독일어 감정사전을 활용한 감성분석

홍문표 (성균관대)

1. 서론

감성분석 혹은 논조분석 Sentimentanalyse은 특정 주제에 대한 텍스트 의미의 긍정성, 부정성을 자동으로 분석하는 자연언어처리 연구분야의 하나이다. 텍스트 의미의 긍정성, 부정성이란 텍스트 주제에 대한 저자의 호감, 비호감을 뜻하며 감성분석은 텍스트를 구성하는 개별 문장의 긍정성, 부정성 분석을 통해 저자의 생각을 간접적으로 알아내는 것을 목표로 한다.

빅데이터 분석의 한 분야로도 볼 수 있는 감성분석 연구의 초기에는 기존의 텍스트 마이닝 분야와 유사하게 학습코퍼스 Lernkorpus로부터 분석을 위한 자질 Merkmal을 추출하고, 추출된 자질을 활용해 새로운 데이터의 긍정성/부정성 또는 논조를 분석하는 학습코퍼스 기반의 분석 방법론이 시도되었다. 학습코퍼스 기반의 감성분석 방법론의 가장 큰 문제점은 감성분석시스템이 학습된 코퍼스의 도메인에 의존적이라는 점이다. 예를 들어, 스마트폰 분야의 학습코퍼스로 개발된 감성분석시스템을 자동차 분야의 텍스트에 적용했을 경우 성능저하가 예상된다. 또 하나의 큰 현실적인 문제는 도메인별로 이러한 학습코퍼스를 구축하는데 많은 비용과 시간이 든다는 점이다.

이러한 문제점을 해결하기 위한 하나의 방안으로 생각할 수 있는 것은 감정사전을 활용하여 감성분석기를 개발하는 것이다. 감정사전을 활용하여 감성분석기를 개발하는 방법에는 크게 두 가지 방법이 있다. 첫째는 감정사전 엔트리 및 감성지수를 의미분석 규칙에 직접 활용하여 분석하는 방법이고, 둘째는 감정사전 엔트리를 기계학습을 위한 학습자질로 활용하는 것이다.

본 연구에서는 두 번째의 방법, 즉 감정사전의 엔트리를 기계학습을 위한 자질로 사용하는 방법을 취하려고 한다. 그 이유는 Taboada et al. (2011)의 연구와 같이 감정사전과 감성지수를 직접 감성분석 규칙에 적용할 경우 감성지수의 타당성에 따라 분석결과가 좌우될 수 있다는 문제가 있기 때문이다. 이

에 대해서는 본문에서 좀 더 자세하게 논의하도록 한다.

본 연구의 또 다른 목적은 도메인별 학습자질의 타도메인 적용가능성에 대한 검증이다. 예를 들어 스마트폰 분야의 감성분석을 위해 선택된 학습자질이 호텔예약 분야에 적용될 경우의 성능 등을 알아봄으로써 감정사전을 활용한 방법론의 타당성을 검증해볼 수 있다. 학습코퍼스에서 추출한 도메인별 학습자질이 타도메인에 적용될 경우 분석성능이 하락할 것이라는 것은 일반적으로 예측이 가능하지만 실제로 실험을 통해 보고된 사례는 아직 없다.

본 논문의 구성은 다음과 같다. 2장에서는 독일어 감정사전의 구조와 특징에 대해 알아본다. 본 연구에서 다루는 독일어 감정사전은 Remus et al. (2010)의 Senti-Wortschatz (이하 'SentiWS')와 Waltinger (2010)의 German Polarity Clues (이하 'GPC')이다. 3장에서는 감정사전을 기계학습에 적용하는 감성분석 방법론에 대해 소개한다. 4장에서는 본 연구에서 제안하는 방법론의 검증을 위한 실험을 다룬다. 끝으로 5장에서는 본 연구의 성과를 정리하고 향후 연구 방향을 소개한다.

2. 독일어 감정사전

2.1 Senti-Wortschatz

감성분석에 활용되는 두 개의 대표적인 감정사전 중 하나는 Remus et al. (2010)의 'SentiWS'이다. 라이프치히 대학에서 개발된 이 사전은 1.8버젼을 기준으로 총 1,650개의 긍정단어, 1,818개의 부정단어엔트리를 수록하고 있다. 이는 단어의 원형을 기준으로 한 규모이며, 단어의 모든 굴절형태를 포함할 경우 15,649개의 긍정단어 엔트리와 15,632개의 부정단어 엔트리를 가지고 있다.[1)]

1) 'SentiWS'는 구축이 완료된 사전이 아니므로 사전의 규모는 계속적으로 변하고 있는 것으로 보인다. 실제로 'SentiWS'의 홈페이지와 Remus et al. (2010)에서는 사전의 규모를 긍정 15,649엔트리, 부정 15,632의 엔트리로 제시하고 있으나, 본 연구의 실험을 위해 다운로드하여 사전을 분석한 결과 총 31,270개의 엔트리로서 논문

Word	POS Tags	Weight	Inflections
harmonisch	ADJX	+0.5243	harmonische, harmonischer, ..., harmonischst
Krise	NN	-0.3631	Krisen

<표 1> 'SentiWS'의 내부구조

표 1[2]에서 보는 바와 같이 'SentiWS'는 키워드 (Word), 품사태그 (POS Tags), 가중치 (Weight), 굴절형태 (Inflections) 등의 정보가 수록되어 있다. 품사태그는 슈투트가르트-튀빙엔 태그세트 (Stuttgart-Tübingen Tagset)를 사용하고 있다. 굴절형태는 Remus et al. (2010)에 따르면 에러가 존재할 수도 있으며 모든 가능한 굴절형태를 포함하고 있지는 않다고 한다.

'SentiWS'의 키워드는 세 개의 서로 다른 출처로부터 컴파일되었다. 첫 번째 출처는 영어의 'General Inquirer'[3]사전으로서, 여기에는 키워드에 '긍정'과 '부정'의 태그가 부착되어 있다. 'SentiWS'의 컴파일을 위해 Remus et al. (2010)에서는 'General Inquirer'의 키워드들을 구글 Google 영-독 자동번역시스템을 활용하여 번역한 후 수정하는 방식을 취하였다. 'General Inquirer' 사전은 1,915개의 긍정 단어와 2,291개의 부정 단어를 엔트리로 가지고 있다.

여기서 한 가지 주목할 점은 'SentiWS'의 엔트리 선정을 위해 'General Inquirer'의 엔트리를 자동번역하여 사용하였다는 점이다. 이 경우 발생할 수 있는 문제점은 영어사전이 단어단위로 엔트리가 구성되어 있으므로 다의어인 경우 여러 개의 독일어 단어로 번역될 수 있다는 점이다. 이 경우에는 영어단어의 '긍정', '부정' 정보가 독일어의 모든 번역된 단어에 잘못 전달될 가능성이 존재한다.

'SentiWS'의 표제어 선정을 위한 두 번째 출처는 '긍정'과 '부정'의 태그가 부착된 상품평 코퍼스이다. Remus et al. (2010)은 각각 5,100개의 '긍정'과 '부정' 태그가 부착된 상품평으로부터 긍정어휘와 부정어휘를 추출하였다. 각

에서 제시하는 사전의 규모와는 약간의 차이가 있음을 밝힘.

2) 표1은 Remus et al. (2010)에서 가져옴.

3) http://www.wjh.harvard.edu/~inquirer/

5,100개의 긍정, 부정 코퍼스는 문장단위로 환산하면 긍정 30,074 문장, 부정 36,743 문장이다. 이들은 각 문장에 등장하는 모든 단어들에 긍정문장의 경우에는 'POS_', 부정문장의 경우에는 'NEG_'와 같은 가상접두사를 부착하였다.[4)]

가상접두사를 부착한 후 '로그우도비 log-likelihood' 측정법을 적용하여 통계적으로 유의미한 범위에서 'POS_', 'NEG_' 접두사와 자주 등장한 긍정의미의 어휘와 부정의미의 어휘를 찾아내었다. 이와 같은 방식으로 코퍼스에서 사전 엔트리로 수록할 단어를 찾아내는 것은 도메인 의존적인 어휘들의 추출에 도움이 되었으며, 'Reklamation', 'Fehlkauf' 등이 이에 속하는 어휘들이다.

'SentiWS'의 표제어 선정을 위한 세 번째 출처는 Quasthoff (2010)의 독일어 연어사전이다. 독일어 연어사전은 의미유사도에 따라 단어들이 그룹으로 나뉘어 있는데, 이 연구에서는 총 25,288개의 의미그룹 중에서 감성정보와 약간의 관련이 있는 6,932개의 의미그룹과 감성정보와 매우 큰 관련이 있는 76개의 의미그룹을 사용하였다. 이 방법을 통해서는 코퍼스 등에서 사용빈도가 매우 낮은 'sonnendurchflutet', 'glasklar', 'bärenstark' 등의 어휘가 'SentiWS'에 추가되었다.

위의 표 1에서 이미 살펴본 바와 같이 'SentiWS'는 긍정, 부정 의미와 관련된 가중치를 가지고 있다. 이 가중치는 다음의 수식을 통해 계산된다.

(1) $$\text{SO-A}(w) = \sum_{p \in P} A(w,p) - \sum_{n \in N} A(w,n)$$

위 수식이 의미하는 것은 단어 w와 긍정적인 단어의 집합 (P)에 속하는 원소 (p)와의 의미연관도에서 단어 w와 부정적인 단어의 집합 (N)에 속하는 원소 (n)와의 의미연관도를 뺀 값이 단어 w의 의미경향성 Semantic Orientation이라는 것이다. 이 때 의미경향성 값이 0보다 크면 단어 w는 긍정의미의 단어이고 0보다 작으면 부정의미의 단어로 분류된다.

이 실험에서 사용한 긍정어휘의 집합 (P)와 부정어휘의 집합 (N)은 (2)와 같다.

---

4) Remus et al. (2010)에서는 'POS_', 'NEG_'이라는 직접적인 가상단어의 형태를 제시하지 않고 '가상단어 (pseudo word)'를 사용했다고만 언급하고 있으나, 본 논문에서는 독자들의 이해를 돕기 위해 'POS_', 'NEG_'의 가상단어를 제시하였음.

(2)

긍정어휘 집합 (P): gut, richtig, schön, glücklich, erstklassig, positiv, großartig, ausgezeichnet, lieb, exzellent, phantastisch

부정어휘 집합 (N): schlecht, unschön, falsch, unglücklich, zweitklassig, negativ, scheiße, minderwertig, böse, armselig, mies

두 단어 간의 의미연관도는 'Pointwise Mutual Information (PMI)'라고 하는 방법으로 계산된다. 이 방법은 Turney (2002)에 의해 제안되었는데 어떤 단어가 다른 단어와의 의미적 연관성이 어느 정도인지를 통계적인 방법에 기반하여 계산하는 방법이다.

(3) $$\mathrm{PMI}(w_1, w_2) = \log_2 \left(\frac{P(w_1 \,\&\, w_2)}{P(w_1) \cdot P(w_2)} \right)$$

수식 (3)이 의미하는 것은 두 단어 w_1과 w_2 간의 의미연관도는 w_1의 출현확률과 w_2의 출현확률의 곱을 분모로 하고 w_1과 w_2가 함께 출현하는 확률을 분자로 하여 로그함수를 적용한다는 것이다.

이 연구에서는 1억 문장 규모의 독일어 코퍼스에 위 방법을 적용하여 단어의 긍정, 부정 가중치를 얻어내었다. 그러나 이러한 방법은 크게 두 가지의 문제를 가지고 있다. 첫째, 단어의 긍정성, 부정성 모두 각각 대표 긍정어휘 11개, 대표 부정어휘 11개와의 의미관련도를 기반으로 계산되기 때문에 대표어휘들의 선정에 따라 가중치가 달라질 수 있다. 예를 들어 특정 어휘가 매우 긍정성이 높은 단어이지만 우연히도 11개의 대표어휘와는 자주 공기하지 않는 어휘라면 가중치가 매우 낮을 수도 있다. 둘째, 빈도계산을 위한 코퍼스의 선정에 따라서도 가중치가 달라질 수 있다. 위 실험에서는 1억 문장 규모의 독일어 코퍼스를 대상으로 하였지만, 코퍼스의 도메인 및 텍스트 레지스터 등에 따라 기대와는 다른 결과가 나올 수 있기 때문이다.

실제로 'SentiWS' 데이터를 좀 더 자세히 들여다보면 가중치의 측면에서 이해하기 어려운 경우를 쉽게 찾아볼 수 있다.

(4)
bärenstark|ADJX 0.0040
beachtenswert|ADJX 0.0040
brandneu|ADJX 0.0040
charmant|ADJX 0.3223
cool|ADJX 0.0040
dauerhaft|ADJX 0.205

(4)는 ‘SentiWS’ 데이터 중 형용사 엔트리 목록의 일부이다. 데이터 포맷은 ‘표제어|POS 가중치’의 형식이다. 논의를 위해 가중치를 살펴보면 최하 0.0040부터 최고 0.3223까지의 값이 분포되어 있다. 여기서 우리의 직관에 따르면 매우 긍정적인 의미를 전달할 것으로 생각되는 ‘bärenstark’, ‘beachtenswert’, ‘brandneu’ 및 ‘cool’은 모두 0.004의 값을 가지는 반면, 문맥에 따라 긍정 또는 부정의 모든 극성을 지닐 수 있는 ‘dauerhaft’는 0.205의 값을 가지게 되어 이들 단어보다 무려 50배가 넘는 가중치를 갖는다. 이는 코퍼스의 규모 및 특성과 대표어휘 목록 때문에 발생한 통계적인 차이일 뿐, 우리의 언어직관과 일치한다고 보기는 어려운 결과이다. 따라서 본 연구에서는 ‘SentiWS’의 가중치를 신뢰하지 않고, ‘긍정’과 ‘부정’이라는 정보만을 사용할 것이다.

이어서 2.2장에서는 ‘SentiWS’와 더불어 독일어 감성분석을 위한 대표적인 리소스인 ‘German Polarity Clues’를 살펴보기로 한다.

2.2 German Polarity Clues (GPC)

Waltinger (2010)에 의해 개발된 ‘GPC’도 ‘SentiWS’와 마찬가지로 영어 감정사전을 기반으로 하고 있다. Waltinger (2010)은 Wiebe et al. (2005)의 ‘Subjectivity Clues’ 사전을 채택하였으며, 자동번역기를 통해 이 사전들의 영어엔트리를 독일어로 번역하였다. 이 때 발생하는 문제점은 역시 하나의 영어단어가 다수개의 독일어 단어로 번역되는 경우이다. 이런 경우에는 최대 3개의 독일어 단어만을 취하였다. 이 결과 영어 감정사전 ‘Subjectivity Clues’와는 엔트리의 수가 달라지게 되었다.

영어 단어를 소스로 하여 번역된 독일어 단어는 영어 단어가 갖는 긍정, 부정의 값을 그대로 전승받았다. 예를 들어 영어의 긍정의미 단어 'brave'는 독일어 'mutig'로 번역되며 'mutig' 또한 긍정의미 태그를 갖게 된다.

그 밖에 'nicht schlecht'와 같이 독일어에서 자주 등장하는 부정표현의 구문을 아예 사전의 엔트리로 추가했다. 이와 같은 과정을 거쳐 'GPC'는 최종적으로 원형기준으로 10,141개의 엔트리를 갖게 되었다.[5] 'GPC'의 엔트리에도 긍정, 부정관련 가중치가 부착되어 있다. 이 가중치는 독일 아마존 사이트의 상품평 코퍼스를 활용하여 계산하였다.

아마존 사이트의 상품평에는 작성자가 별표를 1개에서 5개까지 부여할 수 있는데, 별표 1개와 2개는 대개 부정적인 리뷰, 별표 4개와 5개는 긍정적인 리뷰라고 볼 수 있다. Waltinger (2010)는 어떤 단어의 부정 가중치를 해당 단어가 별표 1개 및 2개로 이루어진 부분코퍼스에 등장한 빈도를 전체 코퍼스에서 등장한 빈도로 나누어 계산하였다. 마찬가지로 긍정 가중치는 해당 단어가 별표 4개 및 5개로 이루어진 부분코퍼스에 등장한 빈도를 전체 코퍼스에 등장한 빈도로 나누어 계산하였다.

그러나 이러한 방법도 코퍼스의 특성에 따라 가중치의 신뢰도가 크게 좌우된다는 문제가 있다. 예를 들어 'empfehlenswert'의 긍정가중치는 0.093인데 반해, 동사형 'empfehlen'은 0.0040으로 약 23배의 차이가 있는데 이는 언어직관에 위배된다고 할 수 있을 것이다. 따라서 본 연구에서는 'GPC'의 경우에도 긍정, 부정 어휘 리스트만을 활용하고 가중치는 사용하지 않기로 한다.

3. 기계학습과 자질

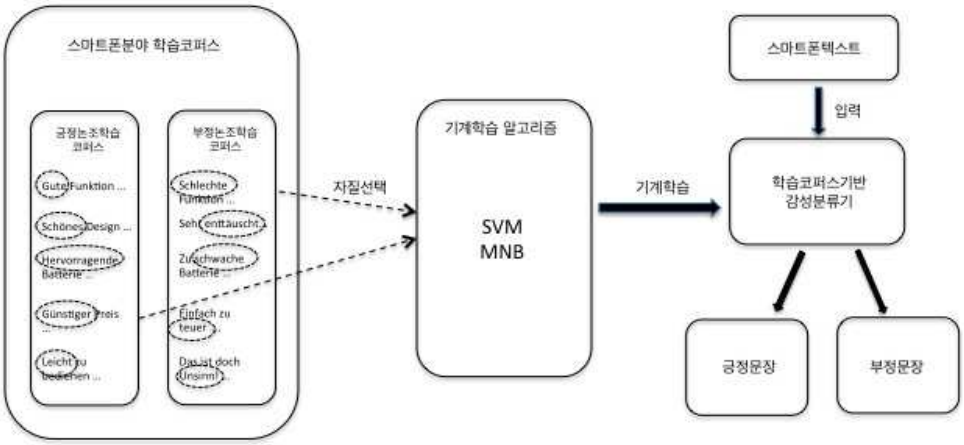
Murphy (2012)에 따르면 기계학습 maschinelles Lernen이란 학습데이터에 존재하는 패턴을 분석하고 이 패턴에 기반하여 새로운 데이터의 클래스를 예측하는 알고리즘에 대한 연구분야라고 볼 수 있다.

---

5) 실제 실험을 위해서는 활용형이 포함된 형태의 사전을 사용하였으며 엔트리의 규모는 37,589개이다.

<그림 1>은 학습코퍼스에 기반한 기계학습 방법을 보여준다. 기계학습을 위해서는 학습을 위한 데이터, 즉 학습코퍼스가 필요하다. 학습코퍼스는 일반적으로 클래스별로 나뉘어 있는데, 본 연구에서는 두 개의 클래스 (긍정/부정)만을 고려하므로 학습코퍼스의 문장들은 '긍정' 또는 '부정'이라는 클래스 정보가 부착되어 있다.

<그림 1> 학습코퍼스 기반 기계학습



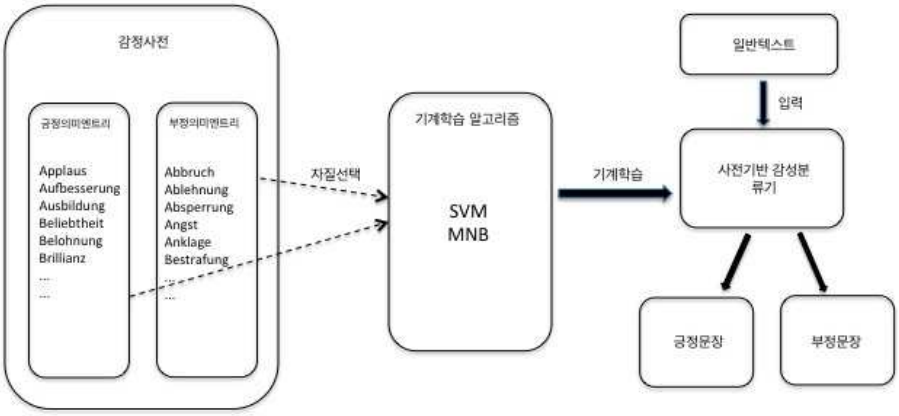
언어학적 직관이 개입하는 부분은 클래스별 학습코퍼스의 특징을 결정짓는다고 판단되는 자질 Merkmal을 선택하는 부분이다. 즉, 긍정논조 학습코퍼스의 '긍정성'을 결정하는 자질이 무엇인지, 부정논조 학습코퍼스의 '부정성'을 결정하는 자질이 무엇인지를 선택하는 것이다. 텍스트의 '긍정성'과 '부정성'을 결정하는 요소는 매우 많지만 일반적으로 클래스별 코퍼스에 상대적으로 자주 등장하는 유니그램이라고 할 수 있을 것이다. 그림 1의 예를 들면, 긍정논조학습코퍼스에서는 'gute', 'schönes', 'hervorragend', 'leicht', 'preiswert'를 자질로 선택하였다.

클래스별로 자질을 선택하게 되면 'SVM (Support Vector Machine)'이나 'MNB (Multinomial Naive Bayesian)'와 같은 컴퓨터공학 분야에서 널리 알려진 기계학습 알고리즘을 적용하여 최적의 함수를 만들어내게 되고 이것을 이용해 새로운 텍스트에 대한 감성분류를 시도하게 된다.

감정사전을 사용하여 감성분류기를 구현하는 방법은 <그림 2>에 나타나있

다. 이 방법도 기계학습을 이용하므로 전체적인 흐름은 학습코퍼스 기반의 방법론과 거의 동일하나 학습코퍼스를 사용하지 않고 감정사전의 엔트리를 자질로 그대로 사용한다는 점이 차이가 있다. 자질이 추출된 이후의 과정은 학습코퍼스 기반의 방법론과 동일하다.

<그림 2> 감정사전기반 기계학습



4. 감정사전 자질을 활용한 감성분석 성능분석

4.1 실험목적 및 방법

본 연구의 목적은 기계학습을 위한 학습코퍼스를 감정사전이 대체할 수 있는지의 여부를 알아보는 것이다. 기존의 방법인 학습코퍼스에서 감성분석을 위한 자질을 추출하고 이에 기반한 감성분석기를 구현하는 것은 도메인별로 많은 학습코퍼스가 필요하다는 한계가 있다. 이를 위해 특정 도메인에 국한되지 않는 독일어 감정사전의 적용방안과 사전별 성능을 비교한다.

연구의 또 다른 목적은 특정 도메인에서 학습된 자질을 다른 도메인에 적용했을 때 생기는 감성분석 성능의 변화를 관찰하는 것이다. 이를 통해 도메인별로 타도메인과 구별되는 특정 어휘들이 감성분석의 성능에 어느 정도 영향을 미치는지 보고자 하였다.

이를 위해 우선 실험도메인으로는 스마트폰과 호텔예약 분야를 선정하였다.

스마트폰과 호텔예약 분야를 실험 도메인으로 선정한 이유는 두 도메인의 연관성이 매우 작을 것으로 판단했기 때문이다. 예를 들어 스마트폰과 디지털 카메라와 같은 도메인을 설정하게 된다면 디지털 카메라가 가지고 있는 기능이 상당 부분 스마트폰에도 포함되어 있기 때문에 도메인 변화에 따른 감성분석 성능의 변화를 객관적으로 판단하는데 문제가 있을 것으로 예상하였다.

스마트폰 도메인 코퍼스는 긍정문장 500개, 부정문장 500개, 총 1천개의 문장으로 구성되어 있다. 마찬가지로 호텔예약 도메인 코퍼스도 긍정문장 500개, 부정문장 500개, 총 1천개의 문장으로 구성하였다.

감정사전 기반 감성분석 성능의 벤치마킹을 위해 학습코퍼스에서 자질을 추출하여 감성분석을 시도하는 Pang et al. (2002) 식의 감성분석을 베이스라인으로 선택하였다. 학습코퍼스로는 앞서 언급한 각각 1천 문장 규모의 스마트폰 분야 코퍼스, 호텔 분야 코퍼스를 사용하였다. Pang et al. (2002)의 방식은 학습코퍼스에 등장하는 유니그램을 학습자질로 택하여 감성분석을 수행한다. 방식이다.

기계학습 및 자동성능평가는 웨카 Weka 3.6.9버젼을 사용하였다. 기계학습을 위한 알고리즘으로는 ‘SVM (Support Vector Machine)’과 ‘MNB (Multinomial Naive Bayesian)’ 알고리즘을 택하였다. 홍문표 (2014)의 연구에 따르면 기계학습에 기반한 독일어 감성분석 알고리즘으로는 ‘SVM’과 ‘MNB’가 가장 높은 성능을 보이는 것으로 알려져 있다.

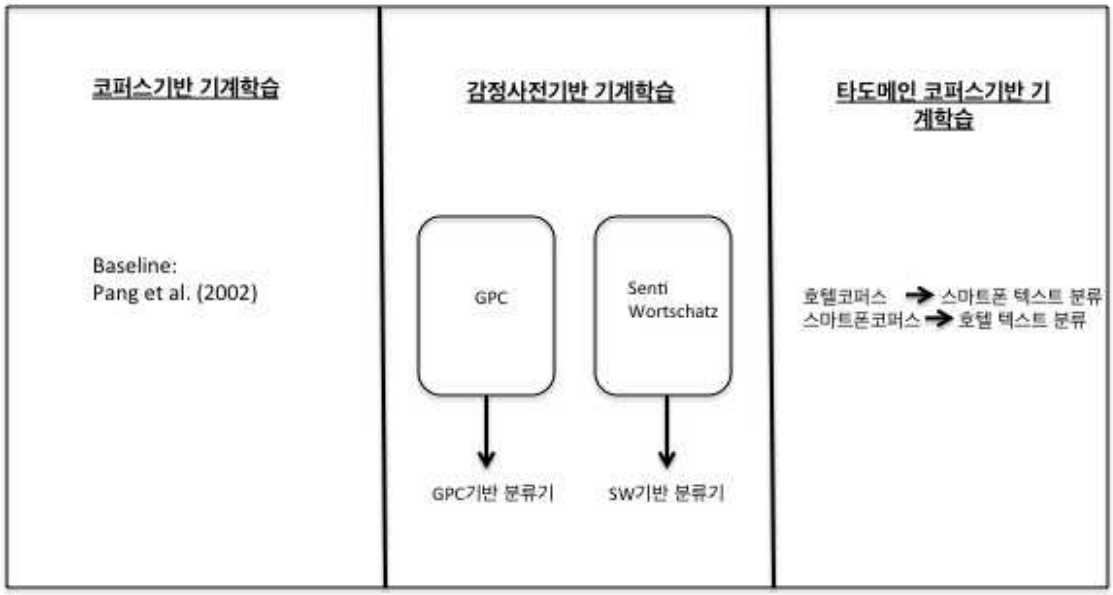
웨카를 사용하여 기계학습을 하기 위해서는 학습코퍼스를 ‘ARFF’ 파일 형식으로 변환하여야 하는데 이를 위해서는 웨카에서 기본으로 제공하는 ‘StringtoWordVector’ 툴을 활용하였다. 이 도구를 활용할 경우 문장 단위로 정렬되어 있는 학습코퍼스가 단어벡터의 형식으로 변환된다. 이 결과 학습코퍼스에 등장하는 유니그램들이 학습자질로 활용된다. 위와 같은 과정을 거쳐 스마트폰 분야의 학습코퍼스로부터는 3,475개의 학습자질이 추출되었으며, 호텔 분야의 학습코퍼스로부터는 3,278개의 학습자질이 추출되었다.

감정사전의 엔트리를 기계학습의 자질로 활용하기 위해서는 약간의 프로그래밍이 필요하다. 본 연구에서는 파이썬 3.3.2를 사용하여 감정사전 엔트리를 기계학습의 자질로 한 ‘ARFF’ 파일 생성 프로그램을 구현하였다. 이 결과 감정사전 엔트리가 자질로 사용된 단어벡터가 생성되었다. ‘GPC’ 사전으로부터

는 37,589개의 어휘를 추출하여 기계학습을 위한 자질로 사용하였으며, ‘SentiWS’ 사전으로부터는 31,270개의 어휘를 추출하여 기계학습을 위한 자질로 사용하였다.

성능평가는 학습코퍼스를 10등분하여 9등분한 부분으로 기계학습을 수행한 후 나머지 1등분된 부분에 대한 감성분석 성능을 평가하는 ‘10등분 교차검증(10-fold cross validation)’ 방식으로 수행하였다. 성능평가를 위한 척도는 정확률 Accuracy을 고려하였다. 정확률은 전체 분석대상 문장 중 올바르게 분석한 문장을 백분율로 나타낸 것으로서, 예를 들어 100문장 중 80문장의 감성 (긍정/부정)을 올바로 분석하였다면 정확률은 80%이다. 다음의 <그림 3>은 실험의 구성을 보여준다.

<그림 3> 실험구성도



4.2 실험결과

4.2.1 스마트폰 분야 감성분석

본 논문에서 제안하는 사전기반 감성분석 방법론의 직접적인 비교대상은 학습코퍼스로부터 자질을 추출하여 분석을 시도하는 Pang et al. (2002) 식의 방법이었다. 먼저 스마트폰 분야의 경우 학습코퍼스로부터 총 3,542개의 자질

을 추출하였고 ‘SVM’ 알고리즘을 적용한 감성분석기와 ‘MNB’ 알고리즘을 적용한 감성분석기를 구현하였다. ‘SVM’ 기반의 감성분석기는 78.8%의 정확률을 보였고, ‘MNB’기반의 감성분석기는 75.4%의 정확률을 보였다.

사전기반의 감성분석기 중 ‘GPC’ 사전에서 자질을 추출한 분석기는 ‘SVM’ 알고리즘을 적용한 경우 스마트폰 분야의 텍스트에 대하여 75.2%의 정확률을 보였으며, ‘MNB’알고리즘을 적용한 경우에는 76.3%의 정확률을 나타내었다. 이는 베이스라인 시스템인 학습코퍼스 기반 시스템과 비교하여 ‘SVM’을 적용했을 경우는 3.6%의 성능하락, ‘MNB’의 경우에는 0.9%의 성능향상을 보인 것이다.

‘SentiWS’에서 자질을 추출한 분석기는 ‘SVM’ 알고리즘을 적용한 경우 스마트폰 분야의 텍스트에 대하여 71.9%의 정확률을 보였으며, ‘MNB’를 적용했을 때에는 69.3%의 정확률을 나타냈다. 이는 베이스라인과 비교해서는 ‘SVM’의 경우는 9.5%의 정확률 하락, ‘MNB’의 경우에는 6.1%의 정확률이 하락한 것이다.

‘SentiWS’는 ‘GPC’와 비교하여서도 상대적으로 낮은 분류의 정확률을 기록하였는데 ‘SVM’ 알고리즘을 적용하였을 경우에는 3.3% 낮았으며, ‘MNB’를 적용하였을 때에는 7%가 낮았다. ‘SentiWS’가 ‘GPC’보다 낮은 성능을 보인 이유에 대해서는 4.3 결과분석에서 좀 더 자세히 다루도록 한다.

이어서 호텔분야의 학습코퍼스로 훈련된 분석기를 스마트폰 분야의 텍스트를 분류하는데 적용하였다. 이 결과 ‘SVM’에 대해서는 73.2%의 정확률, ‘MNB’에 대해서는 74.7%의 정확률이 나왔다. 이는 당초의 예상을 뛰어넘는 것으로서 베이스라인 시스템에 비하여 기계학습 알고리즘별로 각각 5.6%, 0.7%만이 하락하였다. ‘SentiWS’를 기반으로 하는 분류기와 비교해서는 심지어 약 1.3%, 5.4% 만큼의 더 높은 정확률을 보였다. 이를 정리하면 다음 표2와 같다.

	SVM	MNB
Baseline	**78.8**	75.4
GPC	75.2	**76.3**
SentiWS	71.9	69.3
Hotel Features	73.2	74.7

<표 2> 스마트폰분야 텍스트 감성분석 정확률

가장 높은 정확률은 스마트폰 분야의 학습코퍼스를 'SVM' 알고리즘으로 훈련시킨 분류기가 기록했다. 감정사전에서 자질을 추출하여 훈련한 분류기는 'GPC' 사전을 사용했을 때 'SVM'과 'MNB' 알고리즘에 대하여 고른 성능(75.2%, 76.3%)을 보였다. 특히 'MNB' 알고리즘을 사용한 경우에는 가장 높은 정확률 (76.3%)을 기록하여 학습코퍼스에 기반한 방법론보다 오히려 더 좋은 성능을 보였다. 그러나 'SentiWS' 사전으로부터 학습자질을 추출하여 구현한 분석기는 비교대상 중 가장 낮은 정확률을 기록하였다.

실험 결과 중 가장 예상과 다른 결과는 다른 분야의 학습코퍼스로 훈련된 분류기를 스마트폰 분야 텍스트에 적용한 경우이다. 일반적인 예상은 도메인별로 사용되는 어휘의 특징이 다르므로 타도메인에 빈출하는 어휘로 감성분석을 수행할 경우 분석정확률이 낮을 것이라고 예상이지만, 실험결과 베이스라인 대비 크게 뒤지지 않는 정확률을 보였다. 특히 'MNB' 알고리즘을 사용한 경우에는 불과 0.7% 밖에 차이가 나지 않아 학습코퍼스 기반의 방법론에 대한 일반적인 생각이 잘못 되었을 수도 있다는 점을 시사하였다.

4.2.2 호텔 분야 감성분석

호텔 분야 텍스트에 대한 감성분석은 스마트폰 분야보다는 전체적으로 분석 정확률이 조금 낮았다. 먼저 호텔 분야 코퍼스에서 자질을 추출하여 기계학습을 시도한 베이스라인 시스템은 'SVM' 알고리즘에 대해서는 72.3%의 정확률, 'MNB'에 대해서는 75.7%의 정확률을 나타내었다. 이 때 활용된 자질의 수는 모두 3,321개였다.

다음은 'GPC' 사전에서 자질을 추출하여 기계학습을 한 분석기의 성능이다. 이 분석기는 'SVM' 알고리즘을 활용한 경우에는 75.1%의 정확률을 보였

고, ‘MNB’ 알고리즘으로 학습된 경우에는 74.2%의 정확률을 보였다. 최고 정확률에 있어서는 학습코퍼스 기반의 방법론보다 0.6% 정도 낮았지만, 학습알고리즘에 대해 고루 높은 정확률을 보였다.

스마트폰 분야에 대해 가장 낮은 분석정확률을 보였던 ‘SentiWS’ 사전은 이번에도 가장 낮은 성능을 보였다. 호텔 분야에 대해서는 ‘SVM’ 알고리즘 적용시 67.5%, ‘MNB’ 알고리즘 적용시 65.1%의 정확률을 보여 베이스라인 대비 거의 최고 10% 정도의 차이를 보였다.

마지막으로 호텔 분야의 텍스트에 대해서도 스마트폰 분야의 학습코퍼스에서 추출한 자질을 사용하여 개발한 분석기를 적용해 보았다. ‘SVM’ 알고리즘 적용시에는 70.8%의 정확률을 나타냈고, ‘MNB’ 알고리즘 적용시에는 73.9%의 정확률을 보였다. 이는 베이스라인과 비교해서는 ‘SVM’ 적용시 1.5%, ‘MNB’ 적용시 1.8% 밖에 차이가 나지 않는 수치이다.

	SVM	MNB
Baseline	72.3	**75.7**
GPC	75.1	74.2
SentiWS	65.1	67.5
Phone Features	70.8	73.9

<표 3> 호텔 분야 텍스트 감성분석 정확률

<표 3>에서 볼 수 있는 바와 같이 호텔분야에 대한 분석 정확률은 호텔 분야의 학습코퍼스에서 자질을 추출하여 ‘MNB’ 알고리즘을 적용했을 때 얻어진 75.7%였다. ‘GPC’ 사전의 엔트리를 자질로 활용하여 ‘SVM’ 알고리즘을 적용한 경우에는 정확률이 75.1%로서 최고 정확률에 근접하였으며, ‘SVM’ 알고리즘으로 학습한 경우 중에서는 가장 높은 성능을 보였다.

‘GPC’ 사전에서 자질을 추출하여 기계학습을 수행한 경우는 기계학습의 알고리즘에 상관없이 고른 성능을 보였다. 그러나 ‘SentiWS’ 사전에서 자질을 추출하여 분석기를 구현한 경우에는 휴대폰 분야의 학습코퍼스에서 자질을 추출하여 호텔분야 텍스트의 감성분석기를 구현한 경우보다도 오히려 성능이 낮음을 볼 수 있었다. 앞서 스마트폰 분야의 분석에서도 유사한 현상이 나타

났는데 이를 통해 ‘SentiWS’ 사전은 현재로서는 독일어 텍스트의 감성분석에 적합하지 못하다는 것을 알 수 있었다.

4.3 결과분석

4.2장에서 본 바와 같이 도메인별로 감성분석의 최고 성능은 해당 도메인의 학습코퍼스를 활용하여 기계학습을 수행한 시스템들이 기록하였다. 그러나 ‘GPC’ 사전을 활용하여도 도메인별 학습코퍼스를 활용할 때와 크게 차이가 없는 결과를 얻을 수 있음을 보았다. 오히려 호텔분야의 경우에는 평균 정확률이 학습코퍼스 기반의 방법론보다 높았다.

‘GPC’와 거의 비슷한 방식으로 구축된 ‘SentiWS’ 사전의 경우에는 가장 낮은 성능을 보였는데, 이 장에서는 이에 대한 원인을 좀 더 자세히 알아보고자 한다. 또한 다른 도메인으로 학습한 감성분석기가 도메인에 상관없이 높은 성능을 보이는 이유 또한 알아보고자 한다.

이를 위해 웨카 시스템이 제공하는 자질선택 기능을 활용하였다. 웨카 시스템은 감성분석기의 어떤 자질이 분석기의 성능을 높이는데 가장 큰 기여를 하는지를 자동으로 알 수 있게 해주는 기능을 제공한다. 여기서는 ‘InfoGain’ 계산법으로 순위가 매겨진 자질 리스트를 살펴보기로 한다.

다음의 <표 4>는 휴대폰 분야의 학습코퍼스에서 추출된 총 3,542개의 자질 중 기계학습시 긍정과 부정의 논조를 구별하는데 가장 큰 역할을 한 최상위 50개의 자질을 보여준다. 예상과는 다르게 부정부사 ‘nicht’가 가장 큰 역할을 하는 것으로 밝혀졌다. 이는 코퍼스의 크기가 비교적 작은 규모이기 때문에 발생한 일종의 노이즈 noise라고 볼 수도 있겠지만, 동일한 방식으로 수행한 호텔예약 분야의 자질선택 실험에서도 ‘nicht’가 가장 높은 순위를 차지한 것으로 볼 때, 단순히 코퍼스의 규모에서 발생한 노이즈로 무시할 수는 없을 것으로 보인다. 문장의 논조를 결정하는데 ‘nicht’가 보이는 이러한 경향성은 추후의 연구과제로 남겨야 할 것이다.

그 밖에 최상위 50개의 자질들을 살펴보면 예상대로 긍정과 부정을 나타내는 대표적인 어휘 ‘gut’, ‘guten’, ‘gute’, ‘leider’, ‘dank’, ‘unscharf’, ‘klappt’, ‘Manko’, ‘fehlt’, ‘gelingt’, ‘clever’, ‘angenehm’, ‘hervorragend’, ‘Vorteil’, ‘sauber’

등이었다. 이 어휘들은 스마트폰이라는 도메인에 국한된다기 보다는 범도메인적인 어휘라고 볼 수 있을 것이다.

그 밖의 자질로는 ‘galaxy’, ‘s3’, ‘lg’, ‘s4’, ‘x’, ‘g2’, ‘desire’, ‘nokia’ 등과 같은 제품명이나 회사명 등의 고유명사들이었다. 이들은 도메인 코퍼스에서만 찾을 수 있는 어휘들로서 학습코퍼스기반 방법론과 일반감정사전기반 방법론과의 성능차이에 대한 설명이 될 수 있을 것이다. 나머지의 자질들은 기능어들 및 일부 도메인에 빈출하는 일반명사들[6] (예를 들어 ‘speicher’, ‘auflösung’, ‘schärfe’, ‘darstellung’등) 이었다.

	기계학습자질순위		
1	nicht	26	z
2	gut	27	g2
3	sehr	28	insgesamt
4	und	29	möglich
5	guten	30	guten[7]
6	gute	31	auflösung
7	mit	32	manko
8	zu	33	wird
9	das	34	wirken
10	galaxy	35	eher
11	hand	36	schärfe
12	allerdings	37	telefonieren
13	liegt	38	müssen
14	s3	39	desire
15	leider	40	nokia
16	lg	41	fällt
17	s4	42	fehlt
18	lediglich	43	gelingt
19	dank	44	clever
20	unscharf	45	angenehm
21	etwas	46	hervorragend

6) 본 실험에서는 학습코퍼스를 먼저 소문자화한 후 벡터파일로 변환하였으므로 명사들이 모두 소문자로 표기됨을 밝힘.

7) 이 표에는 ‘guten’이 5위와 30위에 등장하는데 30위의 ‘guten’은 눈에 보이지 않는 특수문자가 포함된 경우로 이를 모두 ‘guten’으로 간주한다면 ‘guten’의 순위는 보

22	speicher	47	vorteil
23	klappt	48	unsere
24	einem	49	sauber
25	x	50	darstellung

<표 4> 스마트폰 학습코퍼스에서 추출한 자질 순위

이어서 <표 5>는 'GPC'와 'SentiWS'에서 추출한 상위자질들이다. 'GPC'와 'SentiWS'의 가장 큰 차이는 'nicht'의 포함여부이다. 'GPC'는 'SentiWS'와 다르게 사전 컴파일과정에서 부정부사 'nicht' 및 'nicht'와 빈출하는 형용사 (예를 들어 'nicht schlecht')를 일반 엔트리로 등록하였다. 독일어 텍스트의 감성분석에서 부정어의 처리는 홍문표 (2013)에서 보인 바와 같이 정확률에 매우 큰 영향을 미친다. 'nicht'에 대한 고려가 없다면 'nicht schlecht'는 'schlecht'가 갖는 '부정'의 논조 때문에 '부정'으로 분류될 가능성이 있다. 그러나 'GPC'에는 'nicht schlecht'가 하나의 엔트리로서 '긍정' 논조로 태그되어 있다. 물론 'nicht schlecht'를 하나의 엔트리로 취급함으로써 'nicht'와 'schlecht'가 떨어져서 출현하는 경우는 처리하지 못하는 단점이 있다. 그럼에도 불구하고 부정어에 대한 처리가 전혀 없는 'SentiWS'보다는 정확한 분류가 가능할 것으로 예상되며, 실제로 본 실험 결과에서도 'SentiWS'와 비교하여 스마트폰 분야 텍스트의 감성분석에서 3.3%~7%만큼의 정확률 차이를 보였다.

4.2의 실험결과중 당초의 예상과 달리 비교적 높은 성능을 보인 경우는 다른 도메인의 학습코퍼스로 학습된 분석기로 특정 도메인의 감성분석을 시도한 경우이다. 호텔 분야의 학습코퍼스로 학습된 분석기를 스마트폰 분야 텍스트의 감성분석에 적용한 경우 'SVM'에 대해서는 73.2%, 'MNB'에 대해서는 74.7%의 정확률이 측정되었다. 이는 베이스라인과 비교하여 0.7%~5.6% 정도의 성능차이를 보이는 것으로서 오히려 'SentiWS' 사전을 적용했을 때 보다는 높은 정확률을 보였다.

<표 6>은 호텔분야 학습자질을 사용하여 스마트폰 분야 텍스트의 감성분석을 수행했을 때 사용된 상위 30개의 자질이다. 이 중 굵은 글씨로 표기된 12개의 자질은 스마트폰 분야 학습코퍼스에서 추출한 상위 30개의 자질안에 포

다 높아질 것으로 예상된다.

함되어 있는 것들이다. 즉, 학습코퍼스의 도메인이 바뀌더라도 감성분석에 적용되는 공통의 자질들이 다수 존재하기 때문에 타도메인의 학습자질을 적용해도 일정 수준 이상의 정확률이 유지되는 것으로 판단된다.

5. 결론

본 논문은 학습코퍼스에서 자질을 추출하여 기계학습을 시도하는 기존의 감성분석 방법론의 문제를 해결하기 위해 사전기반 감성분석 방법론을 제안하였다. 기존 감성분석 방법론의 문제는 도메인별로 큰 규모의 학습코퍼스가 필요하므로 분석방법을 다양한 도메인에 적용하기가 현실적으로 어렵다는 점이었다. 본 연구에서 제안한 사전기반 감성분석 방법론은 'GPC'와 ' SentiWS'와 같은 감정사전의 엔트리를 기계학습을 위한 자질로 삼아 학습하는 방법이다. 이 방법론은 도메인에 상관없이 공통적으로 출현하는 감정관련 어휘들이 수록된 감정사전을 활용하므로 도메인 변화에 따른 성능의 변화를 최소화할 수 있을 것이라는 점이 예상되었다. 이 가설의 검증을 위해 스마트폰 분야와 호텔 분야의 텍스트에 대해 실험을 수행하였다. 실험결과 각 도메인별 학습코퍼스에서 자질을 추출하여 학습한 분석기가 가장 높은 성능을 보였지만, 감정사전을 기반으로 학습된 분석기도 학습코퍼스 기반 분석기에 비해 크게 뒤지지 않는 성능을 보였다.

감정사전 중 'GPC'와 'SentiWS'는 서로 매우 큰 성능의 차이를 보였는데, 이는 부정어의 처리에 기인한 것으로 분석되었다. 'GPC'가 'nicht' 및 'nicht'와 자주 공기하는 가치판단 형용사 (예를 들어 'nicht schlecht') 등을 하나의 엔트리로 등록하였기 때문에 이러한 표현들에 대한 감성분석이 'SentiWS'와 비교하여 더욱 정확했던 것으로 판단된다.

본 연구를 통한 또 하나의 발견은 학습자질의 도메인 의존성이었다. 당초 본 연구의 가설은 학습자질이 학습도메인에 크게 의존할 것이기 때문에 어떤 도메인에서 추출된 학습자질들은 다른 도메인에 적용될 경우 분석 정확률이 크게 낮아질 것이라는 것이었다. 그러나 실험결과 호텔분야의 학습자질들을 사용하여 스마트폰 분야의 텍스트에 대해 감성분석을 수행했을 경우 성능의

차이가 예상했던 것보다는 크지 않았다는 점이다. 심지어는 'SentiWS' 사전을 사용했을 때보다도 오히려 정확률이 높았다. 이것에 대한 분석결과 도메인의 변화에도 불구하고 공통적으로 적용되는 자질들이 많이 존재한다는 것을 알게 되었다.

이 점은 본 연구의 후속작업을 위해 시사하는 점이 많다고 본다. 도메인에 상관없이 공통적으로 적용되는 자질들이 많다는 점은 범도메인적인 성격을 지닌 감정사전의 유용성을 보여준다. 감정사전의 엔트리를 확장하고 감정정보 가중치의 직관적 타당성을 높이면 그 유용성은 더 높아질 수 있을 것으로 예상된다.

	GPC	SentiWS		GPC	SentiWS
1	gut	gut	36	dunkle	verzichte
2	nicht	gute	37	schwache	schade
3	gute	ehr	38	problemlos	verzicht
4	sehr	guten	39	unhandlich	last
5	ehr	leid	40	Kauf	wunder
6	Lech	schwach	41	chic	bessert
7	guten	unscharf	42	Wunde	klare
8	Leid	schlecht	43	schick	clevere
9	los	hervorragend	44	schade	erziel
10	leid	leider	45	Last	erzielt
11	schwach	clever	46	Verzichte	erfreu
12	unscharf	dank	47	Verzicht	vorteil
13	schlecht	hervorragende	48	Wunder	top
14	leider	kritik	49	clevere	nahe
15	hervorragend	schlechte	50	intuitiv	fallen
16	clever	dunkel	51	klare	zoll
17	fehl	gelungen	52	hohen	teure
18	dank	tadel	53	erziel	schwere
19	Dank	sauber	54	verbessert	sorgte
20	echt	tadellos	55	bessert	verzichten
21	hervorragende	tadel	56	erzielt	schmerz
22	Kritik	schwer	57	Einerlei	schwerer
23	rage	dunkle	58	erfreu	dunkler

24	gern	fehlt	59	recht	schmerz
25	schlechte	problemlos	60	dunkler	arm
26	dunkel	schwache	61	sorgte	verzichten
27	Dunkel	guter	62	unscharfe	negativ
28	lax	kauf	63	negativ	unscharfe
29	Gala	schick	64	schlechten	schlechten
30	tadel	wunde	65	entfernt	warm
31	tadellos	chic	66	mager	sorgte
32	sauber	recht	67	blau	majestätischer
33	Tadel	fehle	68	Gelingen	ehre
34	gelungen	angenehm	69	anti	wunderbar
35	fehlt	genehm	70	schicke	ehre

<표 5> ‘GPC’와 ‘SentiWS’에서 추출한 상위자질 순위

	기계학습 자질순위		
1	**gut**	16	**mit**
2	**nicht**	17	cher
3	**gute**	18	ch
4	**sehr**	19	schlecht
5	ht	20	**allerdings**
6	**und**	21	**hervorragend**
7	**guten**	22	leider
8	schw	23	vor
9	leid	24	**das**
10	los	25	**dank**
11	hle	26	w
12	ich	27	**einer**
13	schwach	28	karte
14	da	29	arbeit
15	aller	30	uscht

<표 6> 스마트폰 분야에 적용된 호텔 분야 자질 상위 30개

참고문헌

홍문표 (2013): 독일어 텍스트 논조자동분석. 『독어학』 28, 361-383.

홍문표 (2014): 바이그램을 활용한 텍스트 논조자동분석. 『독일언어문학』 65, 27-46.

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.

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

Pang, B., L. Lee & 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. & L. Lee (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.

Quasthoff, U. (2010): *Deutsches Kollokationswörterbuch.* de Gruyter. Berlin, New York.

Remus, R. U. Quasthoff & G. Heyer (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.

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

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & M. Stede (2011): Lexicon-based methods for sentiment analysis. In: *Computational linguistics* 37 (2), 267-307.

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

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

Zusammenfassung

Automatische Sentimentanalyse der deutschen Texte anhand eines Sentiment-Lexikons

Hong, Mun Pyo (Sungkyunkwan Univ.)

Die vorliegende Arbeit stellt eine neue Methode für die Sentimentanalyse deutscher Texte vor. Die meisten bisherigen Ansätze für die Sentimentanalyse oder Stimmungsanalyse sind von Pang et al. (2002) stark beeinflusst. In dieser Arbeit wurde ein statistisches Verfahren vorgeschlagen, das auf annotierten Lernkorpora angelernt wurde.

Ein Problem dieser Methode liegt darin, dass eine große Menge der Lernkorpora für jede Domäne benötigt wird, um einen Klassifizierer für verschiedene Domänen zu implementieren. Generell wird angenommen, dass ein Klassifizierer, der auf Lernkorpora bestimmter Domäne angelernt wurde, eine schlechte Performanz für eine andere Domäne aufweisen würde.

Die neue Methode zieht ein Sentiment-Lexikon statt der Lernkorpora für das maschinelle Lernen heran. Einträge im Sentiment-Lexikon haben Informationen über die Positivität oder Negativität in ihren konnotativen Bedeutungen. Z.B. ist ‘schön’ ein Wort mit einer positiven konnotativen Bedeutung. Dagegen weist z.B. ‘Debakel’ eine negative konnotative Bedeutung auf. In dieser Arbeit wurden ‘German Polarity Clues (GPC)’ von Waltinger (2010) und ‘Senti Wortschatz (SentiWS)’ von Remus et al. (2010) benutzt.

Das Experiment zeigte, dass die Einträge eines Sentimentlexikons gute Merkmale für das maschinelle Lernen unabhängig von verschiedenen Domänen aufweisen können. Jedoch gab es Unterschiede in der Performanz zwischen ‘GPC’ und ‘SentiWS’. ‘SentiWS’ unterliegt ‘GPC’ für fast 5 bis 10 % in der Accuracy der Sentimentklassifizierung. Unsere Analyse zeigt, dass der Unterschied in der Behandlungsweise der negativen Wörter liegen könnte.

[검색어] 감성분석, 논조분석, 감정사전, 학습코퍼스, 기계학습
Sentimentanalyse, Opinion Mining, Sentiment-Lexikon, Lernkorpus, Maschinelles Lernen

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

논문 접수일: 2014. 10. 30
논문 심사일: 2014. 11. 30
게재 확정일: 2014. 12. 17