전은 의미클래스에 따라 구성되어 있는데, 앞선 첫 번째와 두 번째 소스를 통해 얻어진

1) http://asv.informatik.uni-leipzig.de/download/sentiws.html
2) Esuli & Sebastiani (2006).

사전 엔트리가 속하는 의미클래스에 있는 다른 단어들도 ‘긍정’ 또는 ‘부정’의 속성을 공유할 것이라는 가정 하에, 이 단어들도 ‘SW’의 후보 엔트리로 선정한 후, 수작업을 통해 걸러내어 사전을 구축하였다.

위와 같이 컴파일된 사전엔트리들은 어휘 의미의 ‘긍정’, ‘부정’ 속성에 따라 -1부터 1의 값을 갖게 된다. -1에서 1의 값을 정하는 방법은 Turney (2002)가 비지도 학습기반 논조분석에서 사용했던 방법인 ‘Pointwise Mutual Information (이하 PMI)’이다. ‘PMI’는 어떠한 두 단어 간의 긴밀도를 조사할 때 사용되는 방법으로서 a라는 단어가 b라는 단어와 함께 출현할 가능성이 높은 경우, 그렇지 않은 경우보다 ‘PMI’ 값이 높다. 이 연구에서는 사전 엔트리들이 ‘gut’, ‘exzellent’, ‘schlecht’, ‘falsch’ 같은 대표적인 긍정어휘, 부정어휘와의 ‘PMI’ 값을 계산해 −1에서 1의 값을 부착하였다. 예를 들어 어떤 단어와 ‘gut’ 간의 ‘PMI’값과 ‘schlecht’ 간의 ‘PMI’ 값을 비교하여, 그 절대치가 ‘gut’과의 값이 크다면, 이 단어는 긍정적인 의미의 단어일 가능성이 높다는 것이다.

이러한 방식으로 구축된 ‘SW’는 문장의미의 주관성/객관성 분류에서 큰 역할을 할 수 있다. 이러한 단어가 사용된 어휘는 주관적인 의미를 나타낼 가능성이 클 것이기 때문이다.

III. 기계학습 기반 자동분류 방법

III.1. 기계학습

기계학습 maschinelles Lernen이란 학습데이터에 존재하는 패턴을 수학적으로 계산하여 새로운 데이터의 클래스를 예측하는 방법론의 집합이라고 볼 수 있다.[3] 기계학습 기법은 문자 인식 Charaktererkennung, 문서 분류 Dokumentklassifizierung 등과 같은 패턴인식 분야를 위해 개발되었으나 최근에는 그 활용범위가 전산학뿐만 아니라, 언어학, 생물학, 의학, 경제학, 경영학, 사회학 등 거의 전 학문분야에

3) Murphy (2012): p.1.

걸쳐 넓어지고 있다.

기계학습 방법은 크게 지도기반학습방법 Überwachtes Lernen과 비지도기반학습방법 Unüberwachtes Lernen으로 나눌 수 있다. 두 방법의 차이점은 학습을 위한 데이터의 유무이다. 일반적으로 지도기반학습방법은 학습데이터 Lerndaten를 필요로 하고, 비지도기반학습방법은 학습데이터를 필요로 하지 않는다. 패턴인식을 위한 학습데이터는 일반적으로 레이블 label이 부착되어 있는 형태이다.

지도기반학습방법은 학습데이터를 필요로 한다는 점에서 비지도학습방법보다는 시간과 노력이 많이 들지만, 대체적으로 비지도학습방법보다 높은 성능을 보이는 것으로 보고되고 있다.[4)]

지도기반학습방법론에는 'Support Vector Machine (SVM)' 알고리즘, 결정트리 Decision Tree 알고리즘, 'Naive Bayesian (NB)' 등외에도 매우 많은 알고리즘들이 개발되어 있지만, 문장 의미 분석이나 논조 분석 분야에는 'SVM'과 'NB' 알고리즘이 가장 많이 사용되고 있다. 이 중에서도 'SVM'은 다수의 연구에서 가장 높은 성능을 나타내는 것으로 보고되고 있다.[5)]

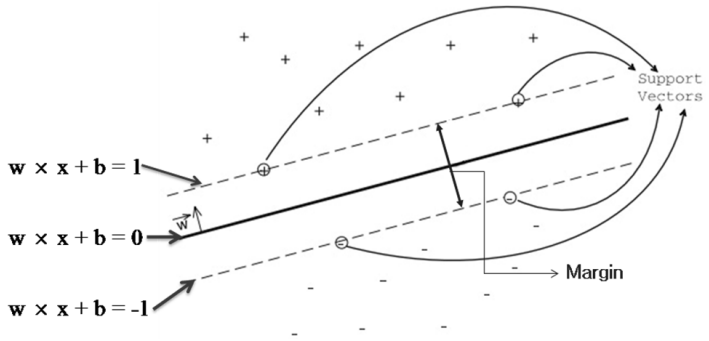
Cortes & Vapnik (1995)에 의해 제안된 'SVM'알고리즘은 두 개의 클래스로 이루어진 데이터를 가장 분명하게 나눌 수 있는 함수를 찾아내는 알고리즘이다. 여기서 두 개의 클래스를 가장 분명하게 나눌 수 있다는 의미는 각 클래스에 속하는 데이터가 함수가 나타내는 선과 최대한의 차이값을 갖도록 함수를 정의한다는 의미이다. 아래의 <그림 1>에서 $w \blacksquare x + b = 0$ 함수는 두 개의 클래스와 최대한의 차이값을 가지면서 두 클래스를 분류하고 있는데, 'SVM' 알고리즘은 결국 여기서의 w와 b 값을 찾아내는 역할을 한다.[6)]

---

4) Alm et al. (2005), Hatzivassiloglou & Mackeown (1997), Pang et al. (2002), Pang & Lee (2004), Popescu & Etzioni (2005).

5) 이와 관련된 대표적인 연구로는 Pang et al. (2002)을 들 수 있다.

6) <그림 1>은 홍문표외(2010)에서 발췌함.



<그림 1> SVM의 선형모델

III.2. 기계학습을 위한 자질선택

기계학습 방법의 적용을 위해 가장 중요한 것은 두 개의 클래스를 구별할 수 있는 적절한 클래스별 자질 Merkmal의 선택이다. 자질이 적절하게 선택되어야 이에 따라 학습데이터를 구성하고 'SVM' 등과 같은 기계학습방법론을 적용하여 두 클래스를 분류하는 최적의 함수를 찾아낼 수 있다.

문장의미의 주관성과 객관성을 판별하기 위해 적용할 수 있는 자질은 주관적 의미의 문장에 주로 등장하는 어휘이다. 예를 들어 'Super', 'Klasse' 등과 같은 어휘는 주로 긍정적인 논조의 문장에서 사용되는 어휘들이다. 이러한 어휘들은 해당 문장이 주관적 의미를 가지고 있음을 나타내는 결정적인 자질이 될 수 있다. 물론 어떤 문장에 'Super', 'Klasse'가 나타나지 않는다고 해서 그 문장이 반드시 객관적인 문장이라는 보장은 없다. 그러나 이렇게 두 클래스를 구별할 수 있는 자질들을 충분히 확보한다면 기계학습 알고리즘을 활용하여 두 개의 클래스를 자동으로 분류할 수 있을 것이다.

자질선택을 위해 가장 먼저 생각해 볼 수 있는 방법은 학습코퍼스에 등장하는 유니그램 unigram을 활용하는 방법이다. 주관적 의미와 객관적 의미의 학습코퍼스에 출현하는 모든 어휘, 즉, 유니그램을 두 클래스의 분류를 위한 자질로 활용할 수 있다. 흔히 문서분류 분야에서 'bag of words' 방식으로 불리는 이 방법은 문서

의 유니그램을 자질로 사용하여, 출현 빈도 또는 출현 여부를 값으로 하는 벡터 Vector를 만들어 학습하는 방법이다. 논조분석을 위한 이 방법은 이미 Alm et al. (2005), Hatzivassiloglou & Mackeown (1997), Pang et al. (2002), Pang & Lee (2004), Popescu & Etzioni (2005) 등의 연구에서 성공적으로 그 성능이 입증된 바 있다.

그러나 유니그램을 기계학습 자질로 선택하기 위해서는 고려해야할 것이 있다. 학습데이터에 등장하는 모든 어휘들을 자질로 선택한다면 기계학습의 계산 복잡도로 인해 효율성도 떨어질 뿐만 아니라, 학습데이터에 등장하는 노이즈 noise 때문에 기대하지 않은 결과가 나올 수도 있다. 따라서 많은 유니그램 중 그 일부만을 자질로 사용하는 것이 더 효과적일 수 있다. 또한 비교적 적은 규모의 학습코퍼스에서 자질을 추출할 경우, 자질선택이 학습코퍼스에 지나치게 의존적이게 되어, 다양한 종류의 테스트 데이터에 대해서는 과적용되어 성능이 떨어지게 되는 'overfitting'의 문제가 발생할 수도 있다.[7)]

따라서 좀 더 일반적인 자질을 선택할 필요가 있는데, 이를 위해 고려해볼 수 있는 것이 'SW'의 엔트리이다. 'SW'의 엔트리들은 앞서 살펴본 바와 같이 긍정과 부정의 의미를 지니는 어휘들이다. 이 사전의 엔트리들은 긍정, 부정의 의미에 따라 정교하게 선정된 어휘들이라 일반적으로 문서 전체에 대한 'bag of words' 방식보다는 노이즈가 적다. 따라서 이 어휘들을 기계학습을 위한 자질로 고려해볼 수 있을 것이다. 유사한 연구로서 Ohana & Tierny (2009)에서도 문서레벨의 영어 논조분석을 위해 'Senti-Wordnet'을 성공적으로 활용할 수 있음을 보였다.

그러나 'SW'는 일반적인 사전을 중심으로 컴파일 된 사전이므로, 트위터 등에 흔히 등장하는 욕설이나 비속어, 외래어 등을 많이 수록하고 있지는 않다. 박신혜 (2011)의 연구에서 밝힌 바와 같이 트위터는 크게 단문메시지적인 특징과 구어체적인 특징을 갖고 있다. 위 연구에 따르면 단문메시지적인 특징은 이모티콘의 사용, 축약형태의 빈번한 사용, 자소 반복을 통한 강조 등이다. 구어체적인 특징으로는 영어표현의 사용과 음운론적 변이형태의 사용, 감탄사나 의성어 등의 사용을 들고 있다. 특히 이모티콘은 이 연구에 따르면 전체 트위터 메시지의 26.4%에서 사용되고 있어, 거의 4문장에 하나 이상의 이모티콘이 사용됨을 알 수 있다. 이모티콘이 사용된 전체 문장중 38.8%는 긍정의 의미로 이모티콘이 사용되었고,

7) Bird et al. (2009): p.225.

24.7%는 부정의 의미로 사용되었으며, 36.4%는 객관적 의미로 사용되었다. 이모티콘이 사용된 전체 문장 중 62.5% 정도는 주관적 의미로 사용되었음을 알 수 있다. 즉, 이 관찰결과를 통해 이모티콘도 문장의미의 주관성을 나타내는 중요한 자질로 사용된다는 것을 추측할 수 있다.

본 연구에서는 위와 같은 점을 고려하여 'SW'의 주요 어휘들과 트위터의 언어학적 특징을 반영한 자질세트를 선정하고, 실험을 통하여 최적의 자질세트를 찾아내고자 한다.

III.3. 방법론의 구현

본 연구에서 제안하는 방법론은 '웨카'를 활용하여 구현되었다. '웨카'는 뉴질랜드의 'Waikato' 대학에서 개발하여 보급한 기계학습을 위한 시스템이다.[8] 이 시스템은 자바로 구현되어 있어, 유닉스 기반의 시스템 뿐만 아니라 일반 PC나 맥 OS에서도 구동이 가능하다. 이 시스템은 'SVM' 뿐만 아니라 'NB', 'Decision Tree' 알고리즘 등과 같은 대부분의 기계학습 알고리즘을 지원하며, 자질선택 기능과 성능평가 기능도 제공한다.

학습데이터를 '웨카' 시스템에 입력하기 위해서는 '웨카' 시스템이 이해할 수 있는 포맷으로 변경되어야 한다. '웨카' 시스템은 'ARFF'라고 불리는 자체적으로 정의한 포맷이나, 컴마로 분리된 'csv' 파일 형태를 입력으로 받아들인다. 'ARFF'는 '관계 relation'의 정의, '자질 Merkmal'의 정의 및 '자질값의 유형 Typ'을 먼저 선언 deklarieren하고, 이에 따른 데이터로 구성된다. 예를 들어, 데이터를 '주관성'과 '객관성'의 두 가지 클래스로 분류하기 위해, 'gut', 'schlecht'라는 두 가지의 자질만을 사용하여 학습한다면, 다음과 같이 'ARFF' 파일을 구성할 수 있을 것이다.

```
@relation subjektivitaet
@attribute gut numeric
@attibute schlecht numeric
@sentiment {o,s}
```

8) http://www.cs.waikato.ac.nz/ml/weka/

@data
0, 0, o
1, 0, s
0, 1, s
0, 0, s
...

위 ‘ARFF’ 파일은 ‘subjektivitaet’라는 이름의 관계를 학습하기 위해, ‘gut’와 ‘schlecht’라는 자질을 사용하고, 두 자질은 각각 수치 numeric 값을 가지며, 최종적으로 분류하고자 하는 클래스는 의미의 객관성과 주관성 (여기서는 ‘o’와 ‘s’의 값으로 표기함)이라는 것을 보여준다. 그리고 이러한 자질과 값을 기반으로 학습데이터는 ‘0, 0’, ‘0, 1’ 등의 벡터로 구성되는데, 이는 예를 들어 첫 번째 데이터의 경우에는 자질 ‘gut’와 ‘schlecht’가 모두 0번 출현하였고, ‘o’, 즉, 객관적인 의미의 문장이라는 것을 의미한다.

본 연구에서는 III.2에서의 관찰을 기반으로 자질을 선택하고 학습데이터를 ARFF 파일 형식으로 구성하여 기계학습 기반의 독일어 문장의미 주관성/객관성 분류기를 구현하였다.

IV. 성능평가

IV.1 실험

이번 장에서는 기계학습기반의 문장의미의 주관성 분류방법 성능을 평가한다. 본 논문의 실험에서는 ‘SVM’의 적용을 위해 웨카 버젼 3.6.9시스템을 활용하였다. 본 실험의 학습데이터 Lerndaten와 테스트데이터 Testdaten로 사용한 문장은 휴대폰 분야에 대한 독일어 트위터 문장이다. 전체 코퍼스의 크기는 630개의 트윗 Tweet이며, 총 861문장, 9,856단어의 규모이다. 630개의 트윗 중 절반인 315개의 트윗은 주관적 의미의 트윗이며, 나머지 절반인 315개의 트윗은 객관적 의미의 트

윗이었다. 트윗당 평균 문장수는 약 1.37개이며, 문장당 단어수는 11.45였다. 630개의 트윗은 저자를 포함한 2명의 검수자가 주관성과 객관성 분류에 모두 동의한 정확도 높은 데이터로 볼 수 있다.

성능평가는 총 630 트윗 규모의 코퍼스를 10개의 세트로 나누어 그 중 9개 세트인 567트윗에 대해 기계학습을 수행하고, 나머지 63개의 트윗에 대해 성능평가를 수행하고, 이 과정을 모든 세트에 대해 반복적으로 수행한 후, 각 평가결과를 모두 합산하여 평균을 구하는 '10-fold' 평가방식으로 수행하였다. 이와 같은 평가방식을 데이터 마이닝 분야에서는 '교차검수 cross validation'라고 부르는데, 이는 학습데이터와 테스트데이터의 분량이 크지 않을 때 효과적으로 사용될 수 있다.[9)]

본 연구에서 제안하는 'SW'와 독일어 트위터 문장의 언어학적 특성을 반영한 자질선택 방법론의 성능을 비교하기 위해 학습데이터의 출현어휘 또는 유니그램 unigram을 모두 자질로 활용한 'bag of words' 방법론의 성능도 평가하였다. 'bag of words' 방법론의 평가를 위해 학습데이터 문장을 웨카 시스템의 'StringtoWordVector' 필터를 활용해 유니그램 자질로 변환하였다. 필터링을 통해 총 2,895개의 자질이 추출되었으며, 이 자질들에 대해 'SVM' 알고리즘을 적용하였다.

IV.2 실험결과 및 분석

각 방법론의 성능평가는 다음의 네 가지 측면에서 이루어졌다.

- 정확도 accuracy
- 정밀도 precision
- 재현율 recall
- f-측정값 f-measure

이 중 정확도는 전체 트윗 중 기계학습에 의해 정확하게 분류된 트윗수의 비율이다. 예를 들어 총 630개의 트윗 중 315개의 트윗에 대해 주관성과 객관성 분류가 정확하게 이루어졌다면, 50%의 정확도이다.

9) Bird et al. (2009): p.241.

이와 유사한 개념으로 정밀도는 어떤 클래스로 분류한 결과 중 올바르게 분류한 비율이다. 예를 들어 어떤 방법론이 300개의 트윗을 '주관적'이라고 분류하였고, 이 중 실제로는 100개의 트윗만이 '주관적' 의미라면, 이 방법론의 정밀도는 100/300=0.333이다.

재현율은 어떤 클래스의 전체 개수 중 올바르게 찾아낸 개수의 비율이다. 예를 들어 '주관적'인 트윗이 300개 있을 때, 어떤 방법론이 150개의 올바른 주관적 의미의 트윗을 찾아냈다면, 이 방법론의 재현율은 0.5 또는 50%이다.

f-측정값은 정밀도와 재현율의 평균적인 값으로서, 일반적으로 '(2 · 정밀도 · 재현율) / (정밀도 + 재현율)'의 공식으로 계산된다.

본 연구에서 제안한 방법론은 630개의 트윗을 무작위로 '주관성'과 '객관성'으로 분류할 경우 확률적으로 기대되는 정확도인 50%를 비교의 베이스라인 baseline으로 삼고, 'bag of words' 방법론과 비교되었다. 또한 제안한 방법론에서 자질의 수를 변화함에 따른 성능의 변화도 관찰하였다.

먼저 'bag of words' 방식으로 기계학습한 방법론은 630개의 트윗에 대해 397개의 주관성과 객관성 레이블을 정확하게 예측하여, 약 63.02%의 정확도를 보였다. 총 315개의 트윗으로 구성된 주관적 의미의 데이터에 대해 살펴보면 정밀도는 0.639였으며, 재현율은 0.6의 성능을 보였다. 역시 총 315개의 트윗으로 구성된 객관적 의미의 데이터에 대해서는 정밀도는 0.623이었으며, 재현율은 0.66이었다. 각 클래스에 대한 f-측정값 수치는 주관적 의미는 0.619, 객관적 의미는 0.641이었다. (<표 1> 참조)

<표 1> 'bag of words' 기반 기계학습 성능평가결과

총 자질의 수	2,395
정확도	63.0159%
주관적 의미 정밀도	0.639
주관적 의미 재현율	0.6
주관적 의미 f-측정값	0.619
객관적 의미 정밀도	0.623
객관적 의미 재현율	0.66
객관적 의미 f-측정값	0.641

‘bag of words’ 방식은 주관적 의미와 객관적 의미에 대해 고른 정밀도와 재현율을 나타내었다. 정밀도 측면에서는 주관적 의미의 분류에서 좀 더 높은 수치를 보였으나, 재현율의 측면에서는 객관적 의미의 분류에서 높은 수치를 나타냈다. 그러나 그 차이는 그리 크지 않다고 볼 수 있다.

두 번째로는 ‘SW’와 트위터의 언어적 특성을 자질로 활용한 기계학습 방법론의 분류 성능을 테스트하였다. 학습코퍼스에 등장하는 많은 어휘들을 모두 의미분류를 위한 자질로 활용할 경우 노이즈 noise가 발생하여, 분류 성능을 떨어뜨릴 수 있으므로 주관적 의미를 나타내는 문장에 주로 사용되는 어휘들이 수록되어 있는 ‘SW’를 자질선택에 활용하였다. 그러나 ‘SW’도 원형을 기준으로 3,500개가 넘는 많은 어휘들을 포함하고 있으므로, 학습코퍼스에 등장하는 어휘 중 ‘SW’에 수록되어 있고 2회 이상 사용된 어휘만을 자질로 선택했다.

그러나 앞서 기술한 바와 같이 ‘SW’에는 트위터에서 사용자들이 자신의 의견이나 감정을 표현하기 위해 흔히 사용하는 비속어나 이모티콘 등이 수록되어 있지는 않으므로, 본 연구에서는 이러한 비속어 등과 이모티콘도 하나의 중요한 자질로 보고 기계학습시 반영하였다.

이와 같은 과정을 거쳐 최종적으로 결정된 자질의 수는 총 130개였다. 총 130개의 자질을 사용한 기계학습 결과 마찬가지로 10-fold로 진행한 실험에서 630개의 트윗중 425개 트윗의 클래스를 정확하게 분류하여 67.46%의 정확도가 측정되었다. ‘bag of words’ 방식으로 2,395개의 자질을 사용한 방법론보다 약 4.4% 이상의 성능향상을 보였다. 의미별로는 주관적 의미에 대해서는 0.877의 매우 높은 정밀도를 보였지만 재현율 측면에서는 0.406으로 ‘bag of words’ 방식보다 낮아졌다. 정확률과 재현율을 평균한 수치인 f-측정값에서도 0.555의 수치를 보여, ‘bag of words’ 방식보다 성능이 낮아졌다.

객관적 의미에 대해서는 0.614의 정밀도를 보였으며, 재현율은 0.943이었으며, f-측정값은 0.743이었다. ‘bag of words’ 방식과 비교해서는 정확률 측면에서는 약 0.01 정도의 근소한 차이로 낮으며, 재현율은 0.28 정도로 월등히 높다. f-측정값 측면에서도 약 0.1 이상으로 높은 향상을 보였다. (<표 2> 참조)

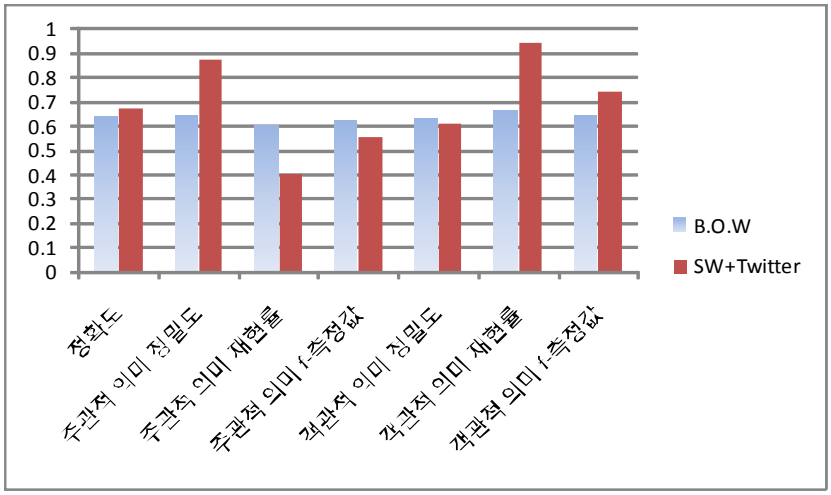
<표 2> ‘SW’와 트위터의 언어적 특성을 반영한 기계학습 성능평가결과

총 자질의 수	130
정확도	67.46%
주관적 의미 정밀도	0.877
주관적 의미 재현율	0.406
주관적 의미 f-측정값	0.555
객관적 의미 정밀도	0.614
객관적 의미 재현율	0.943
객관적 의미 f-측정값	0.743

다음 <그림 2>에서 보는 것처럼 본 연구에서 제안한 방법론은 ‘bag of words’ 방식보다 약 4% 정도의 정확도 향상을 보였다. 또한 주관적 의미에 대해서는 정밀도에서도 0.24 정도 월등히 성능이 향상됨을 보였다. 그러나 재현율의 측면에서는 약 0.22 정도 성능이 떨어졌다. f-측정값의 측면에서는 0.05 정도 성능이 하락했다.

객관적 의미에 대해서는 0.01 정도 정밀도가 하락하였으나, 재현율에서는 0.28 정도로 월등히 성능이 향상하였다. f-측정값에 대해서는 0.1 정도의 성능향상이 나타났다.

위와 같은 결과는 자질의 특성에 기인한 것으로 보인다. 본 방법론에서는 의미의 주관성 자질만으로 분류를 시도하므로, 어떤 문장에 해당 자질이 없으면 객관적 문장으로 분류해버릴 가능성이 높아진다. 따라서 주관적 의미로 분류한 문장의 분류 정밀도는 높지만 재현율은 낮고, 객관적 문장은 정밀도는 떨어지지만 주관적 자질이 발견되지 않으면 자동적으로 객관적 의미로 분류될 가능성이 높으므로 재현율이 높게 된다.



<그림 2> 'bag of words' 방식과 'SW+Twitter' 방식과의 성능비교

이어서 본 연구에서 제안한 자질선택 방법론의 최적화를 위해 자질의 수를 하나씩 줄여가며 성능을 비교 평가하였다. 자질의 수를 줄여 나가는 순서는 웨카시스템이 제공하는 분류성능에 대한 자질별 기여도 순위를 따랐다. 실험 결과 130개의 자질에서 1개씩 자질의 수를 줄여나가더라도 자질의 수가 56개가 될 때까지는 분류 성능의 차이가 없었다. 이후 55개의 자질을 사용할 경우에는 정확도에서 약간의 성능향상이 있은 후, 53개의 자질을 사용했을 때에는 오히려 약간의 성능하락을 보였으며, 최종적으로 34개의 자질만을 사용했을 때 가장 높은 성능을 보인다. 34개의 자질만을 사용했을 때 정확도는 68.4%였으며, 주관적 의미에 대한 정밀도는 0.92, 재현율은 0.403, f-측정값은 0.561이었다. 객관적 의미에 대한 정밀도는 0.618, 재현율은 0.965, f-측정값은 0.753이었다. 자질 수 34개를 기점으로 자질을 줄일수록 성능이 대체적으로 완만하게 하락하였으며, 14개 이하로 내려가면 성능이 급격히 하락하였다. (<표 3> 참조)

<표 3> 자질 수의 변화에 따른 분류성능 변화 추이[10)]

n.. of feat.	56	51	46	41	36	**34**	29	24	19	14
acc.	0.675	0.673	0.675	0.675	0.682	**0.684**	0.678	0.679	0.654	0.575
s.pr.	0.877	0.881	0.893	0.893	0.92	**0.92**	0.918	0.919	0.953	0.98
s.rc	0.406	0.4	0.397	0.397	0.4	**0.403**	0.39	0.394	0.324	0.152
s.f.	0.555	0.55	0.549	0.549	0.558	**0.561**	0.548	0.551	0.483	0.264
o.pr	0.614	0.612	0.612	0.612	0.617	**0.618**	0.613	0.614	0.593	0.54
o.rc	0.943	0.946	0.952	0.952	0.965	**0.965**	0.965	0.965	0.984	0.997
o.f.	0.743	0.743	0.745	0.745	0.752	**0.753**	0.75	0.751	0.74	0.701

최종적으로 34개의 자질만을 사용한 실험의 결과와 ‘bag of words’ 방식의 결과를 비교해보면 <표 4>와 같다.

<표 4> ‘bag of words’ 방식과 34개의 자질만을 사용한 방식의 성능비교

총 자질의 수	‘bag of words’ 방식	본 연구 제안 (34개 자질)
정확도	63.0159%	**68.4%**
주관적 의미 정밀도	0.639	**0.92**
주관적 의미 재현율	0.6	**0.403**
주관적 의미 f-측정값	0.619	**0.561**
객관적 의미 정밀도	0.623	**0.618**
객관적 의미 재현율	0.66	**0.965**
객관적 의미 f-측정값	0.641	**0.753**

10) 셀이 많은 표의 가독성을 위해 자질수를 ‘n. of feat’, 정확도를 ‘acc’, 주관적 의미의 정밀도를 ‘s.pr’, 주관적 의미의 재현율을 ‘s.rc’, 주관적 의미의 f-측정값을 ‘s.f.’, 객관적 의미의 정밀도를 ‘o.pr’, 객관적 의미의 재현율을 ‘o.rc’, 객관적 의미의 f-측정값을 ‘o.f.’로 표기하였음을 밝힌다.

본 연구에서 제안한 방법론은 최종적으로 34개의 자질만을 사용했을 때, ‘bag of words’ 방식과 비교하여 정확도에서는 5% 이상, 주관적 의미의 정밀도에서는 약 0.3이상의 성능향상을 보였으나, 재현율에서는 0.2 정도 낮은 성능을 보였다. f-측정값에서는 0.05 정도 하락하였다. 객관적 의미의 분류에서는 정밀도에서는 0.05 정도 낮은 성능을 보였으나 그 차이는 매우 작았으며, 재현율은 0.3 정도 대폭 상승하였으며, f-측정값에서는 0.11 정도의 성능이 향상되었다.

결론적으로 본 연구에서 제안한 ‘SW’와 트위터의 언어학적 특징을 고려한 자질에 기반하여 분류모델을 구성할 경우, ‘bag of words’ 방식보다 전체적인 정확도가 향상하며, 주관적 의미의 분류에 있어서는 특히 정밀도가 대폭 향상하나, 재현율은 낮아짐을 보았다. f-측정값은 약간 낮지만 그 차이가 그리 크지는 않았다. 객관적 의미의 분류에 있어서 정밀도는 약간 하락하지만 그 차이가 그리 크지 않으며, 재현율은 대폭 상승하였으며, 이로 인해 f-측정값도 우위에 있음을 볼 수 있었다.

끝으로 본 연구에서 제안한 방법론으로 주관적 의미의 분류에 가장 좋은 성능을 보인 34개의 자질은 다음과 같다.

> kack, !!, wtf, leid, gut, schön, lieb, fantastisch, schlecht, scheisse, cool, dank, geil, mag, voll, verbessern, vorteil, ärgern, beschimpfen, blöd, doof, dumm, hass, häßlich, idiot, kaputt, langweilig, nervig, problem, schade, schreck, schwack, stinken,

Ⅴ. 결론

본 연구에서는 문장의미를 주관적 의미와 객관적 의미로 자동분류하는 방법론을 제안하였다. 본 연구에서 제안한 방법론은 ‘SW’의 주요 단어를 자질로 선택하고, 트위터의 언어학적 특성을 반영하여 비속어, 이모티콘, 약어, 외래어 등도 자질로 함께 선택하여 ‘SVM’ 알고리즘을 적용한 방법론이다.

본 방법론은 ‘bag of words’ 방식과의 비교평가를 통해 정확성이 우월하고, 주관적 의미 분류의 정밀도에서 월등함을 보였다. 그러나 주관적 의미 분류의 재현율에서는 거의 0.2 정도나 낮은 성능을 보여, 이는 향후 보완해야 할 과제로 여겨진

다. 객관적 의미의 분류에서는 정밀도는 큰 차이가 없으나 재현율은 크게 높은 수치를 나타내었다. 그러나 이 수치는 분류 모델이 주관적 의미로 분류하지 못한 문장을 모두 객관적 의미라고 분류하였기 때문인 것으로 판단된다.

위 분석결과를 종합해 보면 본 연구에서 제안한 기계학습 방법론과 자질선택은 주관적 의미의 분류에 있어서는 정밀도가 상당히 높으나, 재현율은 'bag of words' 방식보다 떨어진다는 것이다. 이는 본 연구의 주제를 포함한 거의 대부분의 데이터 마이닝 분야에서 관찰되는 정밀도와 재현율의 역비례관계로 해석할 수 있을 것이다. 결국 이 문제를 해결하기 위해서는 정밀도를 최대한 높게 유지하면서 재현율을 최대한 높일 수 있는 접점을 찾아내야 한다. 이를 위한 하나의 방법으로 본 연구에서 시도하였던 'SW'의 확대적용 및 'SW'의 극성값 Polaritätswert에 따라 기계학습시 가중치를 부여하는 것 등이다. 또한 객관적 의미 분류의 정밀도를 높이기 위해 객관적 의미와 관련된 자질을 찾아내는 것도 추가적인 향후 연구의 과제라고 할 수 있다.

참고문헌

박신혜 (2011) 독일어 트위터 메시지의 논조분석을 위한 언어학적 특징연구, 성균관대학교 독어독문학과 석사학위논문

홍문표 (2009) 자동논조분석 시스템 개발을 위한 독일어 텍스트의 논조별 어휘출현양상 연구.『독일문학』. 50-3. 제 111집. 한국독어독문학회. 414-433.

홍문표, 신미영, 박신혜, 이형민 (2010) 구문분석과 기계학습 기반 하이브리드 텍스트 논조 자동분석, 언어와정보 14권 제 2호, 159-181

Alm, C., Roth, D. & R. Sproat (2005) Emotions from text: machine learning for text-based emotion prediction. Proceedings of Joint Conference on HLT/EMNLP, 579-586

Bird, S., Klein, E. & E. Loper (2009) Natural Language Processing with Python, O'Reilly, Sebastopol

Cortes, C & V. Vapnik (1995) Support-Vector Networks, Machine Learning, 273-297

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

Hatzivassiloglou, V. & K. Mackeown (1997) Predicting the Semantic Orientation of Adjectives, Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics, 174-181

Murphy, K.P. (2012) Machine Learning — A probabilistic Perspective, The MIT Press, Cambridge

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

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

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.

Popescu, A.-M. & O. Etzioni (2005) Extracting Product Features and Opinions from Reviews, Proceedings of HLT/EMNLP, 339-346

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

Turney, P.D (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), 417-424.

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

Automatische Klassifizierung der Subjektivität der Satzbedeutungen anhand von maschinellem Lernen

Hong, Mun-Pyo(Sungkyunkwan Univ.)

In der vorliegenden Arbeit wurde ein Ansatz zur automatischen Klassifizierung der Subjektivität der Satzbedeutungen bei Twitter vorgeschlagen. Unter dem Begriff Subjektivität der Bedeutung versteht man einen mentalen Zustand, der durch die Satzbedeutungen direkt oder indirekt ausgedrückt wird.

Der Ansatz stützt sich auf das maschinelle Lernen, das im Bereich der künstlichen Intelligenz für die Patternerkennung erfolgreich herangezogen wird. Bisher wurden verschiedene Algorithmen aus dem Bereich des maschinellen Lernens vorgeschlagen. Dazu gehören u. a. ‘Support Vector Machine (SVM)’, ‘Naive Bayesian’ und ‘Decision Tree’. In dieser Arbeit

wurde der SVM-Algorithmus herangezogen.

Um den SVM-Algorithmus effektiv anzuwenden, müssen passende Merkmale für das maschinelle Lernen ausgewählt werden. Die meisten bisherigen Arbeiten haben die Wörter im Lernkorpus als Merkmale verwendet. Diese Methode wird ‘bag of words’-Ansatz genannt. Sie zeigt einigermaßen befriedigende Ergebnisse in der Klassifizierung. Aber sie leidet unter dem sogenannten ‘overfitting’ Problem der Lerndaten.

Um diese Probleme zu vermeiden, wählen wir die Merkmale aus dem Senti-Wortschatz Lexikon aus. Das Lexikon enthält nur solche Einträge, die eine positive oder negative Polarität der Bedeutung aufweisen.

Im Twitter kann man aber oft vulgären Wörtern oder Emoticons begegnen, die Emotionen des Sprechers oder Verfassers aufweisen. Solche Wörter und Emoticons können auch sehr effektive Merkmale darstellen, die für das maschinelle Lernen erfolgreich angewendet werden.

Das Experiment zeigte, dass unser Ansatz um 5% besser bezüglich der ‘Accuracy’ als der ‘bag of words’-Ansatz ist. Auch bei der Klassifizierung der subjektiven Sätze übertrifft unser Ansatz den ‘bag of words’-Ansatz bezüglich der ‘Precision’. Aber der niedrige ‘Recall’-Wert bleibt als eine zukünftige Aufgabe zu lösen.

핵심어 : 주관성 Subjektivität, 기계학습 maschinelles Lernen,
논조분석 Sentimentanalyse, 트위터 Twitter, 자질선택 Merkmalsauswahl
필자 E-mail : skkhmp@skku.edu
논문투고일 : 2013. 7. 15 / 심사일 : 2013. 8. 14 / 심사완료일 : 2013. 9. 11