독일어 트위터 문장의 규칙기반 논조분석 방안 연구

홍문표 (성균관대학교)

1. 들어가는 말

논조분석 Sentimentanalyse은 텍스트와 같은 비정형 데이터 unstrukturierte Daten로부터 텍스트의 주제 Topik, 주제에 대한 글쓴이의 호/불호 등과 같은 논조 Sentiment, 주제의 특질 Aspekt[1], 논조를 드러낸 시점 등에 대해 구조화된 데이터 strukturierte Daten를 추출하는 것을 말한다. 일반적으로 데이터 마이닝 Data Mining 분야에서 사용하는 '비정형 데이터'라는 용어는 데이터베이스의 각 필드의 값으로 들어갈 수 있는 정보들이 특정한 포맷에 따라 기술되어 있지 않은 데이터를 의미하고, 주로 텍스트를 의미한다. 이에 반해 '구조화된 데이터'라는 용어는 미리 정의된 데이터베이스의 필드에 알맞은 값이 채워진 데이터를 말한다.

다음에서 (1)과 (2)는 비정형 데이터의 예, (1-1)과 (2-1)은 각각 (1)과 (2)로부터 추출한 구조화된 데이터의 예이다.

(1) LG Optimus 7: von vornherein kein Kandidat. Absolut hässliches Design.
(1-1) Topik: LG Optimus 7, Hardware: _, Software: _, Design: Negativ
(2) Gingerbread Keyboard. Die beste Tastatur meiner Meinung nach.
(2-1) Topik: _ , Hardware: Positiv, Software: _, Design: _

(1)의 예문은 주제 ('LG Optimus 7')와 특질 ('Design')에 대한 언급이 되어 있으나 정형화된 포맷을 가지고 있지 못한 반면, (1-1)은 정형화된 포맷을 가

1) 논조분석 분야에서 사용하는 '특질 Aspekt'이라는 용어는 대상이 가지는 여러 가지 성질을 의미한다. 예를 들어 휴대폰의 특질로는 '통화품질', '카메라', '배터리', '디자인' 등을 들 수 있다. 이를 영미권 문헌에서는 'aspect'라고 표기하고 있으며, 현재까지는 한국어로 통일된 번역 용어가 없으므로 본 논문에서는 '특질'이라고 표기하고, 언어학 분야에서 사용하는 시제 개념과 관련된 '양상'과는 다른 개념임을 밝힌다.

지고 있고, 각 특질에 대한 값이 명시적으로 나와 있지 않은 경우에는 '_' 기호를, 명시적으로 나타난 경우에는 그 값을 콜론 기호 다음에 표기하였다.

비정형 데이터로부터 구조화된 데이터를 추출하는 방법은 크게 기계학습기반의 방법론과 사전기반의 방법론으로 나눌 수 있다. 이 중 기계학습기반의 방법론은 논조분석을 위한 방법론의 대표적인 패러다임이라고 볼 수 있다. Pang et al. (2002)의 연구를 시작으로 활발히 논의된 기계학습기반의 논조분석은 영어와 독일어를 비롯하여 많은 언어에 고루 성공적으로 적용될 수 있음이 다수의 연구 논문을 통해 입증되었다. 기계학습기반의 방법론은 레이블이 부착된 학습데이터에 존재하는 패턴을 통계알고리즘을 적용해 학습하여 분류기를 구현한 후 새로운 데이터의 클래스를 분류하는 기법이다.

이에 반해 사전기반의 방법론은 별도의 학습데이터를 사용하지 않고, 기구축된 감정사전의 정보를 이용하여 문장의 긍정성, 부정성을 분류하는 방법론이다. 물론 감정사전을 어떻게 이용하느냐에 따라 감정사전의 엔트리가 기계학습을 위한 자질로 사용되는 경우에는 기계학습기반의 방법론으로 볼 수도 있으나, 본 연구에서 사용하는 사전기반의 방법론이라는 용어는 기계학습을 사용하지 않고 입력문장의 간단한 형태, 통사적 분석 후 감정사전의 정보를 활용하여 문장의 긍정성, 부정성을 분류하는 방법론을 말한다.

본 연구의 목적은 일반 뉴스 텍스트나 전문분야 텍스트 등과는 현격히 다른 성격을 가지고 있는 트위터 텍스트의 논조분석 방법론을 모색하는 것이다. 본 연구에서는 일반 텍스트에 등장하는 문장들에 비해 현저히 짧고, 문어체와 구어체의 중간 성격을 갖고, 의도적이건 무의식 중이건 비문이 자주 사용되며, 이모티콘과 비속어 등의 사용이 빈번한 트위터 텍스트의 논조를 분석하기 위한 방법론을 모색해 보고자 한다.

특히 논조분석을 위해 가장 일반적으로 사용되는 기계학습기반의 방법론을 트위터 데이터에 적용할 경우 일반 텍스트와 비교하여 어떠한 결과가 나타나며 그 원인은 무엇인지를 알아본다. 또한 Klenner et al. (2009) 등의 연구에서 주장하는 바와 같이 트위터 등과 같은 사회관계망 텍스트 레지스터의 논조분석을 위해서는 감정사전에 기반한 규칙기반 논조분석이 더 적합하다는 가설도 검증해볼 것이다. 그리고 규칙기반 논조분석이 성공적으로 적용되기 위해 어떤 점이 보완되어야 하는지에 대한 논의도 이루어진다.

이를 위해 본 논문은 다음과 같이 구성되어 있다. 2절에서는 논조분석을 위한 기존의 방법론을 소개하며 각 방법론의 장단점에 대해 논의한다. 3절에서는 트위터와 같은 종류의 텍스트 레지스터에 대한 언어학적 논의를 다룬다. 특히 논조분석의 관점에서 본 언어학적 특징에 대해 주로 논의한다. 4절에서는 본 연구의 가설을 검증하기 위한 실험을 수행하며, 이의 결과를 다룬다. 끝으로 5절에서는 본 연구의 기여에 대해 논의하며 향후 연구방향에 대해 논의한다.

2. 논조분석 방법론

2.1. 기계학습기반 논조분석

기계학습 maschinelles Lernen은 레이블 label이 부착된 학습데이터 Lerndaten에 존재하는 패턴을 찾아내어 새로운 데이터의 레이블을 예측하는 인공지능 분야의 이론이다. 기계학습의 적용분야는 무궁무진하지만 최초의 적용분야는 패턴인식이다. 예를 들어 농어와 연어를 자동으로 분류하는 시스템을 개발한다고 할 때, 이 시스템은 농어와 연어의 특징에 대해 먼저 학습이 되어야 한다. 농어와 연어를 구별하는 특징에는 몸통의 길이, 지느러미의 길이, 비늘의 크기 등을 들 수 있다. 만약 우리가 많은 수의 농어와 연어에 대해 이와 같은 특징에 대한 값을 가지고 있다면 여기에 지지벡터머신 Support Vector Machine (SVM) 등과 같은 알고리즘을 사용하여 자동분류기를 개발할 수 있다. 이를 통해 우리는 새로 잡힌 물고기가 농어인지 연어인지를 구별할 수 있을 것이다.

이러한 원리를 그대로 언어데이터에 적용한다면 다음과 같을 것이다. 우리에게 긍정 레이블이 부착된 문장과 부정 레이블이 부착된 문장이 대규모로 주어져 있다면 우리는 이로부터 각 카테고리 (‘긍정’, ‘부정’)를 결정짓는 특징들을 찾아낼 수 있을 것이다. 물고기의 종류를 구별하는 특징, 예를 들어 몸통의 길이, 지느러미의 길이, 비늘의 크기 등과는 다르게 언어데이터에서 이러한 역할은 결국 각 문장에 어떠한 단어가 쓰였느냐 또는 어떠한 문장 구조가

사용 되었는가 등과 같은 정보가 맡게 될 것이다.

기계학습을 이용한 논조분석 방법 연구의 초기에는 이러한 특징으로 학습 데이터에 사용된 유니그램 unigram과 바이그램 bigram이 주로 사용되었다. 유니그램은 학습데이터에 사용된 단어와 문장부호 등과 같은 공백으로 구별되는 모든 단위들을 포함하며, 바이그램은 두 개의 연속된 유니그램을 지칭한다. Pang et al. (2002)의 연구에서는 유니그램만을 특징으로 학습한 논조분석기가 가장 높은 성능을 보이는 것으로 보고되었으나, 영어에 대한 Pang/Lee (2004), Pak et al. (2010)의 연구, 독일어에 대한 홍문표 (2014a)의 연구 등에서 보인 바와 같이 일반적으로는 유니그램과 바이그램을 모두 사용할 경우 가장 높은 정확률을 보이는 것으로 알려져 있다.

기계학습기반의 방법론이 갖는 가장 큰 문제점 중 잘 알려진 것은 문장의 구조적 정보를 반영하기가 어렵다는 점이다. 이 방법론에서는 기계학습을 위한 특징으로 유니그램과 바이그램과 같은 단어들의 출현여부만을 사용하는 경우가 많으므로 문장의 구조에 따른 문장 의미의 변화를 고려하기가 단순하지 않다. 예를 들어 'schön'과 같은 긍정의미의 단어도 어떠한 구조상에 나타나느냐에 따라 문장의 극성 Polarität이 달라질 수 있다. 예를 들어 'nicht'와 함께 사용되면 부정의 의미로 변화되고, 그렇지 않더라도 비현실 양상의미를 전달하는 접속법 형태 'wäre'와 함께 사용되어도 부정의 의미를 나타낼 수 있다.

이렇게 구조정보가 문장의미의 긍정성, 부정성을 변화시킬 수 있다는 점을 반영하여 Wilson et al. (2005)의 연구에서는 유니그램과 바이그램 이외에도 수식-피수식구조, 양상동사구문의 사용여부, 능동-수동문장 구조 등과 같은 문장구조에 관한 정보를 특징으로 사용하여 기계학습시 문장의 구조를 간접적으로 반영하고자 하였다.

Josh/Penstein-Rose (2009)의 연구에서는 이보다 더 나아가 문장 내의 의존구조를 일반화하여 기계학습의 특징으로 사용하는 방법을 제안하였다. 이들의 방법론에서는 학습데이터 문장을 의존문법규칙에 따라 구조분석한 후 핵심어 노드와 자녀 노드들의 관계를 일반화하는 방식으로 구문구조를 기계학습에 반영하였다.

Pak et al. (2010)의 연구는 트위터의 논조 분석을 위해 자동으로 코퍼스를 수집하는 방법과 코퍼스의 언어적인 분석을 통해 나타난 현상에 대해 다루었

다. 이 연구에서는 트라이그램과 같은 고차 n-그램으로 문장의 구조적 정보를 반영하는 한편, 유니그램을 사용하여 높은 커버리지 coverage를 얻고자 하였다. 논조 분류기는 'Naive Bayes'와 'SVM', 'CRF'를 사용하여 구현하였으며 이 중 가장 좋은 결과는 'Naive Bayes'에 기반한 분류기에서 얻어졌다.

실험결과 바이그램을 자질로 사용한 실험이 제일 좋은 결과를 보였다. 이는 바이그램이 커버리지와 문장 구조정보 반영을 위한 복잡도 사이의 균형을 잘 맞추기 때문이라고 설명된다. 다음으로 부정어를 부착시킨 n-그램으로 실험하였다. 부정어가 부착된 n-그램이 부착시키지 않았을 경우의 실험보다 더 높은 정확도를 얻을 수 있었다.

Kouloumpis et al. (2011)은 기존 연구들에서 논조 분석에 유용하다고 밝혀졌던 감성 사전과 품사 태그와 같은 자질들이 트위터에서도 동일하게 유용한지에 대해 알아보고자 하였다. 기계학습을 위한 자질로는 유니그램과 바이그램, 'MPQA' 감성사전에 등록되어 있는 정보, 품사, 'Internet Lingo Dictionary'와 그 외 다양한 인터넷 은어 사전들에서 추출한 마이크로블로깅 도메인에 특화된 언어들(이모티콘, 약어 등)을 활용하였다.

실험은 자질들을 각기 다른 4개의 조합으로 묶어 진행하였는데 베이스라인은 n-그램만을 자질로 사용하여 개발한 분류기였다. 4개의 조합은 먼저, n-그램과 사전 자질들(n-gram+lex), n-그램과 품사 자질들(n-gram+POS), n-그램과 사전 자질, 마이크로블로깅에 특화된 자질들(n-grams+lex+twit), 마지막으로 모든 자질을 활용한 조합이 있다. 이 중 'n-grams+lex+twit' 조합이 사용된 분류기가 가장 높은 성능을 보였다. 훈련 데이터에 따라서도 결과에 차이가 존재했다. 해시태그 데이터만 훈련 데이터로 사용되었을 때보다 해시태그 데이터와 이모티콘 데이터가 같이 훈련 데이터로 사용되었을 때 대체적으로 높은 F-측정값을 기록하였다. 그러나 'n-grams+lex+twit' 조합에서는 해시태그 데이터만 사용하였을 때 더 높은 F-측정값을 얻을 수 있었다.

2.2. 사전기반 논조분석

논조분석을 위해 기계학습 방법론을 사용하는 데에는 몇 가지 문제점들이 알려져 있다. 첫째로 기계학습을 위한 학습데이터의 도메인 편향성 문제이다.

예를 들어 스마트폰 분야의 학습데이터로 학습된 분류기로 정치사회 분야의 논조분석을 수행할 경우 정확률이 하락하는 것을 예측해볼 수 있다.

둘째로는 문장구조정보의 반영이 직관적이지 못하다는 점이다. 앞서 2.1절에서 Wilson et al. (2005)의 연구와 Josh/Penstein-Rose (2009) 등의 연구에 대해 소개하면서 문장의 구조정보를 기계학습에 반영하기 위한 노력을 일부 소개하였으나, 이들의 시도는 문장의 구조분석, 의미분석을 수행하는 전통적인 자연언어처리 방법론과 비교해보면 직관성이 떨어진다고 할 수 있다.

마지막으로는 트위터와 같은 짧은 문장에 기계학습 알고리즘을 적용할 경우의 문제이다. 트위터는 140자 이내라는 공간적 제약이 존재하므로 대부분의 문장들을 매우 간략하게 작성하는 경우가 많다. 기계학습을 위한 학습데이터는 대개 n차원의 자질벡터 Vector로 이루어지는데 일반적으로 적게는 수백개에서 많게는 수천개의 자질을 사용하게 된다. 그러나 입력문장이 매우 짧을 경우 n차원의 자질벡터에 모두 0 값만이 할당되어 변별력을 잃을 수 밖에 없고 결국 분류의 정확성은 매우 떨어지게 되는 문제가 있다. 트위터와 같은 사회관계망관련 텍스트 레지스터에 대한 언어학적 분석은 3장에서 좀 더 자세히 다루게 될 것이다.

홍문표 (2014b)의 연구에서는 기계학습 방법론의 도메인 편향성 문제를 해결하기 위해 사전기반의 논조분석 방법을 제안하였다. 이 연구에서 제안한 방법은 ‘German Polarity Clues(GPC)’와 ‘Senti-Wortschatz(SWS)’와 같은 독일어 감정사전의 엔트리를 기계학습을 위한 자질로서 사용하는 것이었다. 이 연구에서는 감정사전의 엔트리로 학습한 논조분석기를 스마트폰 분야와 호텔 분야에 적용할 경우 각 분야별 학습데이터로 학습한 논조분석기의 성능과 비교하여 약 3~7% 정도의 정확률 하락을 보이는 것으로 보고되었다.

그러나 홍문표 (2014b)의 연구는 독일어 감정사전을 논조분석을 위한 기계학습의 자질로 사용하였기 때문에 기존의 기계학습에 기반한 방법론과는 크게 다르지 않다. 이와는 달리 Moilanen/Pulman (2007), Klenner et al. (2009), Taboada et al. (2011) 등의 연구는 기계학습을 사용하지 않고 입력문장에 대한 형태소, 통사 분석 후 감정사전의 정보를 활용하여 문장의 논조를 분석하는 방법을 취하고 있다.

Moilanen/Pulman (2007)은 본 논문에서 제기한 가설과 관련된 연구를 수행

하였다. 이들은 기계학습과 같은 통계기반의 접근법은 입력문장이나 텍스트가 길 경우에는 좋은 성능을 발휘하지만 입력단위가 단어 또는 구 등과 같이 문장의 일부분이어서 짧을 경우에는 정확률이 떨어진다는 점을 지적하였다. 이들은 이러한 문제를 해결하기 위해 몬테규 의미론식의 의미합성계산법을 제안하였다.

이들이 제안하는 논조계산방법은 감정사전의 정보에 기반을 둔다. 즉, 입력문에 등장하는 모든 단어들은 'Senti-Wordnet' 등과 같은 감정사전에서의 값에 따라 긍정과 부정에 해당하는 수치값을 갖는다. 이후 상향식 bottom-up으로 단어와 단어가 결합하면서 더 큰 단위의 논조가 결정된다. 단어와 단어가 결합하여 구를 이룰 경우 의미핵심어 semantischer Kopf 단어가 더 큰 단위의 논조를 결정하는데 영향을 미치게 된다.

예를 들어, 형용사와 명사가 결합하여 명사구[2] 등을 만들어낼 경우 두 품사 중 의미핵심어 역할을 하는 형용사의 논조정보가 명사구의 논조정보를 결정한다는 것이다. 따라서 'schlechte Nachricht'와 같은 경우에는 부정논조의 형용사 'schlecht'와 중립논조의 명사 'Nachricht'가 결합하여 부정논조의 'schlechte Nachricht'가 만들어진다.[3] 이러한 식으로 단어레벨에서 최상위 문장레벨까지 의미(논조)의 합성을 통해 최종적으로는 입력단위의 논조가 계산된다.

그러나 이와 같은 논조계산 방법은 입력문에 대한 정확한 형태소분석, 구조분석, 의미분석과정을 전제로 하기 때문에 만약 이 단계를 거치면서 어디에선가라도 오류가 발생할 경우 최종적인 결과값이 잘못될 가능성이 매우 높아진다. 따라서 이러한 몬테규 의미론 방식의 의미합성 방법론은 트위터 문장의 언어학적 특성을 고려해봤을 때 트위터 텍스트에 적용하기에는 부적합할 것

2) 엄밀히 말하면 N-bar 레벨이 되겠지만, 여기서는 의미적인 면에만 집중하기 위해 명사구라고 지칭함

3) 긍정의미의 형용사가 중립의미의 명사와 결합하여 항상 긍정의미의 명사구만을 만들어내지는 않는다. 예를 들어 긍정의미의 형용사 'warm'이 중립의미의 명사 'Essen'과 결합하여 'warmes Essen'이 되면 이는 긍정의미를 지닌다고 볼 수 있으나, 다른 중립의미의 명사 'Bier'와 결합하면 'warmes Bier'가 되어 부정의미의 명사구를 이루게 된다. 이와 같은 경우는 명사의 어휘정보를 고려하여 예외적인 규칙으로 처리하여야 한다

으로 판단된다.

3. 논조분석 관점에서 본 독일어 트위터의 언어학적 특성

박신혜 (2012)의 연구는 논조분석의 관점에서 트위터의 언어학적 특성을 분석한 연구이다. 이 연구에 따르면 트위터는 작성용량의 제한으로 인해 단문메시지적인 특징을 갖는 한편, 구어체적인 특징도 가지고 있다. 대표적인 단문메시지적 특징으로 볼 수 있는 것은 이모티콘의 사용이다. 이모티콘의 사용은 Read (2005)의 연구에서 볼 수 있듯이 기계학습을 위한 중요한 자질로도 활용되었다. 박신혜 (2012)는 이모티콘의 사용을 140바이트의 작성 용량 제한으로 인해 작성 내용 등을 강조하기 위한 것으로 보았다.

이모티콘의 사용과 관련하여 한 가지 흥미로운 점은 일반적인 예상과는 달리 이모티콘이 긍정문이나 부정문과 같은 주관적인 의미를 전달하는 문장 이외에 중립적인 의미의 문장들에서도 흔히 사용된다는 점이다. 위 연구의 실험에 따르면 이모티콘이 사용된 문장 코퍼스의 38.8%는 긍정문이었으며, 24.7%는 부정문, 36.4%가 중립문장이었다.[4)]

(3) ich glaub ich hab das blackberry curve 3G :D
(4) Tja ... wenn Du keinen Steeve hast, hast du kein iPhone5! x)
(5) Über 350,000 Leute sind schon auf die IPhone 5 Fanseite reingefallen... Da verdient jemand richtig Geld mit Bannerwerbung ... :)

(3)~(5)의 예문에서 확인할 수 있듯이 트위터에서 이모티콘은 중립 문장에서도 사용되며 이는 가벼운 농담이나 유머러스한 글을 작성할 때 글의 문체를 좀 더 자연스럽게 하는 것으로 박신혜 (2012)의 연구는 보고 있다. 또한 많은 경우 한 이모티콘이 긍정과 부정의 논조에 동시에 사용되는 경우도 관찰되었다.

본 연구에서는 박신혜 (2012)의 연구결과를 받아들여 아래 (6)~(8)의 이모티

4) (3)에서 (5)의 예문은 박신혜 (2012)의 48쪽에서 발췌함

콘을 각각 긍정의미의 이모티콘과 부정의미의 이모티콘, 긍·부정 공통의미의 이모티콘으로 간주하고 4장에서 제시할 트위터 논조분석 방법론에 적용하고자 한다.

(6) 긍정의미 이모티콘
:* :-* +_+ ;* *-* *_* ♥ ♥_♥ :3

(7) 부정의미 이모티콘
._. :`( :-( =`{ X_X :S :| :< x/ >:(>:-(|o| T.T T_T :X

(8) 긍.부정 공통 이모티콘
=D -.-* :`) :} :o ;-) ;d =) x,D :-o (: :( :-) :D :`P ;-) ;o >-> xD :-O (`; :(:/ :d ;) ;/ ;P O.o XD -.- -.-“ :-(:-/ :-D ;(;] ^^ oO :P :) :] :O ;) ;D -_- Oo o.O

Tagg (2009)의 연구와 마찬가지로 박신혜 (2012)의 연구에서도 트위터의 단문메시지적 특성으로 축약형의 빈번한 사용을 들고 있는데, 이 중 논조분석과 관련 있는 특성은 독일어 텍스트에 나타나는 외래어의 축약현상이다. 트위터 등과 같은 온라인 커뮤니티의 발전과 함께 몇몇 영어 축약어는 언어에 상관없이 보편적으로 흔히 사용되는 것으로 보이는데 이 중 대표적인 축약어는 다음 (9)~(14)와 같다. 이들은 모두 감정을 전달하는 효과를 가지므로 논조분석에서 반드시 고려해야할 요소로 보인다.

(9) WTF, wtf (what the fuck)
Smartphone: Nicht genügend Ressourcen. Bitte schließen Sie einige Anwendungen
Ich: Wtf? Ich hab nur Musik an!

(10) LOL (laugh out loud)
na wie gut, dass ich ein BB habe & kein iphone XD... Lol

(11) Sry (Sorry)
sry dass ich dir heute mittag nicht geantwortet habe, aber mein Handy hat irgendwie nicht richtig funktioniert :(

(12) OMG, omg (oh, my god)
omg itunes ist total scheisse. Dewegen würde ich nie mehr ein Apple Produkt kaufen

(13) FTW (for the win)
Android Ftw Jungs. Ihr habt keine Ahnung :D

(14) WTH (what the hell)
WTH, Gingerbread ist ja sowas von schnell! Echt viel schneller als vorher!

트위터는 단문메시지적 성격 이외에도 구어체적 특징을 갖는데 그 중 논조분석과 특히 많은 관계를 갖는 것은 영어 표현의 사용이라고 볼 수 있다. 박신혜 (2012)에 따르면 이 현상은 영어어휘를 그대로 사용하거나 영어의 구 또는 문장을 사용하거나 영어단어를 독일어처럼 사용하는 경우로 더 나눌 수 있다. (15)~(17)의 예문은 각각 영어어휘를 사용한 경우, 영어 구를 사용한 경우, 영어단어의 차용현상에 대한 예문이다.[5)]

(15) Grade spiegelbild bei iTunes geladen! Echt nice! Komm doch mal nach Trier!
(16) Stille, ... mitten im Unterricht klingelt das Handy. Ganze Klasse schreit: What the fuck! Jaja BK bei uns xD
(17) Congratulations! Endlich können wir uns von Handy zu Handy twittern!

위와 같은 현상은 독일어 감정사전에는 영어어휘가 누락되어 있는 경우가 많으므로 독일어 감정사전을 활용하여 논조분석을 시도할 경우 반드시 고려해야 할 현상이다. 다음 장에서는 이상에서 살펴본 트위터의 언어학적 특성을

5) 박신혜 (2012)의 69쪽에서 발췌함

반영한 새로운 논조분석 방법론을 소개한다.

4. 논조자동분석 방법제안 및 실험

4.1. 감정사전

지금까지 각각 2장과 3장에서 논조분석 방법론들의 장단점 및 트위터 문장의 언어학적 특성을 살펴보았다. 이에 기반하여 본 장에서는 감정사전에 기반한 논조분석 방법론을 제안하고자 한다. 이를 위하여 사용한 독일어 감정사전은 'GPC(German Polarity Clues)'이다. 'GPC'는 Waltinger (2010)에 의해 개발되었으며 총 10,141개의 레마 Lemma로 구성되어 있다.[6)]

'GPC'는 영어 감정사전 'Subjectivity Clues'를 독일어로 번역하여 만들어진 사전이다. 'Subjectivity Clues'는 Wiebe et al. (2005)이 개발한 대표적인 영어 감정사전인데 'GPC'의 엔트리들은 영어엔트리들을 자동번역을 통해 독일어로 번역한 결과물이다. 이 사전의 또 하나의 특징은 'nicht schlecht'와 같은 일부 바이그램을 하나의 엔트리로 등록해 놓은 점이다. 'nicht schlecht'는 사전에 수록될 수 있는 하나의 엔트리가 아니지만 논조분석에 큰 영향을 미칠 수 있는 것으로 보고 하나의 엔트리로 간주하였다.

'GPC'와 더불어 또 하나의 대표적인 독일어 감정사전으로 들 수 있는 것은 'SentiWS(Sentiwortschatz)'이며, 이는 Remus et al. (2010)에 의해 소개되었다. 본 연구에서는 'SentiWS'를 사용하지 않고 'GPC'를 사용하였는데, 그 이유는 홍문표 (2014b)의 연구결과 'GPC'를 사용한 경우 논조분석의 정확률이 'SentiWS'를 사용한 경우보다 높았기 때문이다. 'GPC'보다 'SentiWS'의 엔트리 수가 더 많음에도 불구하고 'GPC'가 더 높은 정확률을 보인 이유는 'nicht schlecht' 등과 같이 빈번하게 사용되면서 논조의 결정에 큰 영향을 미치는 바이그램들을 사전 엔트리로 등록해 놓았기 때문일 것으로 홍문표 (2014b)에서

6) 이를 단어의 모든 굴절형태가 들어있는 완전형태 full form로 계산하면 긍정논조 엔트리는 17,628개, 부정논조 엔트리는 19,956개이다

는 보고 있다.

‘GPC’를 트위터 논조분석에 바로 사용할 경우에는 치명적인 문제가 존재한다. 3장에서 살펴본 바와 같이 이모티콘과 비속어, 영어축약어, 일부 영어단어 등은 독일어 트위터에도 흔하게 등장하며 문장의 논조를 결정하는데 매우 중요한 단서가 된다. 그러므로 본 연구에서는 ‘GPC’를 보다 확장하여 적용한다. 기존의 ‘GPC’엔트리에 (6), (7)에서 소개한 긍정, 부정 의미의 이모티콘을 추가하였으며, (9)~(17)에서 제시한 영어 축약어 및 흔히 사용되는 영어단어들, 그리고 일부 독일어 비속어도 추가하였다.

4.2. 논조분석 패턴

본 연구에서 제안하는 논조분석 방법은 입력문에 대해 형태소 분석 및 품사태깅 POS Tagging, 레마화 Lemmatisierung의 과정을 거친 후 논조분석 패턴을 적용하는 것이다. 논조분석 패턴은 다음의 표 1~7에서 볼 수 있는 것처럼 총 22개의 패턴으로 구성되어 있고, 그 성격에 따라 7 종류의 규칙으로 나눠볼 수 있다. 입력문의 품사태깅 결과가 제안하는 패턴에 매칭되면 해당 패턴이 적용되어 부분적인 논조가 결정된다.

예를 들어, 부가어적 형용사(ADJA)와 일반명사(NN)가 연속적으로 출현한 문장에서 만약 부가어적 형용사의 논조가 부정(negative)이고 일반명사의 논조가 긍정(positive)이라면 두 단어의 결합은 부정의 논조가 된다. 따라서 ‘eine unbequeme Wahrheit’라는 입력이 들어오면 부가어적 형용사 ‘unbequem’은 ‘GPC’사전의 정보에 따라 ‘부정’의 논조를 갖고, ‘Wahrheit’는 ‘긍정’의 논조를 갖는데, 최종적으로 ‘eine unbequeme Wahrheit’는 규칙 1의 첫 번째 패턴에 의해 ‘부정’의 논조를 갖게 된다.

규칙 2와 3은 ‘und’로 연결된 등위구문의 분석에 관한 것이다. 규칙 4는 존재구문 ‘es gibt’와 관련된 것이고, 규칙 5는 논조의 강조와 관련된 것이다. 규칙 6과 7은 ‘nicht’와 ‘kein(e)’등과 결합하여 논조가 바뀌는 현상을 다루기 위한 규칙 및 패턴이다.

형용사(ADJA/ADJD)	명사(NN/NE)	논조
negative	positive	negative
negative	negative	negative
negative	neutral	negative
positive	positive	positive
positive	negative	negative
positive	neutral	positive
neutral	negative	negative

<표 1> 규칙 1

명사(NN/NE)	und	명사(NN/NE)	논조
negative		positive	positive
negative		negative	negative
neutral		neutral	neutral

<표 2> 규칙 2

형용사(ADJA/ADJD)	und	형용사(ADJA/ADJD)	논조
negative		positive	positive
negative		negative	negative

<표 3> 규칙 3

Es	gibt	명사(NN/NE)	논조
		positive	positive
		negative	negative
		neutral	neutral

<표 4> 규칙 4

zu	형용사(ADJA/ADJD)	논조
	positive	positive
	negative	negative
	neutral	negative

<표 5> 규칙 5

nicht	형용사(ADJA/ADJD)	논조
	positive	negative
	negative	positive

<표 6> 규칙 6

kein	명사(NN/NE)	논조
	positive	negative
	negative	positive

<표 7> 규칙 7

패턴의 매칭을 위해 입력문장에 대해 'WinTreeTagger'를 활용하여 형태소 분석 후 품사 태깅을 수행하였다. 'WinTreeTagger'는 독일어 코퍼스의 효과적인 분석을 위해 매우 손쉽게 사용할 수 있다는 장점이 있다.[7] 품사태깅을 위한 독일어의 품사 정보로는 슈투트가르트/튀빙엔 태그셋(Stuttgart/Tübinger Tagset)이 사용되었다.

'WinTreeTagger'를 통해 품사태깅된 입력문은 표 1~7의 규칙에 적용되는 경우 논조계산이 수행되었다. 입력문에 대해 품사태깅만 수행하고 구조분석을 수행하지 않은 이유는 트위터의 특성상 비문과 미등록어가 많이 등장하기 때문이다. 비문과 미등록어가 등장하는 경우 구문분석기는 그 성능이 급격히 떨어지기 때문에 오히려 수행하지 않는 것보다 논조분석의 결과가 더 나쁠 수 있다. 그러나 문장 전체에 대한 구조분석은 수행하지 않더라도 입력문에서 명사구의 단위와 동사구의 단위 정도를 묶어줄 수 있는 얕은 구조분석 Shallow parsing은 더 좋은 분석결과를 위해 필요할 것으로 예상된다. 본 연구에서는 얕은 구조분석을 수행하지 않고 이를 향후의 연구과제로 남긴다.

7) 'WinTreeTagger'를 활용한 독일어 코퍼스 분석 방법에 대해서는 이민행 (2015) 참조

4.3. 실험

본 연구에서 제안하는 방법론 및 가설을 검증하고자 실험을 수행하였다.[8] 본 연구에서 확인하고자 했던 가설은 일반 텍스트보다 훨씬 짧고 많은 비문 등이 등장하는 트위터 등과 같은 사회관계망 텍스트 레지스터를 기계학습을 통해 논조분석을 수행할 경우 일반 텍스트와 비교하여 분석 정확률이 떨어질 수 있다는 것이었다. 검증하고자 했던 또 하나의 가설은 트위터 텍스트 레지스터에 대해서는 감정사전에 기반한 규칙기반의 논조분석 방법론이 기계학습 기반 방법론 못지 않은 성능을 보일 수 있다는 것이었다.

실험 코퍼스는 총 1,040개의 트윗 Tweet으로 구성되어 있다. 이 중 540개는 긍정 논조의 트윗이고, 500개는 부정 논조의 트윗이다. 모든 트윗은 스마트폰을 주제로 한 것이다. 총 1,040개의 트윗에는 15,192개의 단어가 사용되어 한 개의 트윗당 평균 9.86개의 단어가 사용되었다. 일반적으로 한 개의 트윗이 하나 이상의 문장 또는 구 등으로 이루어졌으므로 트윗당 단어개수를 문장당 단어개수로 환산할 경우 한 문장에는 9.86개보다 훨씬 작은 수의 단어가 사용되었음을 알 수 있다.[9] 이는 신문기사나 전문가 리뷰 등보다 훨씬 짧은 수의 단어이다.

먼저 실험코퍼스의 유니그램만을 학습자질로 하는 기계학습기반의 논조분류기를 구현한 후 분석의 정확률을 측정하였다. 실험은 웨카 시스템 3.6.10버전으로 수행하였으며, 정확률 평가는 10등분 교차검증(10-fold cross validation)의 방식으로 진행하였다. 분류기의 구현을 위한 기계학습 알고리즘으로는 지지벡터머신(Support Vector Machine)을 사용하였다. 실험결과 기계학습기반의 트위터 논조분류기는 70.03%의 정확률을 보였다.

홍문표 (2014b)의 연구에 따르면 스마트폰 분야의 일반 텍스트[10]에 대해 유

---

8) 본 연구의 실험을 수행하는데 큰 도움을 준 성균관대학교 독어독문학과 석사과정 임승희 학생에게 감사의 마음을 전한다.

9) 본 논문에서 사용한 코퍼스의 문장당 단어개수를 정확하게 계산하지 않은 이유는 트윗을 구성하는 단위가 항상 문장만은 아니고, 구어체적 특성으로 인해 문장의 경계가 분명하지 않기 때문이다. 예를 들어 실험 코퍼스에 등장하는 다음의 트윗 “Echt ey :D Ja die ubertreibt -.- “öh dein Handy war aus WIESO?!” nerv nerv ich bin 18 -.- ”에서도 문장의 단위를 어디까지 보아야 하는지가 명확하지 않다.

니그램만을 학습자질로 사용하고 지지벡터머신 알고리즘을 사용하여 구현한 논조분류기는 78.8%의 분류 정확률을 나타냈다. 동일한 방법론을 일반 텍스트와 트위터 텍스트에 적용할 경우 약 8% 이상의 정확률 차이가 나타나는 것으로 보인다.

이어서 논조분석패턴을 적용한 방법론의 정확률을 측정하였다. 이 방법론에서는 입력문장에 먼저 논조분석패턴을 적용한 후, 적용되는 패턴이 있으면 패턴적용 결과값을 논조로 반영하여 전체 문장의 논조를 계산하였다. 만약 적용되는 패턴이 없다면 문장 내에 등장하는 긍정논조 어휘의 수, 부정논조 어휘의 수를 고려하였다. 예를 들어 어떤 문장 내에 긍정 논조의 어휘가 3개 등장하고 부정 논조의 어휘가 1개 등장한다면 이 문장은 긍정논조의 문장으로 간주하였다. 긍정과 부정논조의 어휘가 하나도 나오지 않거나 긍정과 부정의 논조가 같은 수일 경우에는 디폴트값으로서 문장의 논조를 긍정으로 분류하였다.[11)]

먼저 논조분석패턴을 적용하지 않고 'GPC' 사전만을 사용한 경우의 정확률을 측정하였다. 이 경우의 정확률을 64.77%였다. 이 방법론에서는 입력문장에 'GPC'에 등재되어 있는 긍정, 부정 논조의 어휘가 몇 개 출현하는지만을 고려하여 전체 문장의 논조를 결정하였다.

다음으로는 'GPC' 사전에 박신혜 (2012)의 연구에서 제안한 이모티콘 및 축약어, 영어차용어 등의 엔트리를 추가한 확장사전만을 사용한 방법론의 정확률을 측정하였다. 이 방법의 분석 정확률은 64.87%로 순수 'GPC' 사전 대비 0.1%의 정확률 향상을 보였다.

최종적으로는 'GPC' 사전에 이모티콘 및 축약어, 영어차용어 등을 추가하고 표 1~표 7의 모든 규칙을 적용하여 논조를 결정하는 방법론의 성능을 측정하였다. 최종적으로 이 방법론의 정확률은 67.85%였다. 지금까지의 실험결과를 표로 정리해 보면 다음과 같다.

---

10) 전문가의 스마트폰 리뷰

11) 이 경우 문장의 논조를 중립으로 간주하는 것이 알맞겠으나, 기계학습 시스템과의 형평성 있는 비교를 위하여 본 실험에서는 중립논조를 고려하지 않았다. 이는 기계학습을 위한 학습코퍼스에 오직 긍정과 부정의 클래스만이 존재하기 때문이다.

기계학습	GPC	GPC+이모티콘+영어 차용어 등	GPC+이모티콘+영어차용어 등 + 논조분석패턴
70.03%	64.77%	64.87%	**67.85%**

<표 8> 방법론별 논조분석 정확률 비교

실험결과 본 연구에서 제안하는 논조분석 패턴을 적용한 방법론은 기존의 순수사전기반의 방법론과 비교하여 3.08%의 정확률 향상을 가져왔으며 기계학습 방법론과 비교해서는 정확률이 2.18% 떨어지는 것으로 나타났다. 기계학습기반의 방법론은 도메인별로 다른 학습코퍼스가 필요하지만 사전과 규칙을 활용한 방법론은 도메인에 덜 의존적이라는 점에서 본 연구에서 제안하는 방법론은 트위터의 논조분석을 위하여 기계학습기반 방법론의 대안으로 충분히 경쟁력이 있다는 점을 보였다.

실험대상 1,040개의 트윗에 적용된 각 규칙별 회수는 다음의 표와 같다.

규칙	1	2	3	4	5	6	7
적용 횟수	236	41	2	1	21	21	6

<표 9> 규칙별 적용회수

이 결과에 따르면 형용사 + 명사, 명사 + 등위접속사 + 명사, zu + 형용사, nicht + 형용사의 순으로 규칙이 많이 적용되었다. 규칙이 적용되는 과정에서 발생한 오류는 두 가지 종류가 있었다. 하나는 품사태깅결과의 오류로 인한 것이고, 다른 하나는 패턴적용 과정에서의 오류이다. 품사태깅 오류는 대체적으로 'WinTreeTagger'에 등록되어 있지 않은 미등록어가 등장할 경우이다. 예를 들어 고유명사 'chrome'이 형용사로 태깅되어 'chrome tweedeck'이 형용사 + 명사 패턴에 적용되는 오류가 발생하였다.

또한 본 연구에서 제안한 규칙 및 패턴은 단순히 품사태깅 결과에 적용하도록 설계되어 있으므로, 더 넓은 범위의 명사구나 동사구에는 적용되지 못하는 단점이 있었다. 이를 위해 향후 얕은 구조분석 등을 수행하고 규칙과 패턴을 적용한다면 추가적인 정확률의 향상이 기대된다.

5. 맺는말

본 연구에서는 트위터와 같은 사회관계망 텍스트 레지스터의 언어학적 특성을 반영한 논조분석 방법론을 제안하였다. 기존의 기계학습기반의 논조분석 방법론은 높은 분류 정확도를 보이지만 대용량의 학습코퍼스를 필요로 하고 도메인별로 서로 다른 학습코퍼스가 필요하다는 단점이 있다. 또한 입력문이 트위터에서와 같이 매우 짧은 경우에는 분류 정확도가 일반 텍스트와 비교하여 현저하게 낮아진다는 단점이 있다. 이를 극복하기 위해 본 연구에서는 감정사전과 논조분석규칙을 활용한 방법론을 제안하였다.

본 연구에서 활용한 독일어 감정사전은 'GPC'이지만, 이 사전 또한 트위터와 같은 사회관계망 텍스트 레지스터의 특성을 반영하기에는 한계가 있음을 보였다. 이를 위해 박신혜 (2012) 등의 연구에서 밝혀낸 독일어 트위터 텍스트의 언어학적 특성을 반영하여 'GPC'사전을 확장하였다. 확장한 사전의 긍정, 부정 정보를 보다 효율적으로 반영하기 위해 논조분석규칙 및 각 규칙을 적용하기 위한 패턴을 제안하였다.

제안한 방법론은 기존의 기계학습기반의 논조분석 방법론과 거의 대등한 성능을 보였다. 기계학습기반의 방법론이 도메인 의존적이라는 점을 감안할 때, 이 방법론을 기계학습기반 방법론의 대안으로 충분히 고려할 수 있다는 점을 제시한 것이 본 연구의 성과라고 볼 수 있겠다. 마지막으로 향후에는 얕은 구조분석 방법을 도입하여 패턴적용과정을 좀 더 체계화하고 분석규칙 및 패턴을 추가하고자 한다.

참고문헌

박신혜 (2012): 「독일어 트위터 메시지의 논조분석을 위한 언어학적 특징 연구」. 성균관대학교 독어독문학과 석사논문.

이민행 (2015): 『빅데이터 시대의 언어연구』. 21세기북스.

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

홍문표 (2014a): 바이그램을 활용한 텍스트 논조자동분석. 『독일언어문학』

65, 27-46.

홍문표 (2014b): 독일어 감정사전을 활용한 감성분석. 『독어학』 30집, 173-195.

Joshi, M./Penstein-Rose, C. (2009): Generalizing dependency features for opinion mining. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 313-316.

Kouloumpis, E./Wilson, T./Moore, J. (2011): *Twitter sentiment analysis: The good the bad and the omg!. Icwsm (2011) 11*, 538-541.

Klenner, M./Fahrni, A./Petrakis, S. (2009): PolArt: A robust tool for sentiment analysis. In: *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009)*, 235-238.

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

Pak, A./Paroubek, S. (2010): Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC (2010)*. 1320-1326.

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 - Volume 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.

Read, J. (2005): Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: *Proceedings of the ACL Student Research Workshop*, 43-48.

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

Tagg, Cl. (2009): *A Corpus Linguistics Study of SMS Text Messaging.* Ph.D. Thesis, The University of Birmingham.

Taboada, M./Brooke, J./Tofiloski, M./Voll, K./Stede, M. (2011): Lexicon-based methods for sentiment analysis. *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./Cardie, C. (2005): Annotating Expressions of Opinions and Emotions in Language. *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

Regelbasierte Sentimentanalyse von Twitter anhand eines Sentiment-Lexikons

Hong, Munpyo (Sungkyunkwan Univ.)

In der vorliegenden Arbeit wird eine Methode für die Sentimentanalyse von Twitter vorgestellt. Diese Methode stützt sich auf die Regeln für die Sentimentanalyse und ein Sentimentlexikon. Das Sentimentlexikon, das in dieser Arbeit herangezogen ist, ist das sogenannte ‘German Polarity Clues (GPC)’ Lexikon von Waltinger (2010). Das GPC Lexikon führt positiv und negativ konnotierte Wörter auf.

Die meisten Methoden für die Sentimentanalyse ziehen ein statistisches Verfahren heran, so dass dafür immer große Lerndaten benötigt werden. Diese Methoden zeigen zwar die beste Performanz in der Sentimentanalyse aber sind zu domänen-abhängig. Das heißt, für jede neue Domäne werden neue Lerndaten benötigt, was sehr zeit- und kosten-intensiv ist, da die Lerndaten meistens handannotiert sind. Ein anderes Problem dieses Ansatzes ist, dass wenn ein Eingabesatz zu kurz ist, ein maschineller Lern-Algorithmus nicht so erfolgreich funktioniert.

Ein Textregister wie Twitter zeigt andere Charakteristiken als normale Textsorten wie Zeitungen oder wissenschaftliche Arbeiten. Er beinhaltet z.B. viele umgangssprachliche Ausdrücke oder Emoticons und Schimpfwörter oder sogar ungrammatische Satzstrukturen. Im GPC Lexikon werden solche Termini allerdings nicht aufgelistet.

Der vorgeschlagene Ansatz für die Sentimentanalyse ist regelbasiert. In einem Eingabesatz wird nach polaren Wörtern gesucht und Modifikatoren wie Negationen, Abschwächungen und Verstärkungen werden in die Analyse einbezogen.

Das Experiment zeigte, dass der vorgeschlagene Ansatz fast so gut wie der auf maschinelles Lernen basierte Ansatz in der Korrektheit der Klassifizierung (67.85% vs. 70.83%) sein kann. Aber dieser Ansatz kann anders als der auf maschinelles Lernen

basierte Ansatz für jede Domäne generell angewendet werden.

[검색어] 감성분석, 감정사전, GPC, 학습코퍼스, 기계학습, 규칙기반
Sentimentanalyse, Sentiment-Lexikon, GPC, Lernkorpus, Maschinelles Lernen, Regelbasiert

홍문표 110-745
서울 종로구 명륜동 3가 53
성균관대학교 독어독문학과
my1004my@gmail.com

논문 접수일: 2015. 10. 28.
논문 심사일: 2015. 11. 30.
게재 확정일: 2015. 12. 07.