텍스트 논조결정 요인에 대한 전산언어학적 기초연구

홍문표(성균관대)

1. 서론

(1) “Mein Handy ist leider nicht mehr zu benutzen, da beim Aufklappen das Display dank eines Wackelkontaktes abschaltet. Dieser Fehler leider sehr verbreitet und nur durch einen sehr teuren Ersatzteil zu beheben. Also ab in die Tonne.”

위 짧은 텍스트는 독일의 한 휴대폰 관련 리뷰사이트에 어떤 독일 사용자가 국내기업에서 생산된 한 휴대폰에 관해 작성한 댓글이다. 이 글의 작성자는 자신이 구입한 휴대폰의 결함, 즉, 접촉불량에 대해 자신의 부정적인 감정 혹은 견해를 다양한 어휘와 구문(‘leider nicht mehr zu benutzen’, ‘Fehler’, ‘sehr teuer’, ‘ab in die Tonne’)을 통해 직·간접적으로 표출하였다.

블로그, 게시판 등과 같은 인터넷 공간에서 형성되는 여론을 버즈 Buzz 혹은 구전 口傳이라고 한다. 오늘날 많은 기업들은 이와 같은 버즈 정보를 매우 중요한 정보로 받아들인다. 버즈를 통해 생성되는 여론 혹은 입소문은 기업의 브랜드가치 및 소비자의 구매결정에 매우 중요한 역할을 하기 때문이다.[1)]

현재 국내·외의 많은 구전분석 업체는 수많은 인력을 동원하여 각 기업에게 해당 기업의 제품 및 이미지에 대한 구전을 조사하여 정보를 제공하고 있다.[2)] 그러나 많은 인력으로도 제어할 수 없을 정도의 방대한 양의 정보가 인터넷

1) ‘Forrester Research’의 연구에 따르면 입소문은 신문, TV, 라디오 광고 등과 같은 전통적인 형태의 광고보다 전반적인 신뢰도가 높으며, 전적으로 신뢰하는 비율이 높다고 한다. 또한 ‘Jupiter Research Executive Survey’ (2005.10)에 따르면 광고주의 90%이상이 고객들의 추천이 다른 사람들의 제품 구매 결정에 영향을 미친다고 생각한다고 한다.

2) 2008년 3월 7일자 조선일보 기사 “인터넷 댓글을 분석하라 ... 위기 탈출 길이 보인다” 참조.

상에 존재하고 시시각각으로 변화하고 있기 때문에, 텍스트들의 논조를 분석하는 작업의 자동화 필요성이 대두되고 있다. 이를 위해 영어의 경우에는 텍스트의 논조를 자동분석 시스템을 통해 파악하여 기업에 실시간으로 제공하는 서비스가 제공되고 있다.[3] 그러나 독일어의 경우 아직 이와 같은 시스템이 개발되어 있지 못한 실정이다. 따라서 국내의 구전조사 분석기관 및 업체들은 독일어권 정보수집에 많은 어려움을 겪고 있다. 이러한 문제를 해결하기 위한 독일어 텍스트 논조자동분석시스템의 개발이 시급한 실정이다.

본 논문은 텍스트 논조자동분석 시스템 개발을 위한 전산언어학적 토대연구에 관한 것이다. 보다 구체적으로는 텍스트 논조자동분석을 위해 고려해야 할 언어학적 요인을 규명하고, 이러한 요인들을 어떻게 가장 효율적으로 자동분석에 반영할 수 있을지에 관한 논의를 진행하게 될 것이다. 현재 가장 높은 성능을 올릴 수 있는 것으로 알려진 기계학습 Maschinelles Lernen 기반의 자동분석 방법에서 필요한 독일어의 특성이 반영된 기계학습의 자질 Merkmal을 알아내는 것이 본 연구의 목적이다.

이를 위해 본 논문은 다음과 같이 구성되어 있다. 2장에서는 어떤 대상에 대한 감정 혹은 논조가 텍스트 상에서 어떻게 언어적으로 실현되는지에 대한 선행연구를 살펴보고자 한다. 특히 Fries(2006)의 연구를 중심으로 감정 혹은 논조가 독일어 텍스트에서 실현되는 양상에 대한 논의를 진행한다. 3장에서는 현재 영어권을 중심으로 활발히 진행되고 있는 텍스트 논조 자동분석의 방법론을 소개한다. 특히 기계학습에 기반한 방법론을 다루게 될 것이다. 4장에서는 텍스트의 논조와 관련된 독일어의 특성을 고려하여 어떤 언어학적 자질들을 기계학습시 고려해야하는가에 대한 논의를 진행할 것이다. 끝으로 5장에서는 본 연구와 관련하여 향후 진행해야할 연구에 대해 언급할 것이다.

3) 대표적인 영어텍스트 논조자동분석 시스템으로는 미국 Corpora사의 Sentiment 시스템이 있다.

2. 텍스트에 나타난 감정에 대한 연구

감정 Emotion에 대한 연구는 심리학, 의학, 철학, 언어학, 사회학 등의 학문 분야를 아우르는 제 학문간의 협력 interdisziplinär 연구분야이다. 언어학 분야에서 텍스트에 나타난 감정표현에 대한 대표적인 연구로는 Osgood et al.(1957)을 들 수 있다. Osgood et al.(1957)은 '좋음/나쁨', '아름다움/추함', '친절함/잔인함', '정직/비정직' 등과 같은 어휘의미의 평가적 요소 wertende Faktoren에 주목하였다. 이 연구의 영향으로 텍스트논조에 대한 자동분석 분야에서는 이러한 어휘의미의 평가적 요소를 의미적 경향 semantische Orientation이라고 부르기도 한다.

Fries(2007)는 인간의 감정이 부호화되는 양상에 대한 연구를 수행하였다. 그의 연구에 따르면 감정은 문장의 형태로 코드화 Kodierung 될 수 있는데, 이 경우 주어가 감정의 경험자 Emotionsträger이다.[4)]

(1) Ich schäme mich zu Tode!
(2) Ich habe eine riesige Angst
(3) Ich habe seit langem Angst vor der Operation
(4) Du zitterst vor Angst
(5) Er bedauerte das Malheur immer wieder

(1)~(5)에서 감정은 모두 술어부 ('sich zu Tode schämen', 'eine riesige Angst haben', 'seit langem Angst vor der Operation haben', 'vor Angst zittern', 'das Malheur immer wieder bedauern')를 통해 표현되고 있다. 이 술어들은 주어, 즉 감정 경험자의 주관적이고 심리적인 상태를 나타내고 있으며, 상황에 따라서는 'zittern'과 같은 주어의 운동상태를 나타내기도 한다.

감정을 술어부를 통해 나타내는 경우, 감정의 강도 Intensität를 표현할 수도 있으며 ('zu Tode', 'riesig'), 특정 감정을 주어가 갖게 되는 요인 ('vor der

4) 이익환/이민행(2005)의 연구에 따르면, 독일어에서 경험자 논항이 주어로 실현되는 심리동사는 모두 게르마넷 Germanet 구조상에서 동사 'empfinden'을 최단거리 공통 상위어로 갖는다고 함.

Operation', 'vor Angst')도 표현할 수 있고, 감정을 갖은 기간 ('seit langem') 및 빈도 ('immer wieder')도 표현할 수 있게 된다.

반면, 목적어가 감정의 경험자로 인코딩되는 경우도 많은데, 이에 속하는 대표적인 동사들은 'ängstigen, beeindrucken, belasten, belustigen, bewegen, ehren, erfreuen, ergreifen, erschüttern, faszinieren, gefallen, langweilen, verletzen, verwundern' 등이다. 본 논문에서 감정의 경험자가 인코딩되는 형식은 주요 연구테마가 아니므로, 이에 대한 자세한 논의는 여기서 하지 않도록 하고, 관련주제는 이익환/이민행(2005)를 참조하기 바람.

Fries(2007)에 따르면 텍스트를 통해 감정을 표현하는 두 번째 방법은 비명제적인 nicht-propositionale 수단을 통한 것이다.

(6) Führest du (nur) schneller!
(7) Fährst du (aber) schnell!
(8) Was für ein Glück du hast!
(9) Dass du immer Glück haben musst!

감정을 명제를 통하지 않고 표현하는 첫 번째 방법은 '직설법', '접속법' 등과 같은 동사유형을 통해 표현하는 형태론적인 방법이 있다. 위의 예문 (6)에서는 화자의 감정이 동사형태, 즉, 접속법 형태를 통하여 간접적으로 표현되고 있는데, 청자가 조금 더 빨리 운전했으면 하는 아쉬움이나 불만 등이 표출되고 있다.

비명제적인 감정표현 수단의 두 번째 방법은 어순 등과 같은 통사적인 방법이다. 위의 예문 (7)에서 화자는 동사를 문두에 배치하는 구조를 통해 놀라움 등의 감정을 표현하고 있다.

세 번째 방법은 'aber', 'nur', 'was', 'dass' 등과 같은 어휘를 통한 방법이다. 이러한 어휘들은 화자 혹은 저자의 감정을 표출하는데 형태/통사적인 방법들과 더불어 자주 함께 등장하는 요소이다. 감정을 표현하는 어휘 요소로 대표적인 것은 'leider', 'erfreulicherweise', 'beängstigenderweise', 'erstaunlicherweise' 등과 같은 문장부사이다.

네 번째로는 완전한 문장이 아닌 문장의 한 부분 Satzfragment만을 사용하

여 감정을 표현하는 방법을 들 수 있다.

(10) Superfilm!
(11) Du Ignorant!
(12) Entsetzlich für mich!
(13) Mein Beileid!

(10)~(13)에서는 주로 문장의 동사를 생략하였지만 감정의 경험자는 화자나 저자로 한정되고 감정의 유발자는 문맥 등을 통해 파악하게 된다.

마지막으로 발화의 억양, 강세 등을 고려하는 방법이 있으나, 텍스트에서는 이러한 정보가 전달될 수 없으므로 본 연구에서는 고려하지 않기로 한다.

Fries(2007)의 연구에서는 언어정보를 통해 표출되는 내적 감정의 유형을 다음과 같이 네 가지의 종류로 구별한다.[5)]

- 행복/불행(Behagen)

행복/불행의 감정은 우선 문장의 감탄사 Interjektion를 통해 표출될 수 있다. ‘au, brr, hu‘등이 행복/불행 등과 밀접하게 관련된 대표적인 감탄사라고 할 수 있다.

이 유형에 속하는 대표적인 명사로는 ‘Angst’, ‘Beklemmung’, ‘Leid’, ‘Erleichterung’, ‘Freude’, ‘Glück’, ‘Heiterkeit’, ‘Wonne’ 등이 있으며, 동사로는 ‘ärgern’, ‘betrüben’, ‘deprimieren’, ‘erzürnen’, ‘peinigen’, ‘provozieren’, ‘quälen’, ‘rädern’, ‘triezen’, ‘verbittern’ 등이 있다. 또한 형용사로는 ‘deprimierend’, ‘nervtötend’, ‘entzückend’, ‘fröhlich’등을 들 수 있다.

- 감정이입(Empathie)

어떤 대상 혹은 상태에 대한 감정이입도 ‘e, hmm, autsch’ 등과 같은 감탄사를 통해 표출될 수 있다. 감정이입이 의미를 전달하는 대표적인 명사로는 ‘Mitleid’, ‘Mitfreude’, ‘Nächstenliebe’ 등을 들 수 있으며, 동사로는 ’beunruhigen’, ‘einschütern’, ‘aufheitern’, ‘aufrichten’, ‘erquicken’, ‘trösten’ 등이 있

5) Fries(2007)의 이러한 구분은 저자 자신이 밝히듯이 문장의 술어부와 같은 언어정보를 통하여 전달되는 감정유형을 하나도 빠짐없이 포함한다고 할 수는 없을 것이나, 대표적인 감정유형은 모두 포함된다.

다. 형용사로는 ‘goldig’, ‘niedlich’, ‘rührend’ 등이 있다.

- 가치평가(Wertschätzung)

어떤 대상 및 상태에 대한 가치평가에 사용되는 대표적인 감탄사로는 ‘hm, ih, nana, pah, pfui, hm’ 등이 있다. 이와 관련된 대표적인 명사로는 ‘Ekel’, ‘Scham’, ‘Verachtung’, ‘Bewunderung’, ‘Ehrfurcht’, ‘Hochachtung’ 등이 있으며, 동사로는 ’kränken‘, ‘schämen‘, ‘verachten‘, ‘idealisieren‘, ‘verklären‘ 등을 꼽는다. 가치평가를 나타내는 형용사로는 ’abscheulich’, ’eklig’, ’süß’ 등이 있다.

- 관심(Interesse)

가치평가를 하기 위해서는 관심이 있어야 한다는 점에서 관심은 가치평가 유형과 매우 흡사하다. 관심을 나타내는 대표적인 감탄사로는 ‘he, na’ 등을 들 수 있고, ‘Beileid’, ‘Eifersucht’, ‘Furch’, ‘Neid’, ‘Hoffnung’, ‘Liebe’ 등이 관심을 나타내는 대표적인 명사이다. 관심을 나타내는 대표적 동사로는 ’begeistern’, ‘ereifern’, ‘erglühen’, ‘fesseln’, ‘lähmen’, ‘versteinern’ 등이 있다. 관심의 대표적인 형용사로는 ‘atemberaubend’, ‘aufregend’, ‘empörend’, ‘ermüdend’, ‘langweilig’ 등이 있다.

Fries(2007)의 이와 같은 감정유형에 따른 어휘의 분류는 논쟁의 여지는 있으나 텍스트 논조분석을 위해 유용하게 사용될 수 있을 것이다. 특히 ‘가치평가’와 ‘관심’ 유형으로 분류되는 어휘의미의 특성을 좀 더 분석하여, 이를 기반으로 논조정보가 수록된 감정사전을 구축한다면 텍스트 논조 자동분석을 위한 매우 중요한 언어자원으로 사용될 수 있다.

이와 관련하여 게르마넷 GermaNet과 같은 어휘의미망을 활용하는 방안을 생각해볼 수 있다. 영어의 경우 이미 기존의 워드넷에 감정정보를 보충한 감정워드넷 ‘SentiWordnet’이 개발되어 많은 연구에 활용되고 있다.[6] 그러나 독일어에는 이러한 감정정보와 관련된 어휘의미망이 존재하지 않으므로, 게르마넷에서 위 분류의 ’가치평가‘와 ’관심‘과 관련된 어휘들의 의미를 분석해본다면 독일어 텍스트 자동논조분석에 유용하게 사용될 것이다.

6) Vgl. Esuli, A./Sebastiani, F. (2006).

3. 텍스트논조 자동분석

자연언어처리 분야에서는 텍스트 논조의 자동분석을 감정분석 Sentiment-analyse 혹은 의견마이닝 Opinion Mining이라고 부른다. 이 분야는 다른 응용분야와는 달리 2000년대에 들어서야 비로소 많은 연구자들의 관심을 받게 된 최신 연구분야이다. 이 분야의 연구 동향은 텍스트 내용의 논조를 분석하기 위해 언어학 지식에 기반한 규칙기반 방식으로 시작하여 현재는 기계학습 Maschinelles Lernen에 기반한 방식이 연구방법론의 주류를 이루고 있다.

기계학습에 기반한 연구는 태그드코퍼스와 같은 지식 혹은 정보를 필요로 하느냐 아니냐에 따라 지도학습방법 Überwachtes Lernen[7], 및 비지도학습방법 Unüberwachtes Lernen[8]으로 나눌 수 있다. 텍스트 또는 문장단위로 논조정보가 부착된 코퍼스를 학습을 위한 자료로 사용하는 지도학습방법론은 비지도학습방법에 비해 비교적 높은 정확률을 보이나, 태그드코퍼스를 구축하는데 많은 비용과 시간이 소요된다는 단점이 있다. 이러한 단점을 보완하기 위해 비지도학습 방법론에서는 태그드코퍼스와 같은 데이터 없이 분류를 수행한다. 그러나 이 방법론은 정확도가 지도학습방법에 비해 약간 떨어질 수 있다는 단점이 있다.

다음 절에서는 현재까지 제안된 대표적인 지도학습방법 기반 자동분석 방안과 비지도학습방법 기반 자동분석방안을 소개하며, 독일어로의 적용가능성을 검토한다.

3.1 지도학습기반 논조자동분석

자연언어처리 분야에서 텍스트 논조 자동분석은 문장별 혹은 문서별로 긍정/부정/혼합 등의 태그 Tag를 부착한 태그드 코퍼스 getaggtes Korpus로부터 긍/부정/혼합 논조의 주요 자질 Merkmal을 기계학습 Maschinelles Lernen과정을 통해 추출한 후, ‘Support Vector Machine (이하 SVM)’이나 ‘Maximum

7) 영어권 문헌에서는 ‘Supervised Learning’이라 불림.
8) 영어권 문헌에서는 ‘Unsupervised Learning’이라 불림.

Entropy (이하 ME)', 'Naive Bayes (이하 NB)' 등과 같은 기계학습 알고리즘을 통해 테스트 문장이 긍정 논조인지 부정 논조인지 혹은 긍·부정 혼합논조인지를 예측하는 방식으로 이루어진다.

기계학습 과정은 일반적으로 훈련데이터 Training Data라고 불리는 태그드 코퍼스를 m차원의 벡터 Vector로 재구성하여, 데이터를 분류 Klassifizierung하기 위한 정보를 추출하는 과정이다. 이 때 가장 중요한 점은 기계학습을 위해 어떤 자질을 선택하여 다음과 같은 m차원의 벡터를 구성하느냐이다.

$d := (n_1(d), n_2(d), ..., n_m(d))$
$n_i(d)$: 문서 d에서 자질 f_i가 나타난 횟수
하나의 문서 내에서 나타날 수 있는 미리 정의된 m개 자질: $\{f_1, f_2, ..., f_m\}$

Pang et al.(2002)의 연구는 지금까지 수행된 지도학습기반 논조자동분석 분야의 가장 대표적인 연구로 꼽힌다. 이들의 연구에서는 'SVM' 'ME', 'NB'와 같은 3개의 지도학습기반 방법을 사용하여 영화평을 긍정/부정/중립으로 자동분류하는 시도를 다루고 있다.

이들의 실험에서 근본이 되는 가정은 긍정/부정 논조별로 해당 논조에 특징적인 어휘가 존재하고, 이 어휘들이 많이 등장할수록 해당 논조로의 경향성이 클 것이다라는 것이다. 실제로 2명의 전산학 전공 대학원생들에게 영화리뷰에서 긍정과 부정 논조를 전달하는 어휘를 선택하게 한 결과, 긍정적인 어휘로는 'dazzling, brilliant, phenomenal, excellent, fantastic, moving' 등의 어휘를 뽑았으며, 부정적인 어휘로는 'bad, boring, stupid, suck, slow, waste' 등의 어휘를 뽑 았다. 이들은 이러한 어휘만을 가지고 이 어휘들의 출현여부를 단순한 논조판단의 기준으로 삼아 자동분석 실험을 수행한 결과, 약 58~64%의 자동논조분석 정확성을 얻을 수 있었다. 그러나 사람이 선택한 어휘 이외에 자동으로 추출된 고빈도 어휘를 추가했을 경우 정확도는 69%까지 상승됨도 확인될 수 있었다. 자동으로 추출된 고빈도 어휘에는 예를 들어 긍정논조의 대표적인 어휘라고 보기에는 어려운 'still'같은 단어가 포함된다. 그럼에도 불구하고 이러한 어휘를 추가한 상태로 자동분석을 수행하면 오히려 성능이 올라감을 보였다.

이 실험에서 저자들은 유니그램 unigram과 바이그램 bigram, 품사정보를 주요한 자질로 삼아 기계학습을 수행하였다. 그 결과 유니그램만을 사용하였을 때가 가장 높은 정확률을 보였고, 그 중에서도 유니그램의 빈도가 아니라 유니그램의 출현여부를 이용한 결과가 가장 좋았다. 또한 3가지의 기계학습 방법 중 'SVM'을 사용하는 것이 가장 좋은 성능을 얻을 수 있음을 보였다. 이 실험은 그러나 문장레벨에서의 논조분석이 아니라 문서단위에서의 논조분석만을 수행한 것이므로, 단순히 유니그램만이 논조자동분석에 사용될 수 있다고 주장하는 것은 어려울 것이다.

3.2 비지도학습기반 논조자동분석

Pang et al.(2002)의 연구와 같은 지도학습기반 자동분석방법은 태그드코퍼스와 같은 대용량의 학습데이터를 필요로 한다는 문제가 있다. 이러한 문제를 해결하기 위한 방안으로 학습데이터를 필요로 하지 않는 방법론이 비지도학습기반 방법론이다. 논조자동분석 분야에서 비지도학습 방법론에 기반한 연구로는 Turney(2002)의 연구가 대표적이다.

이 연구에서는 품사태거 POS-Tagger와 'PMI-IR'알고리즘을 사용하여 학습코퍼스 없이 영화리뷰를 '추천/비추천'의 카테고리로 자동 분류한다. 먼저 영화리뷰를 영어 품사태거를 사용하여 형태소 분석한 후, 논조가 주로 구문적으로 구현되는 패턴, 예를 들어 "형용사+명사", "부사+형용사" "명사+형용사" 등과 같은 패턴을 추출한다. 추출된 패턴은 '추천'과 '비추천' 카테고리에 가장 대표적인 어휘인 'excellent'와 'poor'와의 밀접도 계산을 통하여 '긍정/부정' 논조를 가려내게 된다.

어떤 패턴이 'excellent' 혹은 'poor'와의 밀접도가 어느 정도인지 알아내기 위해 'PMI(Pointwise Mutual Information)'이라는 개념이 사용된다.

$$PMI(wort_1,\ wort_2) = \log2\{P(wort_1\ \&\ wort_2)\ /\ (P(wort_1)\cdot P(wort_2))\}$$

어떤 두 단어의 'PMI'값은 두 단어가 함께 출현하는 확률($P(wort_1\ \&\ wort_2)$)을 두 단어가 각각 출현할 확률의 곱($P(wort_1)\cdot P(wort_2)$)으로 나눈 것과

비례한다.

이 공식을 사용하여 추출된 패턴들과 ‘excellent’ 및 ‘poor’와의 ‘PMI’값을 각각 구한 후 그 차이를 계산한 것이 어떤 패턴의 의미적 경향 semantische Orientation이다. 추출된 패턴들과 ‘excellent’ 및 ‘poor’와의 ‘PMI’값을 구하기 위해 알타비스타 Altavista 검색엔진에 추출된 패턴에 대한 질의를 던져, 각각 ‘excellent’ 및 ‘poor’와 문장 내에서 인접하여 몇 번 사용되는지의 빈도를 구하였다.

이와 같은 방법으로 분야별 리뷰텍스트에 대해 약 74.39%의 정확률로 ‘긍정/부정’의 논조를 구별할 수 있었다. 이 연구에서 저자들은 단지 품사정보 및 ‘excellent’와 ‘poor’같은 제한된 언어지식만을 활용하였으나, 특정 어휘의 출현 여부 등과 같은 자질을 추가적으로 사용할 수 있을 것이라고 밝히고 있다.

4. 독일어 텍스트 논조결정 요인

본 연구에서는 독일어 텍스트의 논조결정 요인에 대한 연구를 위해 본 연구의 선행연구인 홍문표(2009)에서도 활용된 바 있는 51,314 단어 규모의 논조정보부착 코퍼스를 활용하였다. 이 코퍼스는 17,967단어 규모의 긍정논조코퍼스와 14,509 단어 규모의 부정논조코퍼스, 4,373 단어 규모의 혼합논조코퍼스, 그리고 14,465 단어 규모의 중립논조 코퍼스로 나뉘어진다. 각 문장은 다음의 예와 같이 논조에 따라 <positiv>, <negativ>, <gemischt>, <objektiv>의 태그가 부착되어 있다.

(1) <positiv> Wir als alte Apple bzw. iPhone Hasen, freuen uns natürlich dem Viewty unter den Rock sehen zu können. </positiv>
(2) <negativ> das handy ist eine volle katastrophe!! </negativ>
(3) <gemischt>Das X1 kommt in einer weniger schicken, dafür aber gut aufgeteilten und funktionellen Verpackung daher. </gemischt>
(4) <objektiv> Hallo, habe das Handy nun fast eine Woche. </objektiv>

논조태그는 (1)의 예와 같이 직접적인 감정이 표출되어 있는 경우('wir freuen uns') 뿐만 아니라 (2)와 같이 자신의 감정을 뒷받침하는 근거 혹은 원인('volle Katastrophe')이 드러나 있는 경우에도 부착되어 있다.

다음 절에서는 논조정보부착 코퍼스의 분석을 통해 향후 독일어 텍스트논조 자동분석 시스템을 기계학습기반으로 구현할 경우 특히 고려해야할 언어학적인 자질들에 대해 논의한다.

4.1 어휘 출현여부 및 출현 빈도

일반적으로 텍스트 마이닝 Text Mining과 같은 정보추출/검색 분야에서는 어떤 어휘가 출현하였느냐, 출현했다면 몇 번 출현하였느냐, 그리고 어떤 문서 내에서 얼마나 중요한가 등의 정보가 사용된다.[9] 이에 반해 논조자동분석에서는 Pang et al.(2002)의 연구에서 보인 것과 같이 어휘의 출현빈도 못지않게 어휘의 출현여부 자체도 큰 역할을 하는 것으로 알려져 있다.

홍문표(2009)의 연구에서는 코퍼스 분석을 통해 휴대폰 분야의 사용자 리뷰에서 자주 등장하는 논조별 어휘를 조사하였다. 18,300단어 규모의 긍정논조 코퍼스와 14,533단어 규모의 부정논조 코퍼스에서 관사와 조동사 등과 같은 기능어를 제거한 후 내용어들의 출현빈도를 조사한 결과, 논조별 최고빈도 어휘는 다음과 같았다.

<표 1> 홍문표(2009)에 나타난 휴대폰 분야 긍/부정 고빈도 어휘목록

	긍정논조		부정논조	
순위	어휘	빈도	어휘	빈도
1	gut	245	schlecht	45
2	super	78	leider	43
3	einfach	77	Problem	27
4	gross	49	langsam	15
5	besser	46	enttaeuschen	14
6	schnell	44	schwer	11

9) 이 값은 일반적으로 tf-idf 가중치를 통해 표현된다.

10) 'garnicht'는 'gar nicht'의 띄어쓰기 오류형태로서 하나의 어휘로 볼 수는 없지만,

7	top	37	negativ	11
8	klasse	35	klein	10
9	zufrieden	34	alt	10
10	toll	33	Nachteil	8
11	klein	30	leer	8
12	best	29	schade	8
13	leicht	25	nervig	7
14	echt	24	Kritikpunkt	6
15	schoen	23	stoeren	6
16	schick	22	kurz	6
17	empfehlen	20	schwarz	6
18	absolut	19	scheisse	5
19	voll	18	schwach	5
20	klar	17	garnicht[10)]	5

위의 어휘들을 살펴보면 ‘긍정/부정’ 논조별로 ‘gut’, ‘super’, ‘schnell’, ‘Top’, ‘schlecht’, ‘leider’, ‘Problem’ 등과 같은 문맥에 상관없이 특정 논조를 지니는 어휘가 많이 발견된다. Wilson et al.(2005)에서는 이와 같이 문맥에 상관없이 긍정 혹은 부정과 같은 특정 논조 혹은 감정을 전달하는 어휘를 ‘prior polarity word’라고 부른다. 긍정논조 독일어의 논조별 ‘prior polarity lexicon’을 구축하기 위해서는 논조별로 빈출하는 어휘에 대한 대규모 코퍼스 기반 분석이 선행되어야 한다.

일반적인 문서분류 Dokument Klassifikation에서는 위와 같은 어휘 목록 및 어휘의 출현빈도가 매우 중요한 역할을 한다. 그러나 논조분석의 경우에는 이러한 고빈도 어휘 뿐만 아니라 문서에 한 번이라도 등장하면 문서의 논조를 파악하는데 결정적인 역할을 할 수 있는 어휘들도 고려해야 한다. Yang et al.(2006)의 연구에서는 영어텍스트 논조자동분석을 위해 논조정보가 부착된 사전어휘 뿐만 아니라 ‘bugfested’와 같이 매우 드물게 출현하지만 강한 긍정/부정 논조를 전달하는 어휘도 텍스트 논조자동분석에 유용하게 사용될 수 있음을 보였다.

따라서 독일어 코퍼스에서 단 한번만 출현하는 어휘 ‘hapax legomena’도

10) 코퍼스 분석프로그램에서는 이를 구별할 수 없으므로 여기에서는 하나의 어휘로 취급함.

텍스트의 논조를 결정하는 매우 중요한 요소로 취급될 수 있을 것이다. 51,314 단어규모의 휴대폰 분야 논조정보부착 코퍼스에서 발견되는 'hapax legomena'는 각 논조별로 다음과 같다.[11]

긍정논조 'hapax legomena'

allerbest, ausgiebig, beeindruckt, beliebt, Besonderheit, brilliert, DANKE, erstklassig, fehlerfrei, Freude, funktionstüchtig, Funktionsumfang, Funktionsvielfalt, GEEIIIIILLLL, Genuss, glücklich, HAMMERGEIL, Kompliment, kostenfrei, lobenswert, makellos, nahtlos, perfektioniert, preiswert, Respekt, ruhig, Schnäppchen, Schnellzugriff, stinknormal, supergeil, supergut, SUUUUUPER, üppig, unglaublig, unübertrefflich, Volltreffer, vorbildlich, wackelfrei, wertvoll, wettbewerbsfähig, zeitlos,

부정논조 'hapax legomena'

abschrecken, bekloppt, Defizit, entnervt, erfolglos, erleiden, fehlerbeladen, Fehlkauf, Flop, fragwürdig, genervt, jämmerlich, leistungsschwach, Mist, nutzlos, Problemzone, Punktabzug, Quatsch, ramponiert, Scheiss, Scheissding, schmeissen, Trottel, ungeeignet, unsinnig, verschollen

위의 단어들을 모두 'hapax legomena'로 부르기는 무리가 있으나, 많은 단어들은 무수한 웹사이트를 대상으로 검색을 수행하는 검색엔진에서조차도 검색되는 회수가 매우 낮을 정도로 흔하지 않게 쓰이는 단어이다. 예를 들어, 'gut'이라는 단어를 'Yahoo Deutschland' 사이트에서 검색어로 질의할 경우, 총 5억8천5백만개의 결과가 검색된다. 이에 반해 'GEIIIIIL'이라는 단어를 검색어로 질의할 경우, 12,900개의 결과만이 검색된다.[12]

---

11) 이 논문에서 제시한 'hapax legomena'는 51,314 단어 규모의 소규모 코퍼스에서 추출한 단어들이므로, 보다 큰 코퍼스에서 추출할 경우 더 많은 빈도로 사용될 수 있다.

12) 'GEIIIIIL'은 'geil'의 의도적인 스펠오류형태이므로, 이것을 단어로 파악해야 하는가에 대한 논의가 있을 수 있으나, 논조자동분석을 위한 좋은 단서를 제공한다. 경우에 따라 스펠오류형태가 중간 모음 'I'의 개수에 따라 다양하게 출현할 수 있을 것이나, 이는 'GE*I*L'과 같은 정규식표현 Regulärer Ausdruck을 통해 처리할

이와 같이 'GEIIIIIL'과 같은 단어 혹은 문자열은 매우 드물게 사용되는 단어이지만 아래의 예문 (5)에서 보는 바와 같이 매우 강한 저자의 긍정적인 논조를 전달하므로 자동분석에서는 반드시 고려되어야 할 단어이다.

(5) Habe das Handy jetzt seit ca. 6 Wochen. Ich finde den Lautsprecher relativ leise. Wenn es um die Einstellung, bspw von SMS-Tönen, geht, ist das u900 sehr kompliziert. Die Vorteile überwiegen jedoch deutlich. Tolle Optik, größtenteils übersichtliche Menüführung, für Schnappschüsse durchaus absolut ausreichende Kamera, Touchscreen ist **GEEIIIIILLLL**...

4.2 품사

텍스트의 논조를 결정하는 단어들을 품사별로 분류해보면 형용사가 가장 큰 역할을 함을 알 수 있다. Hatzivassiloglou/Mackeown (1997)는 텍스트 논조의 결정에 영향을 미치는 형용사의 의미적 경향을 분석하는 연구를 통해 텍스트 논조자동분석에 대한 초창기 연구를 주도하였다. Hatzivassiloglou/Wiebe (2000)에서도 형용사를 중심으로 텍스트 논조를 파악하는 방법론을 제안하였다. Mullen/Collier (2004)의 연구에서는 'SVM' 기반 기계학습 과정에서 형용사에 대한 가중치를 높임으로써 논조자동분석의 정확성을 높일 수 있음을 보였다.

독일어의 경우도 영어와 마찬가지로 논조결정에 가장 큰 영향을 미치는 어휘의 품사는 형용사이다. 본 연구에서는 논조정보부착 코퍼스의 분석을 통해 긍정논조와 부정논조에서 모두 논조정보를 전달하는 어휘 중 형용사가 가장 큰 역할을 함을 알 수 있었다.

긍정논조의 경우 코퍼스에서 빈도 2회 이상 출현하는 어휘 179개 중 총 127개, 즉 전체의 약 71% 정도가 형용사였다. 그 외 16.7%는 명사, 12.3%는 동사의 순서였다. 부정논조의 경우 긍정논조보다는 형용사의 비중이 약간 낮기는 하지만 그럼에도 불구하고 빈도 2회 이상 출현하는 어휘 전체 100개의

수 있을 것이다.

어휘 중 53%가 형용사였다. 그 외에는 29%가 명사, 18%가 명사였다.

<표 2> 논조별 품사분포

긍정논조 (179개)			부정논조(100개)		
형용사	명사	동사	형용사	명사	동사
127(70.94%)	30(16.76%)	22(12.30%)	53(53%)	29(29%)	18(18%)

위와 같은 품사분포 정보는 문장의미의 주관성/객관성을 판별하는데 우선적으로 사용될 수 있을 것이다. 텍스트논조 자동분석을 위해 문장이 감정이나 의견 등을 담은 주관적인 의미의 문장인지 아니면 사실과 같은 객관적인 의미의 문장인지를 먼저 구별해야 한다. 이 과정에서 형용사나 부사 등과 같은 품사 정보의 사용은 문장의미의 주관성/객관성을 판별하는데 유용한 데이터로 사용될 수 있다.[13] 이는 다른 문장에 비해 형용사나 부사의 사용빈도가 상대적으로 높은 경우 주관적 의미를 전달하는 문장일 가능성이 높기 때문이다.

4.3 구문구조정보

의존구조 Dependenzstruktur등과 같은 구문구조정보를 논조분석에 적용하는 것은 논란의 여지가 있다. Kudo/Matsumoto(2004)는 n-그램 정보와 함께 의존구조정보를 사용하여 n-그램만을 사용할 때 보다는 더 높은 성능을 올릴 수 있음을 보였다. Wilson et al.(2005)의 연구에서도 어휘가 갖는 논조가 문맥에 따라 변화되는 성질을 파악하기 위해 어휘가 문장의 의존구조에서 갖는 문법적 기능에 주목하였다.

그러나 이에 반해 Dave et al.(2003), Gamon et al.(2005) 등의 연구에서는 의존구조정보가 논조분석의 성능 향상에 큰 효과가 없다는 실험결과를 보였다. 그럼에도 불구하고 일반적으로 의존구조파악과 같은 심층언어구조분석은 부정어와 강조어 등의 수식영역Skopus 등을 정확하게 계산하는데 큰 도움이 될 수 있다.

13) Wilson et al.(2005)의 연구에서는 주관적 의미의 문장과 객관적 의미의 문장을 'prior polarity lexicon'을 사용하여 구별하는 방법론을 제안하였다.

(6) das Bier, das ich in dem warmen Zimmer gefunden habe, war lecker

(6)과 같은 긍정적 논조의 예문을 통사구조에 대한 고려 없이 어휘출현 등만을 학습자질로 삼아 기계학습을 수행할 경우, 'Bier'라는 어휘가 'warm'이라는 어휘와 함께 사용될 경우 '긍정'의 논조를 전달할 수 있다라고 잘못 학습될 가능성이 있다. 따라서 이러한 노이즈를 막기 위해 학습과정에서도 문장의 구조를 함께 고려해야 한다.[14)]

통사분석을 통한 구문구조정보의 사용에 대한 대안으로 바이그램 bigram 혹은 트라이그램 trigram과 같은 고차원 n-그램의 사용을 생각해볼 수 있다. 한 어휘가 문장 내에서 어떠한 역할을 했는지를 알기 위해 문장구조를 파악하는 방법 이외에 바이그램 내지는 트라이그램을 사용할 수 있기 때문이다. 예를 들어 홍문표(2009)의 연구에서 언급한 바와 같이 긍정논조에서 'einfach'라는 어휘가 많이 사용되었을 경우, 이 단어가 'sein'동사와 함께 사용되어 술어의 용법으로 쓰였는지, 아니면 다른 형용사나 부사 앞에서 수식어의 역할을 하는지는 구조분석을 통해서도 파악할 수 있지만, 바이그램 정보만으로도 어느 정도는 알아낼 수 있기 때문이다.

4.4 부정어

'nicht'나 'kein'과 같은 부정어 Negation는 문장논조의 극성 Polarität을 바꾸는 단어이므로 그 처리가 매우 중요하다. 순수 어휘출현 기반 자동분류방식("bag of words" 방식)에서는 논조가 전혀 다른 아래의 두 문장이 매우 유사도가 높은 것으로 판단해버린다.

(7) Ich mag das Handy
(8) Ich mag das Handy nicht

14) 문장구조정보를 기계학습과정에서 반영하는 방법은 여러 가지로 고려될 수 있다. 예를 들어 'prior polarity'를 가진 어떤 형용사가 다른 명사와 한 문장 내에서 사용될 때에는 항상 수식어-피수식어 관계에 있는 경우에만 극성정보를 해당 명사에 반영하므로, 해당 '수식어-피수식어' 문장구조정보를 함께 학습하는 것이 더 정확한 분류를 위해 도움이 될 것이다.

따라서 기존의 논조자동분석 연구에서도 영어의 경우 문장 내에 ‘not’이나 ‘don't’ 등이 출현하면, 이러한 부정어 다음에 나오는 어휘들의 극성을 모두 반대값으로 바꾸어 기계학습을 수행하였다. 예를 들어 “I don’t like the plot” 과 같은 문장의 경우 부정어 ‘don't’가 등장하므로 ‘like’는 ‘like_not’으로 변환되어 기계학습에 적용된다.

이러한 처리방법은 구문분석 결과에 기반하는 것이 아니므로, 부정어의 수식영역을 파악할 수 없다. 따라서 대개의 경우 문장내에 부정어가 등장하면 문장내의 모든 단어의 극성을 ‘_not’을 부착하여 바꾸게 된다. 앞서 4.3에서도 언급하였듯이 구문구조정보를 통해 부정어의 수식영역을 파악하여, 수식영역에 있는 어휘들의 극성만을 바꾸어주는 과정이 반드시 필요하다고 할 수 있다.

독일어에서 기계학습을 위해 부정어로 처리해야할 어휘는 ‘nicht’와 ‘kein(e)’를 우선적으로 들 수 있다. 이 경우 구조분석을 통해 ‘nicht’와 ‘kein(e)’의 수식영역을 파악한 후, 수식영역에 나타나는 어휘의 극성만을 반대값으로 변환해야 한다. (9)에서는 ‘nicht’의 수식영역이 ‘schlecht’까지이므로, ’prior polarity’값이 ‘부정‘인 ‘schlecht’가 ‘긍정’으로 바뀌게 된다.[15] (10)의 경우에는 ‘kein(e)’의 수식영역이 ‘Blöße’까지이므로, ’prior polarity’값이 ‘부정’인 ‘Blöße’가 ‘긍정’으로 바뀌게 된다.

(9) Dabei ist der beiliegende gar nicht mal [schlecht] und wird dem Normalanwender weit genügen.

(10) Auch beim Thema Attachment gibt sich Apple keine [Blöße] und bietet volle Office-Kompatibiliät.

이 외에도 ‘mitnichten’, ‘kaum’, ’ohne’, ’wenig(er)’ 등이 코퍼스에서 자주 등장하는 극성 값을 변화시키는 부정어들이다. (11)~(13) 참조.

(11) Mitnichten [schwierig] gestaltet sich dagegen die Interaktion via USB-Kabel

15) ‘nicht nur ~ sondern auch’와 같은 강조를 위한 표현은 부정어 처리대상에서 제외한다.

(12) Das Neue war aber kaum [besser]
(13) Ich gehe normalerweise nicht sehr sorgfältig mit meinen Mobiltelephonen um, aber bislang hat es alle Stürze ohne [Probleme] überstanden
(14) Weniger [gut] funktionierten die erweiterten Gesprächsoptionen, die Apple anbietet

4.5 문장위치

많은 경우에 텍스트의 논조를 단순히 텍스트 안에 들어있는 긍/부정 어휘의 개수만을 고려하여 판단한다면 아래 예문 (15)는 긍정적이기 보다는 부정적으로 분석될 가능성이 높다. 왜냐하면 아래 예문에서 긍정적으로 볼 수 있는 어휘의 개수는 모두 4개이고, 부정적으로 볼 수 있는 어휘의 개수는 모두 5개이기 때문이다.[16]

(15) Sicherlich ist das Omnia i900 ein *gutes* Handy mit verbundenem *Mehrwert*, aber Datenaustausch per Activesync, ein im Hellen schlecht ablesbares Display und das Starten von Programmen (leicht verzögert) müßten noch verbessert werden, dann wäre es beinahe ein *perfektes* Handy. Ebenso reicht es nicht aus, Programme durch antippen von "x" in der rechten oberen Ecke zu beenden. Hier müssen sämtliche im Hintergrund laufenden Programme erst umständlich durch den Taskmanager abgebrochen werden. **Dennoch habe ich den Kauf dieses Handys *nicht bereut.***

그러나 일반적으로 문장의 주제는 문두 혹은 문미에 강조된다는 점을 기계학습과정에서 고려한다면 좀 더 정확한 논조분석 결과를 얻을 수 있을 것이다. (15)에서도 비록 긍정적 논조 어휘의 사용이 부정적 논조의 어휘보다 많지는 않지만, 저자의 강한 의견이자 텍스트의 주제가 문미에 나타나고 있다. 사용자 후기 등과 같은 텍스트 유형에서는 일반적으로 문두 혹은 문미에 글

16) 위 예문에서 긍정적으로 평가할 수 있는 어휘는 이탤릭체로 표기되어 있고, 부정적으로 평가할 수 있는 어휘는 밑줄로 표시되어 있다.

작성자가 자신의 견해를 정리하여 나타내므로, 텍스트 내에서 논조가 들어간 문장의 위치도 기계학습시 중요한 요소로서 고려되어야 한다. 그러나 좀 더 정확한 결과를 위해서는 사용자 후기나 리뷰 등의 텍스트에서 저자의 주제가 텍스트의 어느 위치에 주로 나타나는가에 대한 경험적인 연구가 선행되어야 한다.

5. 향후연구과제

본 논문에서는 독일어 텍스트의 논조자동분석을 위해 고려해야할 것들에 대해 전산언어학의 관점에서 논의하였다. 어떤 텍스트 혹은 문장이 주제에 대해 긍정적이냐 부정적이냐 혹은 중립적이냐 등을 구별하는 것은 결국 데이터를 기준에 따라 분류하는 작업과 크게 다르지 않다. 이러한 이유로 대부분의 고성능 논조자동분석은 기계학습 방법론을 따르고 있다.

기계학습이란 결국 학습데이터로부터 분류를 위한 자질을 추출하고, 분류대상이 되는 데이터를 학습된 자질과 비교하여 어느 쪽으로 분류할 수 있을지를 결정하는 과정이라고 할 수 있다. 이를 위해 본 연구에서는 우선 Fries(2006) 등의 연구를 통해 감정 혹은 논조가 언어적으로 실현되는 양상에 대해 검토하였다. 이러한 언어적 실현 양상은 ‘어휘출현빈도 및 출현여부’, ‘품사’, ‘구문구조정보’, ‘부정어’, ‘문장위치’ 등과 같은 기계학습을 위한 자질로 요약될 수 있었다.

향후 연구에서는 본 연구를 통해 제안된 언어학적 특성들을 기계학습 자질로 고려하여, 최근 가장 높은 성능을 보이는 것으로 알려진 ‘SVM’ 알고리즘에 기반하여 실제 독일어 텍스트의 논조를 자동으로 분석하는 실험을 수행하고자 한다. 이 실험에서는 본 연구에서 제안한 언어학적 자질 외에도 실제 텍스트 논조분석시 큰 문제로 대두되는 멀티토픽 문제(예를 들어 아이폰과 햅틱폰을 비교하는 문서), 형용사 논조의 문맥 의존성 문제(예를 들어 ‘warm’이 ‘warmes Brot’에서는 긍정적이지만, ‘warmes Bier’의 경우는 부정적인 점) 등도 함께 고려되어야 할 것이다.

참고문헌

이익환/이민행 (2005):『심리동사의 의미론 - 영어, 한국어와 독일어의 대조연구』. 도서출판 역락.

홍문표 (2009): 독일어문장 논조자동분석을 위한 어휘분포양상 연구.『독일문학』.

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

Aman, S./Szpakowicz, S. (2008): Using Roget's Thesaurus for Fine-grained Emotion Recognition. *Proceedings of IJCNLP*, 312-318.

Butscher, R. (2005): *Text Mining in der Konsumentenforschung unter besonderer Berücksichtigung von Produktontologien,* Ph.D. Dissertation, Universität Erlangen-Nürnberg.

Dave, K. Lawrence, S./Pennock, D. M. (2003): Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web.*

Ding, X./Liu, B. (2007). The Utility of Linguistic Rules in Opinion Mining. *SIGIR* 2007, 812-812.

Ekman, P. (1992): An Argument for Basic Emotions: *Cognition and Emotion* 6, 169-200.

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.

Fang, W. Hsu, Y./Sung, K. (2008): Electronic word-of-mouth and purchase intentions: The mediating role of conformity tendency. *International Journal of Psychology* 43(3), 137.

Feldweg, H. (1997): GermaNet - ein lexikalisch-semantisches Netz für das Deutsche. In: Ludewig, P/Geurts, B. (eds.): *Lexikalische Semantik aus kognitiver Sicht - Perspektiven im Spannungsfeld linguistischer und psychologischer Modellierungen*. Tübinger Beiträge zur Linguistik (TBL) Bd. 439, Tübingen: Gunter Narr Verlag.

Fellbaum, C./Gross, D./Miller, K. (1993): *Adjectives in WordNet.* http://www.cosgi.princeton.edu/~wn.

Fries, N. (2007): Die Kodierung von Emotionen in Texten. *Journal of Literary*

Theory.

Gamon, M. (2004): Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: *Proceedings of the 20th International Conference on Computational Linguistics* (COLING-2004), 841–847.

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

Helbig, H. (2001): *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet.* Berlin et al.: Springer.

Jahr, S. (2000): *Emotionen und Emotionsstrukturen in Sachtexten: ein interdisziplinärer Ansatz zur qualitativen und quantitativen Beschreibung der Emotionalität der Texten.* Walter de Gruyter.

Kamps, J./Marx, M./Mokken, R.J./de Rijke, R. (2002): Words with attitude. In: *Proceedings of the 1st International Conference on Global Word-Net*, 332-341.

Kanayama, H./Nasukawa, T./Watanabe, H. (2004): Deeper sentiment analysis using machine translation technology. *Proceedings of the 20th International Conference on Computational Linguistics.*

Kunze, C. (2000): Extension and Use of GermaNet, a Lexical-Semantic Database. In: *Proceedings of the Second International Conference on Language Resources and Evaluation* Vol. II, 999-1002.

Kunze, C. (2001): Lexikalisch-semantische Wortnetze. In: Carstensen, K.-U. et al. (Eds.): *Computerlinguistik und Sprachtechnologie: eine Einführung.* 386-393.

Liu, B./Hu, M./Cheng, J. (2005): Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th international World Wide Web conference*, 342-451.

Martin, J. R./White, P. R. (2005): *The Language of Evaluation: Appraisal in English.* Palgrave Macmillan.

Miller, G./Charles, W. (1991): Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1-28.

Morinaga, S./Yamanishi, K./Tateishi, K./Fukushima, T. (2002): Mining Product Reputations on the WEB. *Proceedings of 8th ACM SIGKDD International Conference on Knowledge. Discover and Data Mining,* 341-349.

Osgood, C.E./Suci, G.J./Tannenbaum, P.H. (1957): *The Measurement of Meaning.* University of Illinois Press, Chicago.

Osswald, R. (2004): Die Verwendung von GermaNet zur Pflege und Erweiterung des Computerlexikons. *HaGenLex, LDV Forum* Band 19, 43-51.

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.

Pang, B./Lee, L. /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, 2002.

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

Scaffidi, C./Bierhoff, K./Chang, E./Felker, M./Ng, H./Jin, C. (2007): Red Opal:product-feature scoring from reviews. *Proceedings of the 8th ACM Conference on Electronic Commerce*, 182-191.

Siegel, M./Xu, F./Neumann, G. (2001): Customizing Germanet for the use in deep linguistic processing. In: *Proceedings of the NAACL Workshop*, 1-7.

Strapparava, C./Valitutti, A. (2004): WordNet-Affect: an affective extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004), 1083-1086.

Sugitani, Y. (2008): The validity of 'word-of-mouth' on the web: the nonverbal cues can interfere with consumers' memory. *International Journal of Psychology.* 43(3), 656.

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), Philadelphia, Pennsylvania, 417-424. (NRC #44946).

Valitutti, A./Strapparava, C./Stock, O. (2004): Developing Affective Lexical Resources. In: *PsychNology Journal.* 2(1).

Wang, B./Wang, H. (2008): Bootstrapping Both Product Features and Opinion Words from Chinese Customer Reviews with Cross-Inducing. *Proceedings of IJCNLP* 2008, 289-295.

Wilson, T./Wiebe, J./Hoffmann, P. (2005): Recognizing Contextual Polarity in

Phrase-Level Sentiment Analysis, *HLT/EMNLP*, 347-354.
Yang, K./Yu, N./Valerio, A./Zhang, H. (2006): WIDIT in TREC-2006 blog track. *Proceedings of TREC*.
Yi, J./Niblack, W. (2005): Sentiment Mining in WebFountain: *Proceedings of the 21st International Conference on Data Engineering*, 1073-1083.

Zusammenfassung

Computerlinguistische Studie zu den Entscheidungsfaktoren der Textsentiments

Hong, Munpyo (Sungkyunkwan Univ.)

Die vorliegende Arbeit beschäftigt sich mit den linguistischen Faktoren, die in der automatischen Analyse des Textsentiments berücksichtigt werden sollen. Texte oder Sätze nach dem Sentiment zu unterscheiden ist nicht viel anders als Daten nach gegebenen Kriterien zu klassifizieren. Aus diesem Grunde stützen sich die meisten Ansätze für automatische Sentimentanalyse auf maschinelles Lernen.

Unter dem Begriff „maschinelles Lernen„ versteht man einen Prozess, in dem Merkmale für die Klassifikation aus dem Lernkorpus extrahiert werden und die neuen Daten im Hinblick auf die Merkmale neu klassifiziert werden. Um die „Merkmale„ für die automatische Klassifikation zu identifizieren, wurde die Arbeit von Fries(2006) herangezogen. Zu den wichtigsten Merkmale für das maschinelle Lernen gehören u.a. „Wortfrequenz und Wortpräsenz„, „POS„, „syntaktische Struktur„, „Negation„ und „Position im Text„

Auf der Basis der Ergebnisse der vorliegenden Arbeit soll noch ein Experiment anhand des „SVM„-Algorithmus durchgeführt werden. In dem Experiment müssen noch die Themen wie „Multi-Topik„ und „Kontextabhängigkeit des Sentiments„ mit berücksichtigt werden.

[검색어] 텍스트논조, 감성분석, 의견마이닝, 기계학습
Textsentiment, Sentimentanalyse, Opinion Mining, Maschinelles Lernen

홍문표 110-745
서울시 종로구 명륜동 3가 53번지
성균관대학교 문과대학 독어독문학과
skkhmp@skku.edu

논문 투고일: 2009. 10. 22
논문 심사일: 2009. 11. 20
게재 확정일: 2009. 11. 29