독일어 텍스트 논조자동분석*

홍문표 (성균관대)

1. 서론

텍스트 논조분석 Sentimentanalyse은 텍스트에 나타난 저자의 감정이나 의견 등을 추출하여 '긍정'과 '부정'으로 분류하는 자연언어처리 기술을 말한다. 본 논문에서는 휴대폰 분야의 리뷰 텍스트를 논조분류의 대상으로 한다.

논조분석은 의견마이닝 Opinion Mining, 감성분석 등으로 불리기도 하는데 최근 빅데이터 Big Data 처리의 중요성과 함께 학문적 관심이 높아지고 있는 분야이다. 논조분석은 주로 영어텍스트에 대한 분석기술 개발로 연구가 시작되었고 최근에는 그 적용언어가 독일어, 일본어, 중국어, 한국어 등으로 확대되고 있다.

지금까지 개발된 논조분석을 위한 방법은 크게 두 가지로 나눌 수 있다. 하나는 기계학습에 기반한 분류방법이며, 다른 하나는 의미분석 기반의 분류방법이다. 기계학습에 기반한 분류방법은 학습코퍼스에 존재하는 패턴을 기계학습 알고리즘을 적용하여 알아낸 후 새로 입력되는 문장이나 텍스트에 적용하여 논조분류를 시도하는 방식이다. 의미분석 기반의 분류방법은 기존의 자연언어처리 분야의 파이프라인 Pipeline 방식을 적용한 방법으로서 문장의 형태소분석, 구조분석, 의미분석을 거쳐 문장의 논조를 계산해내는 방식이다. 각 방식에는 장단점이 존재하는데, 본 연구에서는 기계학습에 기반한 방법론을 제안하고자 한다. 의미분석 기반 방식 대신 기계학습 기반 방식을 채택한 이유는 3장에서 좀 더 자세히 기술할 것이다.

논조분석을 위한 이 두 가지의 방식은 모두 문장의 논조를 전달하는 요소는 단어라는 전제에서 출발한다. 즉, 모든 단어는 긍정, 부정, 중립 등과 같은 극성 Polarität 정보를 가지고 있고, 문장의 논조 혹은 극성은 이러한 단어들의

* 이 논문은 2012년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012S1A5A2A01020054).

개별 극성의 합으로 이루어진다는 것이다. 그러나 문장의 논조 계산을 위해서는 단순히 문장을 구성하는 단어들이 가지고 있는 긍정과 부정 같은 극성정보 뿐만 아니라, 중립적인 논조를 지니는 단어도 문맥에 따라서는 긍정이나 부정과 같은 극성을 가지게 되는 특성도 고려해야 한다. 뿐만 아니라 긍정과 부정 같은 극성정보가 사전에 주어진 것으로 고정되는 것이 아니라 통사구조에 따라 변화가 가능하다는 점은 논조분석을 매우 어렵게 만드는 언어적 특성 중의 하나이다.

본 논문의 연구목적은 문맥에 따라 극성이 변화되는 현상을 기계학습에 반영한 논조자동분석 방법을 제안하는 것이다. 본 논문의 구성은 다음과 같다. 2장에서는 단어극성의 문맥의존성에 대해 설명한다. 이를 위해 먼저 단어의미의 극성 개념에 대해 논의한다. 3장에서는 논조분석을 위한 기계학습기반 방식과 의미분석 방식의 주요 기존연구를 소개하며, 단어극성의 문맥의존성을 다루는 방식의 장단점에 대해 논의한다. 4장에서는 본 논문에서 제안하는 단어극성의 문맥의존성 처리방안을 다룬다. 5장에서는 제안 방안의 타당성 검증을 위한 실험을 소개하며 끝으로 6장에서는 본 연구의 결과를 정리하고 향후 연구방향을 제시한다.

2. 단어극성의 문맥의존성

2.1. 단어 의미의 극성

단어 의미의 극성 Polarität이란 일반적으로 언어사회의 구성원들이 어떤 단어를 접했을 때 느끼는 감정을 긍정과 부정이라는 두 개의 극 Pol 중 어디에 더 가까운가에 따라 구분한 것이다. 어떠한 단어에 노출되었을 때 '기쁨', '놀람', '슬픔', '지루함', '재밌음', '역겨움' 등과 같은 다양한 감정유형이 발생할 수 있으나, 논조분석의 목적을 위해 여기서는 극성을 간단히 '긍정', '부정', '중립'으로 분류한다. 이와 같은 단순한 분류에 따르면 '기쁨', '재밌음' 등은 '긍정' 극성으로 분류될 수 있을 것이고, '슬픔', '지루함', '역겨움' 등은 '부정' 극성으로 분류될 수 있을 것이다. '놀람'은 문맥에 따라 '긍정'으로도 분류

될 수 있을 것이며 때로는 '부정' 극성으로 분류될 수도 있을 것이다.

극성정보는 품사의 측면에서 볼 때 개체나 사건의 성상 등을 나타내는 형용사와 밀접한 관련이 있지만 명사 및 동사와도 관련이 있다. 긍정극성을 가진 대표적인 형용사로는 'gut', 'schön', 'wunderbar', 'hübsch' 등이 있고 부정극성을 지닌 대표적인 형용사로는 'schlecht', 'häßlich', 'untreu', 'kaputt' 등을 들 수 있다. 명사의 경우도 긍정, 부정 극성을 지닌 단어들을 쉽게 찾아볼 수 있는데, 예를 들어 'Polizist'는 가치중립적인 의미의 단어이지만 'Bulle'는 부정극성의 명사이다. 동사의 경우는 Wiebe et al. (2005)의 연구에서 지적한 바와 같이 감정표출의 양상이 다른 품사들과는 다른데, 'hassen', 'freuen', 'mögen' 등과 같이 심리상태를 직접 언급할 수도 있고, 'weinen', 'lachen' 등과 같이 화행을 통해 간접적으로 심리상태를 나타낼 수 있다.

모든 단어에 극성정보를 부착하게 되면 논조분석을 위한 매우 유용한 데이터로 사용될 수 있다. 극성정보가 논조분석을 위해 어떻게 활용될 수 있는지 다음의 예를 통해 보도록 한다.

(1) Das Galaxy Note 3 ist das derzeit am besten(+) ausgestattete Smartphone auf dem Markt. Potente(+) Hardware, lange Laufzeit und ein tolles(+) Display - wenn man auch mit der etwas unhandlichen(-) Größe leben muss.

(2) Das LG G2 überzeugt(+) mit seinem nahezu randlosen Bildschirm, der nicht nur groß, sondern auch hochauflösend(+) ist. Zudem arbeitet das Smartphone aufgrund seiner starken(+) Hardware in der Praxis sehr performant(+). Zudem ist das Gerät mit knapp 500 Euro Straßenpreis vergleichsweise günstig(+) und damit empfehlenswert(+).

(1), (2)의 예에서 밑줄이 그어진 단어들은 모두 '긍정' 또는 '부정'의 극성을 가진 단어들이다. 그 외의 모든 단어들은 '중립'극성의 단어들로 볼 수 있다. 밑줄이 그어진 단어 중 (+) 표시가 되어 있는 단어는 긍정극성, (-) 표시가 되어 있는 단어는 부정극성의 단어들이다. (1) 예문은 긍정극성의 단어, 즉 (+) 표시가 되어 있는 단어가 모두 3개, 부정극성의 단어가 모두 1개 등장한다. 따라서 이 예문은 전체적으로 볼 때 긍정적 논조의 텍스트라고 분류할 수 있

을 것이다. (2)의 예문도 마찬가지로 총 6개의 긍정극성의 단어가 출현하고 부정극성의 단어는 하나도 출현하지 않는다. 따라서 이 텍스트를 긍정적 논조의 문장으로 분류할 수 있을 것이다. 이 예문들에서 본 것과 같이 단어의 극성정보를 알고 있다면 이를 문장이나 문서의 논조분류에 유용하게 활용할 수 있다.

논조분석 분야에서 위와 같은 극성정보를 부착한 사전을 선극성사전 prior-polarity lexicon[1]이라 부른다. 선극성사전은 어떤 단어가 문맥에 상관없이 근본적으로 가지고 있는 극성정보를 부착한 사전이다. 영어와 관련된 대표적인 선극성사전으로는 'SentiWordnet' (Esuli/Sebastiani 2006, Ohana/Tierney 2009), 'Wordnet-Affect' (Strapparava/Valitutti 2004), 'General Inquirer'[2] 등이 있으며, 독일어와 관련된 대표적인 선극성사전으로는 'Senti-Wortschatz' (Remus et al. 2010), 'German Polarity Clues' (Waltinger 2010) 등이 있다. 일반적으로 선극성사전에는 표제어에 대해 긍정, 부정 등과 같은 극성정보가 레이블로 붙어 있으며, 극성의 정도가 수치화되어 부착되어 있기도 하다.

최근의 논조분석 연구에는 선극성사전이 기계학습을 위한 자질로 많이 사용되고 있으나 단어의 도메인 의존적인 극성을 알아낼 수 없다는 단점이 있다. 본 연구에서는 기계학습을 위한 학습코퍼스를 직접 구축하여 학습코퍼스에서 학습자질을 추출하는 방식을 택하였다.

2.2. 문맥의존성

문서나 문장의 논조를 긍정, 부정으로 분류하는데 2.1에서 예를 든 것처럼 단순하게 긍정, 부정 어휘의 개수만을 파악하여 분류를 시도한다면 논조분석은 매우 단순한 작업이 될 것이다. 그러나 문장의 논조는 그리 단순하게 결정되지 않는다. 선극성사전에서 긍정 또는 부정 레이블이 부착되어 있는 단어도 문맥에 따라서는 긍정이 부정으로 바뀌기도 하고 부정이 긍정으로 바뀌기도 한다. 이렇게 단어의 극성을 뒤바꾸는 대표적인 단어는 부정어 Negation이다. 부정어가 문장에 사용될 경우 부정어의 수식영역 Skopus 안에 들어가는 단어

1) '선극성사전'을 다룬 독일어 관련 문헌에서도 'prior polarity lexicon'을 독일어로 번역하지 않고 그대로 사용하므로, 본 논문에서도 영어표기를 그대로 사용함을 밝힘.
2) http://www.wjh.harvard.edu/~inquirer/

들은 극성이 뒤바뀌게 된다.

(3)의 예문에서 긍정극성의 단어 'cool'은 부정어 'nicht'의 수식을 받음으로써 극성이 부정으로 바뀐다.

(3) Das neue Handy ist cool! (positiv)
(4) Das neue Handy ist nicht cool! (negativ)

단어의 극성을 바꾸는 단어로는 부정부사 'nicht'이외에도 부정관사 'kein-'도 있다. 부정관사 'kein-'과 극성을 가진 명사가 결합하면 극성이 반대극성으로 바뀐다.

(5) Das macht mir Spaß (positiv)
(6) Das macht mir kein Spaß (negativ)

부정어는 수식하는 단어의 바로 앞에 오는 것만은 아니다. 특히 독일어의 문장부정 같은 경우는 부정어가 수식하는 단어의 바로 앞에 오는 것이 아니라 문미에 온다. 이 경우 부정어의 수식영역에 들어가는 요소는 전체 동사구로 보는 것이 타당할 것이며, 따라서 (8)의 예문에서는 동사구의 핵심어인 'mag'이 부정어의 수식을 받는다고 할 수 있다.

(7) Ich mag das neue Design (positiv)
(8) Ich mag das neue Design nicht (negativ)

또한 문장의 주어가 부정어로 수식받을 때에도 부정어의 수식 영역은 인접한 단어가 아닌 넓은 수식영역을 갖는다고 보는 것이 합당할 것이다. 따라서 (9)의 예에서 부정관사 'kein'의 수식영역은 'Mensch'라기 보다는 동사구 'liebt das neue Design'으로 봐야할 것이다. 이 결과 긍정속성단어 'liebt'는 부정속성으로 바뀐다고 할 수 있다.

(9) Kein Mensch liebt das neue Design (negativ)

단어의 극성을 바꾸는 역할을 하는 단어는 부정어로만 국한되지 않는다. ‘kaum’, ‘wenig’, ‘selten’ 등의 부사도 준부정어의 역할을 하여 수식받는 단어의 극성을 뒤바꾼다 (예문 10, 11).

(10) Mir gefällt das neue Design kaum (negativ)
(11) Mir gefällt das neue Design ganz wenig (negativ)

접속법 구조도 부정어와 마찬가지로 단어의 극성을 바꾸는 역할을 한다. 다음 예문(12)를 보면 긍정극성의 단어 ‘freuen’이 사용되었으나 이 문장이 긍정 논조의 문장이라고 말하기는 어렵다. ‘freuen’은 조동사 ‘würde’와 함께 접속구문을 이루면서 긍정의 극성은 사라지고 중립 내지는 부정의 논조로 바뀌었다. (13)번 예문에서도 마찬가지로 접속법 형태로 굴절된 동사 ‘käme’ 때문에 긍정극성의 단어 ‘hochwertig’의 극성이 긍정으로 해석될 수 없다.

(12) Ich würde mich über eine neues Design freuen, weil ich gespannt bin wie Samsung das löst.
(13) Käme das Samsung Galaxy S4 in einem ähnlich hochwertigen Gehäuse wie beispielsweise das HTC One, würde es auf den ersten Platz der Handy-Bestenliste von CHIP Online landen.

그 밖에 ‘lösen’, ‘beheben’, ‘aufklären’, ‘heilen’, ‘abstellen’, ‘in Ordnung bringen’ 등과 같은 단어들은 모두 부정극성의 논항을 취해 긍정극성으로 바꿔주는 역할을 한다. Choi & Cardie (2008)의 연구에서는 이와 같이 논항명사의 극성을 바꿔주는 역할을 하는 단어들을 ‘content word negator’라고 부른다.

예문 (14)에서 동사 ‘fehlen’은 논항명사구내의 긍정극성어 ‘sinnvoll’과 결합하여 전체 문장의 극성을 부정으로 바꾸는 역할을 한다.

(14) In Pages fehlen einige durchaus sinnvolle Funktionen der Version 09

비슷한 예로 (15)의 예문에서 동사 ‘lösen’은 부정극성의 명사 ‘Problem’의

극성을 변화시켜 문장 전체의 극성을 긍정으로 바꿔준다.

(15) Update soll Problem mit SSD-Datenverlust lösen

마지막으로 전치사 'ohne'도 논항명사의 극성을 바꾸는 대표적인 전치사이다.

(16) Das iPhone 5 bietet einen hohen Bedienkomfort und kann dank seines handlichen Formats ohne Probleme mit einer Hand bedient werden

위 예문 (16)에서 'Probleme'는 부정극성의 명사이지만 전치사 'ohne'와 결합하여 'ohne Probleme'는 긍정의 극성으로 바뀐다. 다음 3장에서는 2.2장에서 소개한 극성변화현상을 다룬 기존 연구들을 살펴보기로 한다.

3. 관련연구

기계학습을 기반으로 하여 문장의미의 극성 분류를 시도한 가장 대표적인 연구로는 Pang et al. (2002)를 들 수 있다. 이 연구는 문서단위의 극성분류를 시도하였는데, 문서의 의미를 긍정논조와 부정논조 같은 극성에 따라 분류하는 것이 기존의 주제별 분류 Topik-Klassifizierung와 크게 다르지 않을 것이라는 가설이 연구의 출발점이었다. 이들은 이를 위해 학습코퍼스의 유니그램을 기계학습을 위한 자질로 사용하였다. 이 연구에서는 유니그램을 자질로 삼아 'bag of words' 방식의 벡터표상을 만들어 기계학습을 시도했다. 즉, 개별 자질들의 출현여부 내지는 출현빈도만이 중요할 뿐 자질 간의 관계는 고려되지 않았다.

이들의 실험결과 유니그램만을 자질로 사용하여 분류하는 것이 유니그램과 바이그램을 모두 사용하는 경우, 품사정보를 사용하는 경우, 형용사만을 사용하는 경우 등보다 더 높은 정확률을 보이는 것으로 나타났다. 특히 유니그램을 자질로 사용하더라도 유니그램의 출현빈도와 출현여부를 나누어 생각해볼

수 있는데, 이 연구에 따르면 출현빈도를 기계학습에 고려하는 것보다는 출현여부자체만을 고려하는 것이 오히려 분류에서 더 높은 성능을 보이는 것으로 보고되었다. 아마도 이는 상품평이나 리뷰 등과 같은 텍스트의 특성상 특정 어휘를 반복해서 사용하기 보다는 감정을 나타내는 어휘가 문서 내에서 단 한 번만 사용되더라도 저자의 감정이 강하게 나타나는 특성에 기인하는 것으로 생각된다.

이들의 연구에서는 문맥에 따른 단어의 극성변화현상을 본격적으로 다루지는 않았다. 이들은 부정어에 따른 극성의 변화만을 고려하였는데, 부정어 'not'과 문장종결부호 사이에 등장하는 모든 단어에 'NOT_'이라는 가공단어를 부착하여 학습을 시도하였다.

(17) I like this new Nokia model.
(18) I do not NOT_like NOT_this NOT_new NOT_Nokia NOT_model.

위와 같이 'NOT_'이 부착된 단어들 (예를 들어, 'NOT_like')은 부착되기 이전의 단어들과는 다른 별개의 어휘로 학습에 적용된다. 그러나 (17), (18)의 예에서 보는 바와 같이 부정어의 수식영역 Skopus에 대한 고려가 거의 없으므로 높은 성능향상을 가져오지는 못하였다.

문맥에 따른 극성의 변화를 본격적으로 다룬 것은 Wilson et al. (2005)의 연구이다. 이들의 연구도 기계학습방식 기반의 논조분류방법을 다루었다. 이들은 유니그램 이외의 자질로 통사적 속성도 함께 고려하였다. 이 연구에서는 유니그램을 학습자질로 고려할 때, 학습코퍼스에 등장하는 유니그램을 학습자질로 그대로 반영한 것이 아니라 감정사전정보를 반영하여 학습하였다.

이 연구에서는 또한 문장을 구조분석한 후 구조분석 결과를 기계학습에 반영하였다. 논조분석을 위해 특히 이들이 주목한 통사구조는 수식-피수식어 구조인데 어떤 단어가 의존구조상에서 다른 단어를 수식하고 있거나 다른 단어에 의해 수식받고 있는지의 여부도 학습을 위한 자질로 고려하였다.

부정어의 처리를 위해 이들은 단어의 극성을 변화시키는 '극성변화어' Polarity Shifter를 세 가지 종류로 나누었다.

- ‘일반극성변화어’ General Polarity Shifter
- ‘부정극성변화어’ Negative Polarity Shifter
- ‘긍정극성변화어’ Positive Polarity Shifter

일반극성변화어는 영어의 ‘not’, 독일어의 ‘nicht’와 같은 부정어 및 영어의 ‘never’, ‘hardly’, ‘rarely, 독일어의 ‘nie’, ‘kaum’ 등과 같은 부정부사를 가리킨다. 부정극성변화어는 부정부사와는 달리 동사나 명사 등과 같은 내용어들인데 긍정극성을 갖는 단어와 결합하여 긍정극성단어를 부정으로 변화시킨다. 예를 들어 앞서 소개한 (14)의 예문에서 부정극성변화어 ‘fehlen’은 긍정극성단어 ‘sinnvoll’이 포함된 명사구를 논항으로 취하면서 긍정극성을 부정으로 변화시키는 역할을 한다.

(14) In Pages fehlen einige durchaus sinnvolle Funktionen der Version 09

긍정극성변화어는 부정극성변화어와는 반대로 부정극성의 단어와 결합하여 부정극성의 단어를 긍정으로 변화시킨다. (15)의 예에서 ‘lösen’은 부정극성단어 ‘Problem’을 논항으로 취하여 부정극성을 긍정으로 바꾸게 된다.

(15) Update soll Problem mit SSD-Datenverlust lösen

이들은 어떤 단어가 이러한 극성변화어들인지의 여부를 모두 학습자질로 사용하였다. 이 결과 영어 문장 극성분류의 정확도를 약 4% 정도 향상할 수 있었다. 이들의 연구는 극성변화어의 종류를 부정어에만 한정하지 않고 다른 품사들로 확대했지만 부정어의 수식영역을 언어학적으로 타당하게 결정하지 않고 부정어를 중심으로 한 특정 윈도우 사이즈로 한정하여 계산한다는 단점이 있다.

이 밖에 Gamon (2004), Matsumoto et al. (2005), Joshi/Penstein-Rose (2009) 등의 연구에서도 기계학습기반 방식에서 부정어의 처리를 다루었는데 기본적으로 Wilson et al. (2005)의 연구와 처리방식이 크게 다르지는 않다.

의미분석기반 방식은 기존의 자연언어처리 응용시스템과 마찬가지로 언어

분석과정을 거쳐 문장의 논조를 계산해내는 방식을 의미한다. 일반적으로 언어분석과정에는 형태소분석, 구조분석, 의미분석 단계가 포함된다. 논조분석을 위한 의미분석은 일반언어학에서 익숙한 몬테규 의미론 Montague Semantik의 '의미합성성의 원리 Kompositionalitätsprinzip'를 따른다. 의미합성성의 원리를 논조분석에 적용하면 문장의 논조, 즉 문장의 극성은 문장을 구성하는 단어의 극성과 통사결합구조간의 함수관계이다.

Moilanen/Pulman (2007)은 각 통사규칙에 대해 극성정보해석규칙을 제시했다. 이들이 제안한 극성분석 규칙은 의미핵심어 Semantischer Kopf를 중심으로 의미정보의 상향전파 규칙, 극성정보의 변경규칙, 극성정보가 충돌할 때를 위한 조정 규칙 등으로 구성되어 있다.

의미분석기반 방식은 언어학적으로 타당한 분석과정을 거치지만 분석단계가 많기 때문에 중간분석과정에서 오류가 일어나면 극성분류에도 오류가 일어날 가능성이 높다는 치명적인 약점이 있다. 그럼에도 불구하고 입력문장이 매우 짧거나 정형화된 문장의 경우에는 도메인에 상관없이 비교적 높은 정확도를 보인다는 장점도 있다. 그러나 본 연구에서 다루는 문장들은 비교적 긴 문장들이어서 분석단계의 오류 가능성이 높기 때문에 본 연구에서는 기계학습기반의 분류방법을 채택하도록 한다.

4. 구조분석 기반 기계학습

4.1. 구조분석

본 연구에서는 부정어를 중심으로 한 단어극성의 변화현상을 문장논조분석에 반영하기 위해 구조분석 정보를 활용한다. 부정어 이외에도 문장의 의존관계를 기계학습에 반영하는 것이 논조분석의 성능을 향상시킨다는 연구결과가 영어에 대해서는 많이 보고되었지만, 본 연구에서는 구조분석 결과 중 극성변화어의 수식영역과 접속법2식 konjunktiv II 동사가 사용된 동사구조 정보만을 활용하였다.

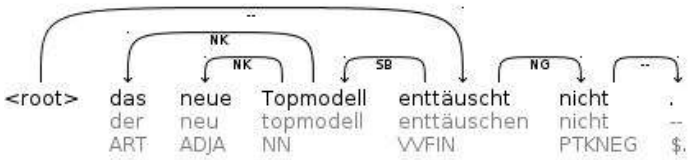
구조분석 결과를 기계학습에 반영하기 위해 슈투트가르트 대학에서 개발한

독일어 의존문법 파서를 활용하였다.[3] 학습코퍼스를 의존문법 파서로 분석한 후 파싱결과를 부정부사 'nicht', 'kein-' 등의 수식영역을 결정하는데 활용하였다. 또한 접속법 2식의 동사가 사용된 경우에는 조동사 'würde'가 사용된 경우와 동사가 접속법 굴절변화를 통해 사용된 경우로 나누었다. 그 밖에 극성변화전치사 'ohne'가 사용된 경우에는 'ohne'와 의존관계에 있는 논항명사를 찾아내었다. 마지막으로 긍정극성변화어와 부정극성변화어가 사용된 문장도 구조분석 결과를 활용하였다.

먼저 부정부사 'nicht', 'nie', 'kaum'이 사용된 문장은 구조분석을 통해 부정부사를 지배하는 동사를 찾아내고 동사들의 극성을 반대극성으로 바꾸었다. 그러나 본 연구에서는 선극성사전을 사용한 것이 아니므로 해당 동사의 선극성정보를 알 수가 없다. 따라서 Pang et al. (2002)의 연구에서처럼 해당 동사의 앞에 'not_'이라는 인공접두사를 부착하여 새로운 어휘로 바꾸었다.

예를 들어 다음의 문장 (19)는 <그림 1>과 같이 분석된다.

(19) Das neue Topmodell enttäuscht nicht.



<그림 1> : 'das neue Topmodell enttäuscht nicht.'의 구조분석결과

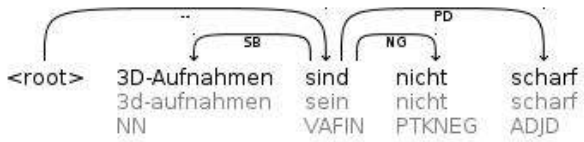
<그림 1>의 파싱결과에서 'nicht'는 자신으로 향하는 아크의 출발점, 즉 자신의 지배소인 'enttäuscht'를 의미적으로 수식하게 되는데 이 점을 기계학습에 반영할 때에는 문장의 벡터표상에 'enttäuscht' 대신에 새로운 자질인 'not_enttäuscht'를 사용한다.

단, (20)번 예문에서처럼 'nicht'의 지배소가 'sein' 동사일 경우에는 'sein' 동

3) 독일어 의존문법 파서는 http://de.sempar.ims.uni-stuttgart.de/에서 직접 테스트할 수 있다. 파서에 관한 좀 더 자세한 내용은 Seeker/Kuhn (2012) 참조.

사가 지배하는 형용사 (ADJD)가 ‘nicht’에 의해 의미적으로 수식받는다고 보고 형용사의 극성을 ‘not_’ 접두사를 사용하여 변경하였다. 이에 따라 <그림 2>의 경우 파싱결과를 기계학습에 반영할 때에는 ‘scharf’ 대신에 ‘not_scharf’를 사용한다.

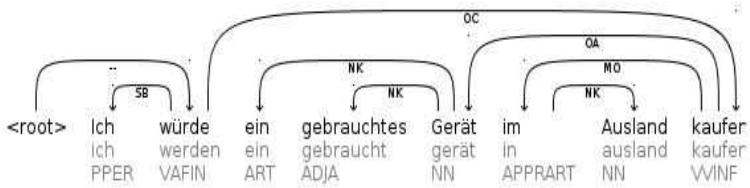
(20) 3D-Aufnahmen sind nicht scharf



<그림 2> ‘3D-Aufnahmen sind nicht scharf’의 구조분석결과

접속법 2식 동사가 사용된 경우는 앞서 언급한 바와 같이 조동사 ‘würde’가 사용된 경우와 본동사가 굴절된 경우로 나누었는데, 조동사 ‘würde’가 사용된 경우는 조동사와 호응하는 원형동사의 극성을 인공접두사 ‘not_’을 사용하여 변경하였다. 아래의 <그림 3>에서는 ‘kaufen’ 대신 ‘not_kaufen’이 기계학습에 적용된다.

(21) Ich würde ein gebrauchtes Gerät im Ausland kaufen

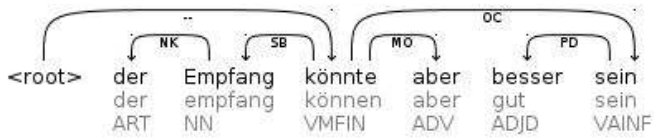


<그림 3> ‘Ich würde ein gebrauchtes Gerät im Ausland kaufen’의 구조분석결과

문장의 본동사가 접속법 굴절을 통해 변화되었을 때는 본동사의 원형에 ‘not_’접두사를 부착하여 극성변화를 기계학습에 반영한다. 단 접속법 굴절이 일어난 문장의 동사가 ‘würde’가 아닌 다른 조동사인 경우 조동사와 호응하는

동사의 극성이 바뀌게 되는데, 만약 이 동사가 ‘sein’ 동사라면 ‘sein’ 동사의 극성이 변화되는게 아니라 <그림 2>의 예와 유사하게 ‘sein’과 의존관계에 있는 형용사의 극성이 변화하게 된다. 따라서 아래 <그림 4>에서는 ‘besser’가 아니라 ‘not_besser’로 학습하게 된다.

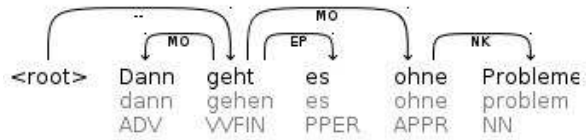
(22) der Empfang könnte aber besser sein



<그림 4> ‘der Empfang könnte aber besser sein’의 구조분석결과

극성변화전치사 ‘ohne’의 경우도 구조분석 후 ‘ohne’와 의존관계에 있는 논항명사의 극성을 바꾼다. 아래 예문의 경우 ‘Problem’ 대신 ‘not_Problem’이라고 하는 새로운 어휘가 긍정논조의 학습데이터로 반영된다.

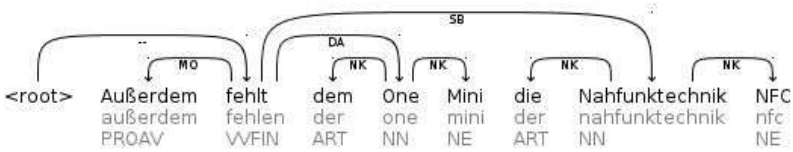
(23) Dann geht es ohne Probleme



<그림 5> ‘Dann geht es ohne Probleme’의 구조분석결과

그 밖에 긍정극성변화어와 부정극성변화어의 경우에는 구조분석을 수행한 후 술어의 의미적 특성에 따라 주어논항이나 목적어 논항의 극성변화를 시도하였다. 아래의 예문에서 부정극성변화어 ‘fehlt’는 주어논항의 극성을 변화시키는 단어이므로 문장의 주어인 ‘Nahfunktechnik’을 ‘not_Nahfunktechnik’으로 바꾼다.

(24) Außerdem fehlt dem One Mini die Nahfunktechnik NFC



<그림 6> ‘Außerdem fehlt dem One Mini die Nahfunktechnik NFC’의 구조분석결과

4.2. 기계학습

기계학습 maschinelles Lernen이란 데이터에 존재하는 패턴을 인식하기 위한 방법론들을 총칭하는 것으로서 지도기반학습 Überwachtes Lernen과 비지도기반학습 Unüberwachtes Lernen 방법으로 나눌 수 있다 (Murphy 2012: 10). 이 중 논조분석과 같은 데이터 마이닝 분야에서는 지도기반학습을 더 많이 적용하는데, 지도기반학습이란 학습데이터 Lerndaten로부터 학습데이터를 구성하는 입력 x와 출력 y간의 함수관계를 학습하여 새로운 입력 x’에 대한 출력 y’을 예측하는 것이라 할 수 있다. 기계학습 관점에서의 논조분석이란 문장 단위에 ‘긍정’이나 ‘부정’의 레이블이 부착되어 있는 학습데이터로부터 문장과 레이블 사이의 함수관계를 계산해내고 이 함수관계를 이용하여 새로운 입력문장의 레이블값을 예측하는 것이다.

기계학습을 위한 알고리즘은 ‘Support Vector Machine (SVM)’, ‘Naive Bayesian (NB)’, ‘Maximum Entropy (ME)’ 등이 있는데 Pang et al. (2002), Pang/Lee (2004), Wiebe et al. (2005), Wilson et al. (2005), Joshi/Pennstein (2009) 등의 연구에서 모두 ‘SVM’ 알고리즘이 극성분류에서 가장 높은 성능을 보이는 것으로 보고되었다. 따라서 본 연구에서도 기계학습 알고리즘은 ‘SVM’을 채택하였다.

‘SVM’ 알고리즘은 Cortes/Vapnik (1995)에 의해 처음 소개되었는데 특히 2개의 클래스를 분류하는데 탁월한 성능을 보이는 것으로 알려져 있다. 이 알고리즘은 ‘Support Vector’라고 불리는 인스턴스 instance들과 두 클래스를 나누는 가상 경계면 Hyperplane이 최대 절대값을 갖도록 함수를 구한다. 본 연구에서는 ‘SVM’ 알고리즘의 적용을 위해 기계학습 오픈소스 프로그램 ‘웨카 WEKA’를 사용하였다.

5. 실험

5.1. 학습코퍼스

본 연구에서 제안하는 독일어 텍스트 논조분석 방법론은 기계학습 방식을 따른다. 기계학습을 위한 학습코퍼스는 영어와는 달리 독일어의 경우에는 연구목적을 위해 공개된 것이 저자의 지식범위에서는 현재까지 없으므로 본 연구에서는 직접 구축하여 사용하였다.

학습코퍼스의 구축을 위해 독일에서 가장 활발히 휴대폰 신제품 리뷰 등이 소개되는 'www.connect.de' 사이트와 'www.chip.de' 사이트에서 문장들을 수집하였다. 수집 문장의 개수는 총 1,000개였으며, 문장당 평균 단어수는 13.57개였다. 각 문장은 일반 사용자가 아닌 휴대폰 분야의 전문가들에 의해 작성되어 비속어나 비문법구조 등이 사용되지 않는 비교적 깨끗한 문장구조를 가진다고 할 수 있다.

1,000개의 문장은 모두 긍정 또는 부정의 논조를 가진 문장들이고 중립 논조나 혼합논조의 문장은 배제하였다. 이 문장들에 '긍정' 또는 '부정'의 주석부착을 위해 2 명의 독어독문학과 대학원생이 투입되었다. 이 학생들은 모두 'ZMP' 이상의 독일어 수준으로 독일 대학에서의 수학경험도 있으므로 위 사이트의 기사들을 읽고 정확한 의미를 파악하는데 문제가 없다고 할 수 있다.

2 명의 학생이 각각 모든 문장에 대해 '긍정' 또는 '부정' 레이블을 부착한 후 두 작업자간의 태깅 일치도 Inter-Annotator Agreement Rate를 카파값 Kappa-Value으로 계산한 결과 0.956이 나왔다. 통상적으로 카파값 0.7 이상이면 상당히 일치율이 높은 데이터로 간주할 수 있으므로 이 태깅결과는 신뢰성 있는 학습데이터로 사용하기에 문제가 없을 것으로 판단되었다.

최종적으로 504개의 긍정논조 문장과 496개의 부정논조 문장이 학습데이터로 구축되어 기계학습과 성능평가에 활용되었다.

5.2. 실험

본 연구에서 제안한 방법론의 성능평가를 위해 실험을 수행하였다. 부정어 수식관계에 대한 구조분석 정보를 활용한 기계학습 방법론의 성능을 평가하

기 위해 Pang et al. (2002)에서 소개된 유니그램만을 자질로 한 'bag of words' 방식의 기계학습 방법을 베이스라인으로 삼았다.

베이스라인에 대한 비교대상으로 우선 'nicht', 'kein-', 'nie', 'kaum'와 같은 일반극성변화어만을 고려한 구조분석 결과를 기계학습에 반영하였다. 최종적으로는 일반극성변화어 및 긍정극성변화어, 부정극성변화어를 모두 고려한 방식의 성능을 평가하였다.

실험은 기계학습 오픈소스 프로그램인 웨카 Weka 버전 3.6.10을 사용하였다.[4] 그리고 논조분석기 모델링은 웨카가 제공하는 다양한 기계학습 알고리즘 중 데이터마이닝과 특히 논조분석 분야에서 가장 높은 성능을 내는 것으로 보고되고 있는 'SVM' 알고리즘을 사용하여 시도했다. 웨카 프로그램에서 학습데이터를 불러오기 위해서는 'ARFF'라고 하는 포맷으로 학습데이터를 구성해야 한다. 'ARFF' 파일은 분류하려고 하는 관계의 이름, 학습을 위한 자질의 정의 및 자질의 값 유형을 먼저 선언하고 이어서 데이터로 구성한다. 학습코퍼스에서 구축한 문장들은 단어벡터로 변환되어야 하는데 이를 위해 웨카 시스템은 'StringtoWordvector'라고 하는 전처리기를 제공한다. 이 전처리기를 적용하면 학습코퍼스에 문장단위로 정렬되어 있는 데이터를 단어벡터의 형식으로 변환할 수 있다.

실험은 학습코퍼스 1천 문장을 10등분하여 900문장으로 기계학습을 수행하여 논조분석 모델을 만들고 그 모델의 분류성능을 나머지 100문장에 적용하여 테스트하고 최종적으로 10등분한 것에 모두 테스트해보는 방식인 '10-fold' 교차검수 cross validation 방법을 시도하였다. 교차검수를 통한 최종성능은 분류의 정확도 Accuracy, 정밀도 Precision, 재현률 Recall, f-측정값 f-measure로 평가하였다.

분류의 정확도란 전체 분류의 대상이 되는 문장 중 정확하게 분류한 문장 개수의 비율이다. 예를 들어 1천개의 문장에 대해 논조분석기가 극성분류를 시도하여 총 700문장의 극성을 정확하게 예측했다고 하면 이 분석기의 정확도는 70%이다. 많은 논조분석 관련연구에서는 분석기의 성능을 언급할 때 정확도만을 언급하기도 한다.

이와 유사한 개념인 정밀도는 어떤 클래스에 대해, 예를 들어 '긍정' 클래스라고 한다면 '긍정'으로 분류기가 분류한 것 중 실제로 '긍정'인 것의 비율

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

이다. 어떤 분석기가 총 500문장을 '긍정'이라고 분류했는데 실제로 이 중 300문장만이 '긍정'이라고 한다면 이 분석기의 '긍정' 클래스에 대한 정밀도는 60%이다. 재현률은 어떤 클래스의 전체 개수 중 정확하게 찾아낸 개수의 비율이다. 예를 들어 500개의 '긍정' 클래스가 존재하는데 논조분석기가 200개만의 올바른 '긍정' 클래스를 찾아냈다면 이 분석기의 재현율은 40%이다. f-측정값은 정밀도와 재현율을 정규화한 값인데, 이를 구하는 수식은 '(2 × 정밀도 × 재현율) / (정밀도 + 재현율)'이다.

5.3. 평가결과

위와 같은 평가 기준에 따라 베이스라인 시스템, 즉 학습코퍼스의 유니그램만을 고려한 Pang et al. (2002) 방식의 시스템을 평가한 결과 77.6%의 정확도가 측정되었다. Pang et al. (2002)의 연구에서 동일한 방식으로 영어문서의 논조분석을 수행한 결과 약 82% 정도의 정확률이 측정된 바 있는데, 본 실험의 결과도 독일어 문장에 대해 적용한 차이는 있지만 영어의 그것과 크게 다르지 않았다. 긍정 논조의 경우 정밀도는 0.783, 재현률은 0.768, f-측정값은 0.776이 측정되었다. 부정 논조의 경우는 정밀도는 0.769, 재현률은 0.784, f-측정값은 0.776으로 측정되었다.

<표 1> 베이스라인 시스템 성능평가 결과

정확도		77.6%
긍정	정밀도	0.783
	재현률	0.768
	f-측정값	0.776
부정	정밀도	0.769
	재현률	0.784
	f-측정값	0.776

이어서 학습코퍼스를 구조분석한 결과에서 'nicht', 'kein-', 'nie', 'kaum' 등과 같은 일반극성변화어의 수식영역만을 고려하여 기계학습에 반영한 분류모델의 성능은 다음 <표 2>와 같았다.

<표 2> 일반극성변화어를 고려한 모델 성능평가 결과

정확도		79.7%
긍정	정밀도	0.805
	재현률	0.788
	f-측정값	0.796
부정	정밀도	0.789
	재현률	0.806
	f-측정값	0.798

실험 결과 일반극성변화어를 고려한 모델은 베이스라인 대비 정확도는 2.1% 향상되어 79.7%를 기록하였다. 긍정극성의 정밀도는 2.2% 향상하여 0.805였으며 재현률은 2% 향상하여 0.788로 측정되었다. f-측정값도 2% 향상하여 0.796을 기록하였다.

부정극성에 대해서도 전체적으로 성능이 향상됨을 볼 수 있었는데 정밀도는 2% 향상하여 0.789였으며, 재현률은 2.2% 향상하여 0.806을 기록하였다. f-측정값은 2.2% 향상하여 0.798로 측정되었다.

마지막으로 본 논문에서 제안하는 방법인 일반극성변화어 및 긍정극성변화어, 부정극성변화어를 모두 고려한 방법론의 성능평가 결과는 다음 표 3과 같았다. 이 방법론은 정확도 측면에서는 베이스라인 시스템 대비 3.2%가 향상되었으며, 긍정논조의 정밀도는 2.8% 향상되어 0.811을 기록했다. 재현률은 4% 향상되어 0.808로 측정되었고, f-측정값은 3.3% 향상되어 0.809를 기록했다. 부정극성에 대해서는 정밀도가 3.6% 향상하여 0.805였으며, 재현률은 2.4% 향상하여 0.808을 기록했고, f-측정값은 3.1% 향상하여 0.807로 측정되었다.

이 방법론은 일반극성변화어만을 고려한 모델과 비교하여도 정확도의 측면에서는 1.1%, 긍정극성의 정밀도 측면에서는 0.6%, 재현률은 2%, f-측정값은 1.3%의 성능향상이 관찰되었다. 부정극성의 정밀도 측면에서도 1.6%, 재현률은 0.2%, f-측정값은 0.9%만큼의 성능이 향상되는 것으로 보인다.

<표 3> 모든 극성변화어를 고려한 모델 성능평가 결과

정확도		80.8%
긍정	정밀도	0.811
	재현률	0.808

	f-측정값	0.809
부정	정밀도	0.805
	재현률	0.808
	f-측정값	0.807

실험 결과 본 연구에서 제안한 구조분석 결과 중 부정어의 수식영역을 고려하여 기계학습을 시도하는 방법론은 기존의 유니그램만을 고려한 방법론보다 정확도, 정밀도, 재현률, f-측정값에서 성능향상을 가져올 수 있음을 확인하였다.

본 연구에서 제안한 기계학습에 기반한 분류시스템은 일종의 블랙박스 시스템이라고 할 수 있다. 즉, 어떤 문장이 어떠한 이유로 특정 클래스로 분류되었는지를 알기가 거의 불가능하다. 규칙기반의 시스템이었다면 문장극성 오분류의 원인을 분석할 수 있겠지만, 기계학습 기반의 시스템은 이러한 원인분석이 매우 어렵다는 단점이 있다. 그럼에도 불구하고 본 방법론이 잘못 분류하는 문장을 살펴보면 향후 보완해야할 점을 추론해 볼 수 있다.

(25) Wie schon beim Galaxy S4 Active bleibt auch beim Z1 der Klinkenanschluss offen, kann also auch unter Wasser genutzt werden.

위 예문 (25)는 긍정논조로 분류될 수 있는 문장이다. 그럼에도 본 연구에서 개발한 분류기는 위 문장을 부정논조로 분류하였다. 위 문장을 어떤 계산과정에 의해 부정으로 분류하였는지 정확하게 파악하기는 어렵지만 위 문장에 긍정극성을 확실하게 전달하는 단어가 없음은 분명하다. 위 문장에서는 'kann also auch unter Wasser genutzt werden' 부분이 긍정적인 논조를 전달하는 부분인데, 이 부분이 긍정적인 논조 혹은 정보라는 것을 알기 위해서는 세상지식 Weltwissen이 반영되어야 한다. 앞으로는 세상지식을 어떻게 논조분석의 성능향상을 위해 반영할 것인지에 대한 논의가 좀 더 이루어져야 할 것이다.

6. 결론

본 연구에서는 단어의 극성이 문맥에 따라 변화하는 현상을 기계학습 기반의 논조분석에 반영하기 위해 구조분석 결과를 도입하는 방법론을 다루었다. 본 연구에서 고려한 문맥은 일반극성변화어, 긍정극성변화어, 부정극성변화어 등이었다. 이러한 극성변화어들은 수식영역에 따라 피수식어의 극성을 변화하게 만드는 역할을 하였다. 이 점을 기계학습에 반영하기 위해 학습코퍼스의 유니그램만을 학습자질로 선택하는 기존의 방법론에 구조분석 결과를 추가적으로 반영하여 베이스라인 대비 3.2%의 정확도 향상을 이루었다.

기계학습기반의 논조분석 방법은 그 장단점이 명확하다. 우선 입력문장의 문법성, 비문법성에 크게 구애받지 않는 강건함 Robustheit이 가장 큰 장점이다. 따라서 본 연구에서 제안한 방법론은 트위터나 블로그 게시물 등과 같이 철자법이나 문법구조가 잘 지켜지지 않는 입력텍스트에 대해서도 어느 정도의 분류정확도를 보장할 가능성이 높다.

그러나 큰 단점 중의 하나는 학습코퍼스의 도메인에 매우 의존적이라는 점이다. 본 연구에서 개발한 휴대폰 분야 논조분석 모델은 아마도 자동차 분야에 적용하게 되면 분류의 정확도가 훨씬 낮아질 것이다. 이 문제를 해결하기 위해서는 도메인별 코퍼스를 대량으로 구축해야한다. 그러나 이것 또한 많은 시간과 노력을 필요로 하는 작업이므로 순수하게 학습코퍼스에만 의존하는 방식은 실용적이지 못할 것이다.

이 문제를 해결하기 위해서는 'Senti-Wortschatz'나 'German Polarity Clues' 등과 같은 선극성사전의 어휘를 학습의 자질로 활용하는 것이다. 이 사전에 등재되어 있는 엔트리들은 도메인에 상관없이 긍정, 부정 속성을 지닌 어휘들일 것이기 때문에 이 어휘들을 학습자질로 하여 분류모델을 만든다면 도메인 의존성이 낮아질 가능성이 있다. 그럼에도 불구하고 이 방법은 단어 극성의 도메인 의존성은 전혀 파악할 수 없다는 단점도 있다.

따라서 향후 연구방향은 학습코퍼스와 선극성사전에서 학습자질을 고루 추출하여 도메인 의존성이 낮은 분류모델을 고안하는 것이 될 것이다.

참고문헌

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.

Cortes, C/Vapnik, V. (1995): Support-Vector Networks. In: *Machine Learning*, 273-297.

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

Moilanen, K./Pulman, S. (2007): Sentiment composition. In: *Proc. of RANLP-2007*, 378-382.

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

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

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

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

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

Seeker, W./Kuhn, J. (2012): Making Ellipses Explicit in Dependency Conversion for a German Treebank. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey: European Language Resources Association (ELRA), 3132-3139.

Strapparava, C./Valitutti, A. (2004): WordNet-Affect: an affective extension of WordNet. In: *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. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC), 1638-1642.

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

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

Zusammenfassung

Automatische Sentimentanalyse deutscher Texte

Hong, Munpyo (Sungkyunkwan Univ.)

Die vorliegende Arbeit befasst sich mit der Sentimentanalyse deutscher Texte. Unter dem Begriff „Sentiment“ versteht man im Bereich der Sprachverarbeitung eine positive oder negative Neigung oder Stimmung der Sätze. Das Sentiment eines Satzes wird durch die Polarität der Wörter im Satz bestimmt. Die Polarität eines Wortes ist normalerweise apriorisch. Aber die Polarität des Wortes kann in einem bestimmten Kontext verändert werden. Vor allem wenn ein Wort mit einer bestimmten Polarität negiert wird, wird die Polarität umgesetzt.

In den meisten Arbeiten über die Sentimentanalyse deutscher Texte wird das Phänomen nicht berücksichtigt. Sie versuchten den sogenannten „Bag of Words“ Ansatz mit Unigrams als Merkmale für das maschinelle Lernen. In diesem Ansatz werden nur die Wörter, d.h. Unigrams als Merkmale verwendet, ohne dass beim maschinellen Lernen auf die Satzstruktur Rücksicht genommen wird. Hier werden die negierenden Wörter wie „nicht“, „kein“, „nie“, und „kaum“ nur als ein Merkmal behandelt, das sich nicht von anderen Wörtern unterscheiden lässt.

Um dieses Problem zu lösen, schlugen wir in der vorliegenden Arbeit eine neue Methode für das maschinelle Lernen vor. Um den Skopus von den negierenden Wörtern zu bestimmen, wurde die Parsing Resultate eines Dependenzparsers herangezogen. Durch die Einführung des Negationsskopus ins maschinellen Lernen wurde die Accuracy des Sentimentanalysesystems um 3.2% auf 80.8% erhöht.

[검색어] 논조분석, 감성분석, 기계학습, 선극성사전, 데이터마이닝, 빅데이터
Sentimentanalyse, Opinion Mining, Maschinelles Lernen, Prior-Polarity Lexicon, Data Mining, Big Data

홍문표 110-745
서울특별시 종로구 명륜동 3가 53
성균관대학교 독어독문학과
skhmp@skku.edu

논문 접수일: 2013. 10. 31
논문 심사일: 2013. 11. 29
게재 확정일: 2013. 12. 10