자동논조분석 시스템 개발을 위한 독일어 텍스트의 논조별 어휘출현양상 연구*

홍문표 (성균관대)

1. 서론

본 논문은 특정 주제에 대해 표출되는 독일어 텍스트의 감정 Sentiment 내지는 논조를 자동으로 분석하는 시스템의 개발을 위한 기초연구로서, 보다 구체적으로는 특정 논조에서 나타나는 독일어 어휘들의 분포 양상을 조사하는 것이 연구의 목표이다.

자동논조분석 Automatic Sentiment Analysis[1]은 어떤 주제에 대해 텍스트가 긍정적인 내용을 담고 있는지, 부정적인지, 혹은 중립적인지를 밝히는 일종의 문서분류 Klassifikation작업이다. 자동논조분석에 관한 연구들은 모두 문장의 논조를 결정하는 어휘가 있다는 가정으로부터 출발한다. 즉, 긍정적인 내용의 문장에는 긍정적인 어휘, 예를 들어, 'gut', 'schön' 등이 많이 등장하고, 부정적인 내용의 문장에는 'schlecht', 'hässlich' 등과 같은 부정적인 어휘가 많이 출현할 것이라는 가정이다. 언어학 지식에 기반한 기호적 접근법 Symbolic Approach이나, 통계지식에 기반한 통계학적 접근법 Statistical Approach 모두 긍정/부정의 어휘리스트를 기반으로 출발한다는 점이 공통적이다.

본 논문에서는 독일어 블로그 및 독일 웹사이트 등에서 특정 주제에 대해 개인의 의견 및 감정이 노출된 문장들을 수집하고, 수집한 문장에 대해 '긍정적'/'부정적'/'혼합'/'객관적' 등과 같은 감정표지를 부착한 후, 코퍼스 처리 프로그램

* 이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음

1 자동논조분석은 자연언어처리 분야에서는 'opinion mining'이라고 불리기도 한다.

을 활용하여 감정 또는 논조에 따라 어휘 출현의 양상이 어떻게 다른지 분석한다.

2장에서는 본 연구와 관련된 연구들을 소개한다. 3장에서는 본 연구를 위해 구축한 독일어 논조정보부착 코퍼스와 코퍼스 처리를 위해 사용한 프로그램을 소개한다. 4장에서는 본 연구에서 수행한 실험결과를 보이고 결과를 분석한다. 5장에서는 연구의 결과 및 의의, 향후 연구방향 등을 다룬다.

2. 관련연구

자동논조분석과 관련된 연구동향은 크게 세가지로 나눌 수 있다. 우선 기존의 워드넷에 감정 혹은 논조정보를 추가적으로 부착한 'SentiWordnet'이나 'Wordnet-Affect' 등에 관한 연구이다. 이러한 확장 워드넷에는 표제어에 기존의 의미관계 이외에 논조분석 등에 활용될 수 있는 긍정/부정과 관련된 극성 Polarität정보가 추가로 부착되어있다. Kamps et al. (2002), Strapparava et al. (2004), Esuli & Sebastiani(2006) 등의 연구가 이와 관련된 대표적인 연구이다.

자동논조분석의 두 번째 연구동향은 'SentiWordnet' 등을 활용한 자동논조분석 알고리즘 연구이다. 이와 관련된 연구는 다시 크게 두가지로 나뉜다. 첫째는 문장의미의 주관성/객관성 판별에 관한 연구이다. 텍스트의 논조를 판단하기 위해서는 우선적으로 텍스트를 구성하는 각 문장의 주관성/객관성 판단이 필요하다. 이에 관한 대표적인 연구로는 Pang & Lee(2004), Riloff et al. (2003), Liu et al. (2005) 등을 들 수 있다. 자동논조분석 알고리즘 연구의 두 번째 연구경향은 문장의 긍정/부정 극성을 판별하여 그 강도 Intensität를 판별하는 연구이다. 이와 관련된 연구로는 Pang & Lee (2005), Wilson et al. (2005), Butscher(2005), Scaffidi et al. (2007) 등이 있다.

마지막으로 자동논조분석과 관련된 세 번째 연구경향은 논조정보부착 말뭉치 구축에 관한 것이다. 본 연구에서 밝히는 바와 같이 논조정보가 부착된 말뭉치는 자동논조분석을 언어학 지식에 기반한 규칙기반 방법론으로 시도하건, 통계기반 방법론으로 시도하던지 상관없이 공통적으로 필요한 언어자원이다. 따라서 각 언어별로 이러한 코퍼스를 구축하기 위한 기초 연구가 진행되어 왔다. 대표적인 논조정보부착 말뭉치로는 영어권에서는 영화에 대한 관객들의 평가정보가 부착되

어 있는 ‘Cornell movie-review datasets’[2]가 있고, ‘NTCIR’ 코퍼스에는 영어 뿐만 아니라 일본어와 중국어에 대해서도 논조정보가 부착되어 있다.

3. 독일어 논조정보부착 코퍼스

3.1. 코퍼스 개요

현재까지는 논조정보가 부착된 공개 독일어 코퍼스가 존재하지 않으므로, 본 연구에서는 우선 논조정보를 부착한 소규모 코퍼스를 구축하였다. 논조정보 혹은 논조태그 Sentiment Tag가 부착된 코퍼스는 순수 언어학 관점에서의 연구 뿐만 아니라 논조정보 사전을 구축하거나 기계학습 등을 통해 문서의 논조정보를 파악하는 방법론을 개발하기 위해서도 반드시 필요한 언어자원이다.

본 연구를 위해 구축된 코퍼스는 신규 휴대폰에 대한 리뷰 review 및 사용경험기 등에서 추출한 문장들로 구성되었다. 휴대폰 분야를 코퍼스 구축의 대상으로 선정한 이유는 독일시장이 국내 휴대폰 생산업체들에게 매우 중요한 시장 중의 하나이고, 실제로 이 때문에 국내기업들이 독일 네티즌들의 여론 동향을 면밀히 모니터링하고 있는 분야이기 때문이다. 이를 위해 독일의 휴대폰 분야 주요 웹사이트 및 블로그를 참조하였으며, 일부 오스트리아와 스위스에서 운영되는 사이트들의 텍스트들도 포함되었다.

이러한 사이트 등에서 추출된 텍스트를 문장단위로 정렬한 후, 2명의 작업자가 ‘positiv/negativ/gemischt/objektiv’의 태그정보를 문장단위로 부착, 검수하였다.[3] ‘positiv’ 태그는 특정 주제 (특정 휴대폰)에 대해 긍정적인 논조를 담고 있다고 판단되는 문장에 부착되었으며, 그 예는 다음과 같다.

긍정논조문장:

<positiv>Die sanft abgerundete Metalleinfassung des Displays perfektioniert die edle

2 URL: http://www.cs.cornell.edu/people/pabo/movie-review-data

3 보다 정교한 태깅작업을 위해서는 다수명의 작업자가 태깅을 한 후, 상호교차검수를 해야하나, 본 연구에서는 시간과 비용의 문제로 2명만이 이 작업을 수행하였다.

Haptik.</positiv>
<positiv>In der Tat harmonieren alle Tasten perfekt mit der Gehäuseoberfläche, sind aber trotzdem gut erfühlbar und bieten erstklassige Druckpunkte.</positiv>
<positiv>Die Bildqualität ist großartig, was sicher an dem hochauflösenden Display mit 160 Punkten pro Zoll liegt.</positiv>

'negativ' 태그는 특정 휴대폰에 대해 부정적인 논조 및 내용을 담고 있는 문장에 부착되었으며, 그 예는 다음과 같다.

부정논조문장:

<negativ>Ich war schon so weit, das Gerät zurückzugeben, weil mir die Bedienung gar zu mysteriös erschien.</negativ>
<negativ>Das herunterladbare von der Webseite ist völlig unzureichend und beschränkt sich im wesentlichen auf die Auflistung der Menüpunkte.</negativ>
<negativ>Das Niveau und die Detailtiefe ist lächerlich.</negativ>

'gemischt' 태그는 한 문장 내에 긍정과 부정 두 가지의 논조가 섞여 있을 경우에 부착되었다. 대개의 경우 긍정적인 논조를 담고 있는 단문과, 부정적인 내용을 담고 있는 단문으로 구성된 복문이므로, 긍정이나 부정의 논조만을 포함하는 문장들보다는 평균 문장길이가 길었다.

혼합논조문장:

<gemischt>Der wird praktischerweise von Lochrastern für Lautsprecher und Mikrofon flankiert - sieht hübsch aus, mitunter kann es aber vorkommen, dass die Hände die Boxen komplett verdecken, so dass das Handy keinen Mucks mehr von sich gibt.</gemischt>
<gemischt>Das X1 kommt in einer weniger schicken, dafür aber gut aufgeteilten und funktionellen Verpackung daher.</gemischt>
<gemischt>Ich bin auch mit sehr vielen Dingen zufrieden, nur finde ich Wiederholungen in Threads nicht so toll.</gemischt>

'objektiv' 태그는 긍정, 부정 등과 같은 주관적인 의견이 아니라, 객관적 사실

등만을 담은 문장에 부착되었다. 어떤 제품에 대한 의견을 수록한 텍스트에도 이러한 객관적인 문장은 다수 등장한다. 이 태그는 본 연구에서 직접적으로 활용되지는 않았으나, 향후 자동논조분석을 위해서는 매우 유용하게 사용될 것으로 예상된다. 왜냐하면 자동논조분석을 위해서는 우선 입력문장이 주관적인 내용을 담고 있는지 객관적인 사실만을 담고 있는지 분석을 해야 하는데, 이를 위한 통계자료로 활용될 수 있기 때문이다.

객관적 문장:

<objektiv>Ich habe mir das Handy vor fast eineinhalb Jahren gekauft.</objektiv>
<objektiv>Ich habe dieses Handy von meiner Mutter geschenkt bekommen weil sie damit nicht klar kam, nun habe ich es 4 Monate!!!!</objektiv>
<objektiv>Für diejenigen die sich beschweren der Akku halte nicht lange, wie wärs mit Bluetooth ausschalten oder diejenigen die nicht checken.</objektiv>

문장의 논조를 정확히 파악하는 것은 매우 어려운 작업이다. 논조란 주관적인 정보이므로 태깅작업을 하는 사람의 판단에 따라 다를 수 있다. 예를 들어 다음과 같은 문장은 문맥에 대한 정확한 이해가 없으면, 긍정적인 논조인지 혹은 부정적인 논조인지를 파악하기조차 어렵다.

(2) Das iPhone ist das iPhone.
(3) In der kleinen Box steckt viel Zubehör.

(2)번 문장에서 “iPhone은 iPhone일 뿐이다”라는 의미는 문맥에 따라 긍정적인 의미일 수도 있으며, 경우에 따라서는 부정적인 뉘앙스를 전달할 수도 있다. 또한 (3)번 문장은 객관적인 사실을 전달한다고 볼 수도 있지만, 때에 따라서는 ‘viel Zubehör’가 긍정적인 논조를 전달한다고 볼 수도 있을 것이다. 본 연구에서는 긍정 혹은 부정에 대한 명백한 단서가 없는 한, 위 문장들과 같은 경우는 모두 ‘objektiv’ 정보를 갖는 것으로 간주되었다.

위와 같은 방식으로 본 연구를 위해 구축된 논조부착 독일어 코퍼스는 총 3,667문장 규모이다. 단어수로는 총 51,314단어이며, 문장당 평균 13.99단어이다. 이 중 긍정적인 논조를 갖는 문장은 1,350문장이며, 총 17,967단어로서 문장당

평균 13.30단어이다. 부정논조를 갖는 문장은 1,062문장으로, 총 14,509단어로 이루어져 있으며, 문장당 평균 13.66단어를 갖는다. 혼합 논조의 문장은 총 222문장이며, 4,373단어이다. 문장당 평균 단어수는 16.70으로 앞서 언급한 바와 같이 긍정과 부정 논조만을 갖는 문장들보다 평균 단어수가 많다. 객관적인 문장은 1,033문장이며, 총 14,465단어로 이루어져 있고, 문장당 평균 14.01단어였다.

종합해보면, 긍정과 부정 논조만을 지닌 문장들은 평균단어수가 코퍼스 전체 평균단어수보다 약간 적은 반면, 객관적인 문장들은 코퍼스 전체 평균단어수와 거의 단어수가 비슷하며, 혼합논조의 문장들은 코퍼스 전체 평균단어수보다 월등히 평균 단어수가 많음을 알 수 있었다. (표 1 참조)

	문장수	총단어수	평균단어수/문장
전체코퍼스	3,667	51,314	13.99
긍정논조	1,350	17,967	13.30
부정논조	1,062	14,509	13.66
혼합논조	222	4,373	16.70
객관적 사실	1,033	14,465	14.01

표 1 : 독일어 '논조' 정보부착 말뭉치 규모

3.2 코퍼스 분석도구

구축한 코퍼스에서 단어의 빈도수와 바이그램 bigram 빈도를 추출하기 위해 공개 코퍼스분석 소프트웨어인 'Antconc' 시스템을 사용하였다. 'Antconc'는 일본 와세다대학의 로렌스 안토니 Laurence Anthony에 의해 개발되었으며[4], 유니코드를 처리할 수 있으므로, 독일어의 움라우트와 에스체트 및 한글의 처리에 유리하다.

이 프로그램은 입력파일에 대해 코퍼스 분석프로그램이 일반적으로 제공하는 'Concordance' 정보 이외에 단어빈도수 및 바이그램 빈도까지 자동으로 분석하여 출력하는 장점을 가지고 있다. 그림 1은 이 프로그램을 사용하여 바이그램 정보를 추출하는 모습을 보여준다.

4 'Antconc' 프로그램은 http://www.antlab.sci.waseda.ac.jp/에서 무료로 다운로드 받을 수 있다.

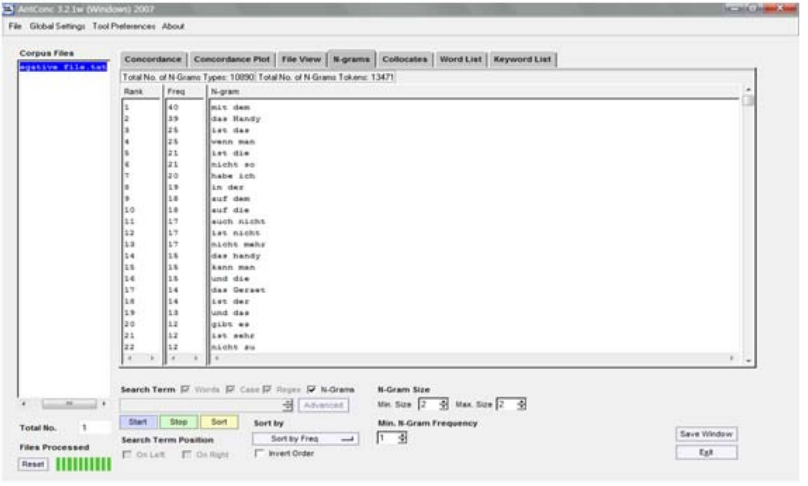

그림 1 : 'Antconc'를 통한 바이그램 추출모습

4. 실험

4.1. '긍정' 논조 코퍼스

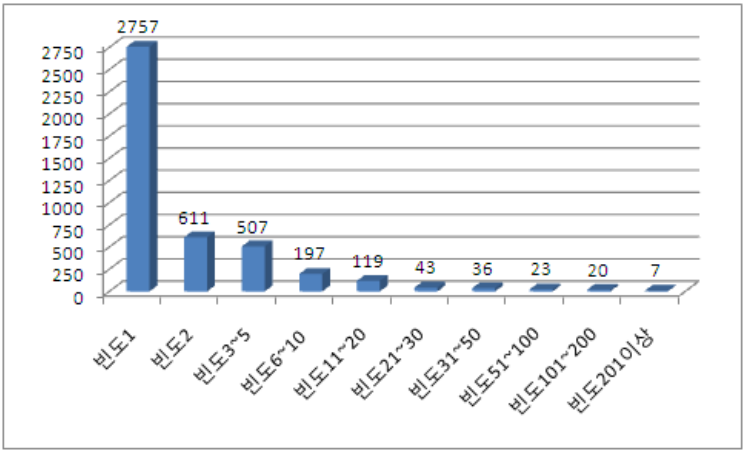

그림 2 : '긍정' 논조 코퍼스 출현 빈도수별 어휘수

본 연구의 실험에서는 우선 '긍정', '부정', '혼합', '객관적'으로 분류된 휴대폰 분야의 문장들에서 가장 빈번히 출현하는 어휘를 파악하였다. 이를 위해 각 태그 값 별로 구분된 하위코퍼스에 대해 'Antconc' 프로그램을 이용해 코퍼스에 등장하는 어휘를 추출한 후 빈도순으로 소팅하였다.

일차적으로 '긍정' 논조 코퍼스의 1,350문장을 분석하였다. 이 코퍼스에는 총 4,320개의 단어가 등장하였다. 이 단어들의 전체 출현빈도는 18,300번이었다. 이 중 단 한번만 등장하는 빈도1의 단어가 총 2,757개로서, 전체 출현단어의 절반을 넘었다. '긍정' 논조 코퍼스의 빈도수별 출현 단어수는 그림 2와 같다.

이 중 최상위 20개 어휘 및 이의 빈도수를 살펴보면 표 2와 같다.

순위	빈도	어휘
1	491	und
2	428	ist
3	402	das
4	354	die
5	289	der
6	245	gut
7	209	mit
8	191	auch
9	179	ich
10	176	Handy
11	175	es
12	163	dem
13	157	man
14	155	ein
15	149	gut
16	148	sehr
17	143	zu
18	141	auf
19	141	Das
20	134	in

표 2 : '긍정' 논조 코퍼스 최고빈도 어휘

위 표의 결과는 'gut'을 제외하면 관사, 접속사 등과 같은 기능어가 대부분의 최상위 순위를 차지하고 있음을 보여준다. 실제로 상위 20개 어휘 중 내용어라고

할 수 있는 것은 ‘gut’, ‘ich’, ‘Handy’, ‘man’ 등이고 이 중 실제로 문장의 논조를 결정하는 어휘는 ‘gut’ 밖에 없음을 알 수 있다. 따라서 위와 같이 추출된 어휘에서 관사, 접속사, 대명사 등을 제거하는 작업이 필요하다. 이러한 ‘stopword’[5]들을 제거한 후 빈도수 1의 단어도 제거하였다. 이렇게 하여 얻어진 출현빈도 2 이상의 단어갯수는 총 156개이었으며, 이들의 총 출현빈도는 1,509번이었다. 이 단어들이 코퍼스에 등장하는 절대빈도와 이 절대빈도를 총 어휘 누적출현빈도인 18,300으로 나눈 상대빈도를 제시한 표는 다음과 같다.

순위	어휘	절대빈도	상대빈도
1	gut	245	0.013388
2	super	78	0.004262
3	einfach	77	0.004208
4	gross	49	0.002678
5	besser	46	0.002514
6	schnell	44	0.002404
7	top	37	0.002022
8	klasse	35	0.001913
9	zufrieden	34	0.001858
10	toll	33	0.001803
11	klein	30	0.001639
12	best	29	0.001585
13	leicht	25	0.001366
14	echt	24	0.001311
15	schoen	23	0.001257
16	schick	22	0.001202
17	empfehlen	20	0.001093
18	absolut	19	0.001038
19	voll	18	0.000984
20	klar	17	0.000929
21	richtig	17	0.000929
22	begeistern	16	0.000874
23	perfekt	15	0.00082
24	neu	15	0.00082

5 본 연구에서 ‘stopword’로 분류된 품사는 관사, 전치사, 대명사, 조동사, 접속사, 기호, 문장부호 등이다.

25	moeglich	14	0.000765
26	edel	13	0.00071
27	lang	13	0.00071
28	scharf	12	0.000656
29	positiv	12	0.000656
30	genial	12	0.000656

표 3 : 'stopword'를 제거한 '긍정' 논조 고빈도 어휘

표 3에서 상위에 위치하는 단어들을 보면 'gut', 'super', 'einfach' 등과 같이 주로 휴대폰 분야에서 긍정적으로 사용되는 어휘들이 있음을 알 수 있다. 특히 'gut'의 경우 다른 단어들과는 구별될 정도로 월등히 절대빈도와 상대빈도가 높음을 알 수 있었다.

어떤 어휘가 '긍정' 논조의 코퍼스에 자주 등장한다고 해서 무조건 그 어휘가 '긍정' 논조만을 전달한다고 할 수는 없다. 예를 들어 11번째로 자주 등장하는 'klein'의 경우, 휴대폰 분야라 할지라도 항상 긍정적인 논조에만 사용되는지 알기 어렵다. 일반적으로 휴대폰 사용자들은 휴대폰이 작고 휴대성이 좋은 경우 그 휴대폰을 긍정적으로 평하게 되지만, 휴대폰의 액정이 작은 경우에는 그 휴대폰을 부정적으로 평하게 된다. 이러한 어휘에 대해서는 다음 장에서 '부정' 논조 말뭉치의 분석이 끝난 후 언급하게 될 것이다.

위의 분석결과를 품사별로 살펴보면 'empfehlen', 'begeistern' 등을 제외하면 거의 형용사/부사가 긍정논조를 전달하는 어휘임을 알 수 있다. '긍정' 논조 고빈도 50개 어휘를 품사별로 구분하면 다음과 같다.

	형용사/부사	동사	명사
개수	44	3	3
비율	88%	6%	6%

표 4 : '긍정' 논조 고빈도 50개 어휘 품사별 개수

한 어휘가 문장 내에서 어떤 의미로 사용되었는지를 알기 위해서는 그 어휘를 포함하는 앞, 뒤의 단어, 즉, 바이그램을 살펴보는 것이 도움이 된다. 예를 들어 위의 표에서 3번째로 빈번히 사용된 'einfach'의 경우 '사용법이 간단하다' 등의 의미로 사용된 것인지, 다른 형용사를 강조하는 부사로 사용된 것인지 단순 빈도

정보만으로는 알기 어렵다. 이를 위해 ‘Antconc’ 프로그램으로 코퍼스에서 바이그램을 추출하고, 바이그램 중 ‘stopword’로만 구성되어 있는 것들은 제외한 후 빈도순으로 정렬하였다.

전체 코퍼스에서 총 12,830개의 바이그램이 추출되었고, 이 중 ‘stopword’로만 구성된 바이그램을 제외한 나머지 바이그램 중 빈도 2 이상의 총 개수는 140개였다. 이 중 상위 20개 바이그램은 다음과 같다.

순위	빈도	바이그램
1	62	sehr gut
2	16	besser als
3	16	sein gut
4	13	sein einfach
5	12	sehr zufrieden
6	10	kein Problem
7	8	zu bedienen
8	8	zufrieden damit
9	8	sein top
10	7	beste Handy
11	7	gute Bilder
12	6	ganz gut
13	6	gut verarbeitet
14	6	sein super
15	6	verfuegt ueber
16	6	klasse Handy
17	6	sein zufrieden
18	6	stechen scharf
19	5	gut fuer
20	5	super aus

표 5 : ‘긍정’ 논조 고빈도 바이그램

바이그램 추출결과를 살펴보면 ‘긍정’ 논조를 전달하는 개별 어휘들이 어떻게 사용되는지 잘 보여준다. 예를 들어, ‘einfach’의 경우 13회나 ‘sein’ 동사와 결합하여, 사용법 등이 간단하다는 긍정적인 논조를 전달하고 있음을 알 수 있다. 또한 ‘einfach’는 ‘einfach genial’, ‘einfach klasse’, ‘einfach super’ 등과 같이 긍정논조의 형용사들과 함께 사용되어 강조의 역할을 함을 알 수 있었다.[6]

6 본 실험에서 ‘einfach genial’, ‘einfach klasse’, ‘einfach super’는 모두 각 3회씩 사용되었다.

이렇게 추출된 바이그램은 '긍정' 논조에서 자주 사용되는 어휘의 쓰임을 보여주므로, 논조자동분석 시스템에서 매우 유용한 리소스로 사용될 수 있다. 이러한 바이그램이 많이 출현하면 해당 문장이 긍정적인 논조를 담고 있다고 단순 어휘의 출현에 근거할 때보다는 좀 더 정확성 높게 추정할 수 있기 때문이다.[7] 그러나 이를 위해서는 현재 구축한 코퍼스를 대규모로 확장하여 바이그램을 추출할 필요가 있다.

4.2 '부정' 논조 코퍼스

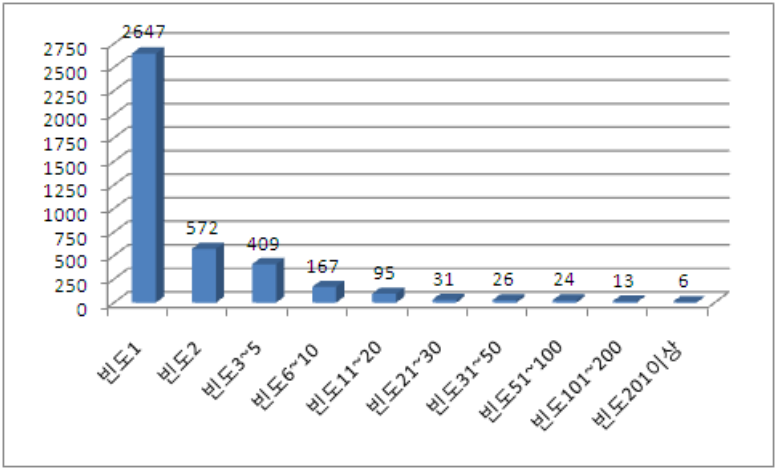

그림 3 : '부정' 논조 코퍼스 출현 빈도수별 어휘수

부정논조 코퍼스의 1,052 문장에 대해서도 '긍정' 논조 코퍼스와 동일한 실험을 수행하였다. 'Antconc' 프로그램을 이용하여 출현어휘를 빈도순으로 추출한 후, 'stopword'를 제거하였다.[8] 'stopword'를 포함한 총 출현어휘의 개수는 3,988

7 특정 논조에 자주 사용되는 어휘 혹은 바이그램이 등장했다고 해서 무조건 해당 문장을 특정논조로 자동 분류하는 방식(bag of words 방식)에 대해서는 많은 문제점이 지적되고 있으나, 논조자동분석 알고리즘 자체는 본 논문의 연구목표가 아니므로, 여기에서는 이에 대한 논의를 차후 연구로 미루기로 한다.

8 'nicht'와 같은 부정어도 기능어에 속하지만, 부정어는 문장의미의 극성을 바꿀 수 있기 때문에, 논조판단에 있어서는 결정적인 역할을 하는 어휘이다. 따라서 이러한 단어들은 'stopword'로 분류하지 않았다.

개였으며, 이들의 전체출현빈도는 14,533번이었다. 이 중 단 1회만 출현하는 어휘는 2,647개였으며, 2회 이상 출현하는 어휘의 개수는 1,341개였다. 단 1회만 사용된 어휘의 비율은 '긍정' 논조의 경우보다 많았다. (63.8% vs. 66.4%) '부정' 논조 코퍼스의 빈도수별 출현단어수는 그림3과 같다.

이 중 최상위 20개 어휘 및 이의 빈도수를 살펴보면 표 6과 같다.

순위	빈도	어휘
1	333	die
2	314	das
3	291	ist
4	286	nicht
5	276	und
6	207	ich
7	198	der
8	155	man
9	142	es
10	142	zu
11	138	mit
12	117	auf
13	116	nur
14	113	Handy
15	111	sich
16	109	ein
17	107	den
18	104	dem
19	101	aber
20	95	auch

표 6: '**부정**' **논조 코퍼스 최고빈도 어휘**

위와 같이 추출된 어휘에서 '긍정' 논조의 실험에서와 마찬가지로 'stopword'를 제거하고 얻어진 어휘 중 빈도 2 이상의 어휘는 총 77개였다. 이 어휘들의 전체출현 빈도는 409였다.

본 실험에서 활용된 '부정' 논조 코퍼스의 규모가 '긍정' 논조 코퍼스보다 작았음을 고려하더라도, '부정' 논조 코퍼스에서 사용된 빈도 2 이상의 어휘개수 및 그 출현빈도는 '긍정' 논조 코퍼스보다 훨씬 작았다는 점이 특징적이라 할 수 있다. 표 7을 통해 알 수 있듯이, 독일어 블로그나 사용후기 등에서는 부정적인

논조의 글에서 긍정적인 논조의 경우보다 자주 반복하여 사용되는 어휘가 드물고 따라서 어휘의 종류가 다양하게 사용되는 것으로 드러났다.

	빈도2 이상 어휘수	출현빈도	문장당 빈도 2이상 어휘수	코퍼스문장개수 (단어수)
긍정	156	1,509	0.116	1,350(17,967)
부정	77	409	0.073	1,062(14,509)

표 7 : 논조별 빈도2이상 출현 어휘수

표 8은 상위 30개 어휘의 절대출현 빈도 및 상대빈도를 보여준다.

순위	단어	절대빈도	상대빈도
1	schlecht	45	0.003096
2	leider	43	0.002959
3	Problem	27	0.001858
4	langsam	15	0.001032
5	enttaeuschen	14	0.000963
6	schwer	11	0.000757
7	negativ	11	0.000757
8	klein	10	0.000688
9	alt	10	0.000688
10	Nachteil	8	0.00055
11	leer	8	0.00055
12	schade	8	0.00055
13	nervig	7	0.000482
14	Kritikpunkt	6	0.000413
15	stoeren	6	0.000413
16	kurz	6	0.000413
17	schwarz	6	0.000413
18	scheisse	5	0.000344
19	schwach	5	0.000344
20	garnicht	5	0.000344
21	aergern	5	0.000344
22	Schrott	5	0.000344
23	beschraenken	4	0.000275
24	kompliziert	4	0.000275
25	Manko	4	0.000275

26	unbrauchbar	4	0.000275
27	unzufrieden	4	0.000275
28	Katastrophe	4	0.000275
29	langweilig	4	0.000275
30	wenig	4	0.000275

표 8 : 'stopword'를 제거한 '부정' 논조 고빈도 어휘

4.1.에서 언급하였던 'klein'의 경우, 예상대로 '부정' 논조에서도 10회나 사용됨을 볼 수 있다. '긍정' 논조 코퍼스에서 빈출하는 최상위 20개 어휘의 경우, 'klein'을 제외하고는 단 하나의 어휘도 '부정' 코퍼스에 등장하지 않았다. 반면 '부정' 코퍼스에서 빈출하는 최상위 20개 어휘의 경우, 'klein'과 'kurz'는 '긍정' 코퍼스에서도 각각 30회와 3회 출현하였다.

본 연구의 실험을 통해서 본 결과, '긍정' 코퍼스의 고빈도 어휘들은 'klein'을 제외하면 거의 모두 휴대폰 분야에서 '긍정' 논조를 전달하는 어휘라고 볼 수 있으며, '부정' 코퍼스의 고빈도 어휘들은 'klein'과 'kurz'를 제외한다면, 거의 모두 '부정' 논조를 전달하는 어휘라고 볼 수 있을 것이다.

물론 본 연구에서 다루지 못한 '혼합' 논조의 문장에서도 이러한 강한 긍정/부정 성향의 어휘가 나올 수도 있지만, 이러한 경우에도 이 어휘들은 각각 긍정/부정의 논조를 전달할 것으로 예상된다. 이에 대한 실험은 추후 연구로 미루기로 하고, 본 논문에서는 다루지 않는다.

'부정' 논조의 코퍼스에 자주 등장하는 단어들은 품사의 측면에서 볼 때 '긍정'의 논조와 마찬가지로 형용사/부사가 많이 등장하지만 그 비중은 '긍정' 논조의 경우보다는 적음을 알 수 있다 (88% vs. 52%). '부정' 논조 코퍼스에서는 예를 들어 'enttäuschen', 'stören' 등과 같은 동사 및 'Problem', 'Nachteil', 'Kritikpunkt' 등과 같은 명사도 상대적으로 많이 등장한다.(표 9 참조)

논조에 따른 이와 같은 출현어휘 품사의 분포양상도 향후 더 큰 코퍼스에 기반하여 검증되어야 하겠지만, 논조 자동분류를 위해 충분히 활용될 여지가 있을 것으로 보인다.

	형용사/부사	동사	명사
개수	26	9	15
비율	52%	18%	30%

표 9 : '부정' 논조 고빈도 50개 어휘 품사별 개수

'부정' 논조 코퍼스에서도 바이그램을 추출하였다. 추출된 전체 바이그램의 개수는 10,890개였다. '긍정' 논조의 경우와 마찬가지로 'stopword'를 제거한 후 빈도2 이상의 바이그램을 추출한 결과 단 37개의 바이그램만 추출되었다. 표 7에서 확인된 바와 같이 '부정' 논조의 텍스트에서는 반복적으로 사용되는 어휘가 상대적으로 적고, 이에 따라 반복적으로 사용되는 바이그램도 작은 것으로 분석된다. 부정논조의 상위 20개 바이그램을 소개하면 다음과 같다.

순위	빈도	바이그램
1	4	nicht moeglich
2	4	Nie wieder
3	4	sehr langsam
4	3	kann nicht
5	3	Nie mehr
6	3	nur schlecht
7	3	nur telefonieren
8	3	sehr enttaeuscht
9	3	so schlecht
10	2	beschraenkt sich
11	2	etwas enttaeuscht
12	2	FINGER WEG
13	2	groesste Problem
14	2	ist grottenschlecht
15	2	ist schlecht
16	2	keine Kamera
17	2	Nachteil des
18	2	Nachteil ist
19	2	nicht automatisch
20	2	nicht besonders

표 10 : '부정' 논조 고빈도 바이그램

5. 결론

본 논문에서는 휴대폰 분야의 사용자 후기 혹은 블로그 등의 긍정/부정 논조 문장에 출현하는 어휘의 특성을 코퍼스언어학 관점에서 살펴보았다. 이를 위하여 긍정/부정 등의 정보가 부착된 소규모 코퍼스를 구축하였으며, 'Antconc'라는 프로그램을 활용하여 출현 어휘 및 바이그램의 빈도수를 조사하였다.

'긍정' 논조의 텍스트에서는 주로 형용사가 많이 사용되었으며, 대부분 긍정적 의미를 지니는 형용사가 사용되었으나, 'klein'과 같은 상대적 개념의 형용사의 경우는 '부정'의 논조에도 상대적으로 많이 출현하였으므로 이 단어를 적어도 휴대폰 분야에서는 '긍정' 논조의 단어로 분류하기는 어려움을 보았다. 그러나 유사한 기능을 하는 'groß'의 경우는 '긍정' 논조의 코퍼스에서는 빈번하게 사용된 반면, '부정' 논조의 코퍼스에서는 한 번도 사용되지 않음을 볼 수 있었다. 물론 이는 좀 더 큰 규모의 코퍼스를 통한 실험으로 검증되어야 할 것이다.

'부정' 논조의 텍스트에서도 형용사가 많이 사용되었으나, '긍정' 논조의 텍스트와 비교해서는 상대적으로 명사와 동사가 많이 사용되었다. 또한 반복적으로 사용되는 어휘의 수도 '긍정' 논조보다는 상대적으로 적음을 알 수 있었다. 이를 통해 사용자들은 '부정' 논조의 글을 쓸 때에는 좀 더 구체적인 문제점을 기술한다는 사실을 알 수 있었다.

이 연구의 결과물이 자동 논조분석시스템 개발에 바로 사용될 수 있기 위해서는 좀 더 큰 규모의 코퍼스에 기반하여 어휘를 추출할 필요가 있다. 관련된 앞으로의 연구로는 도메인의 변화에 따른 긍/부정 어휘양상의 변화 및 도메인 비의존적인 '긍/부정' 어휘에 대한 연구이며, 이에 기반한 자동분석 방법론의 개발이다.

참고문헌

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.

Aman, S. & S. Szpakowicz (2008) Using Roget's Thesaurus for Fine-grained Emotion

Recognition, Proceedings of IJCNLP 2008, 312-318.

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

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

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

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

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. & 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.

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

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

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

Morinaga, S., Yamanishi, K., Tateishi, K. & T. Fukushima (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., G.J. Suci & P.H. Tannenbaum (1971): The Measurement of Meaning. University of Illinois Press.

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.

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, 2002.

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

Riloff, E., Wiebe, J. & T. Wilson (2003) "Learning subjective nouns using extraction pattern bootstrapping", CONLL-03, 25-32.

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

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.

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

Yi, J. & W. Niblack (2005) "Sentiment Mining in WebFountain", Proceedings of the 21st International Conference on Data Engineering, 1073-1083.

Zusammenfassung

Untersuchung zur Wortfrequenz in positiven/negativen Texten für eine automatische Sentimentanalyse

HONG Munpyo (Sungkyunkwan Uni)

Die vorliegende Arbeit beschäftigt sich mit der Frequenz und Charakteristik der Wörter, die in einem positiven/negativen Text vorkommen. Dafür wurde ein Sentiment-getaggtes Korpus in der 'Handy'-Domäne erstellt. Anhand des 'Antconc' Programs wurden die Frequenz der Wörter bzw. Bigram ermittelt.

In 'positiven' Texten wurden meistens solche Adjektive verwendet, die positive

Stimmungen überbringen. Bei dem Fall ‘klein’, das einen hohen Rang in der Frequenzliste der positiven Gruppe besetzt, konnte jedoch nicht behauptet werden, dass dieses Wort unbedingt zur positiven Wortgruppe gehört, weil es auch in negativen Texten relativ häufig vorkommt.

Im Gegensatz dazu wurde das Wort ‘groß’, das eine ähnliche Funktion wie ‘klein’ zu haben scheint, nur in positiven Texten gefunden. Allerdings muss es noch mit größerem Korpus bestätigt werden.

In ‘negativen’ Texten wurden auch Adjektive am meisten gebraucht, aber Substantive und Verben wurden im Vergleich zu den positiven Korpus viel mehr benutzt. Die Anzahl der wiederholt benutzten Wörter war auch geringer als bei den positiven Texten. Man scheint bei der Verfassung der negativen Texten eine reichere Wahl der Wörter zu treffen.

Damit das Ergebnis der vorliegenden Arbeit direkt für die Entwicklung eines automatischen Sentimentanalyse Systems herangezogen werden kann, muss es mit größerem Korpus bestätigt werden.

주제어: 논조분석, 의견마이닝, 단어빈도, 논조부착코퍼스

Schlüsselbegriffe: Sentimentanalyse, Opinion Mining, Wortfrequenz, Sentiment-tagged Corpus

필자 이메일 주소: skkhmp@skku.edu

투고일: 2009. 6. 30. / 심사일: 2009. 8. 10. / 심사완료일: 2009. 9. 6.