구문분석과 기계학습 기반 하이브리드 텍스트 논조 자동분석

*홍문표†, 신미영‡, 박신혜†, 이형민‡

성균관대학교†, 경북대학교‡

Munpyo Hong, Miyoung Shin, Shinhye Park & Hyungmin Lee. 2010. Hybrid Approach to Sentiment Analysis based on Syntactic Analysis and Machine Learning. *Language and Information 14.2*, 159–181. This paper presents a hybrid approach to the sentiment analysis of online texts. The sentiment of a text refers to the feelings that the author of a text has towards a certain topic. Many existing approaches employ either a pattern-based approach or a machine learning based approach. The former shows relatively high precision in classifying the sentiments, but suffers from the data sparseness problem, i.e. the lack of patterns. The latter approach shows relatively lower precision, but 100% recall. The approach presented in the current work adopts the merits of both approaches. It combines the pattern-based approach with the machine learning based approach, so that the relatively high precision and high recall can be maintained. Our experiment shows that the hybrid approach improves the F-measure score for more than 50% in comparison with the pattern-based approach and for around 1% comparing with the machine learning based approach. The numerical improvement from the machine learning based approach might not seem to be quite encouraging, but the fact that in the current approach not only the sentiment or the polarity information of sentences but also the additional information such as target of sentiments can be classified makes the current approach promising. (**Sungkyunkwan University, Kyungpook National University**)

Key words: Sentiment Analysis, Opinion Mining, Text Sentiment, Pattern Matching, Machine Learning, Support Vector Machine, Polarity, Sentiment

* This work was supported by the Industrial Strategic technology development program, 2009-S-034-01,"Development of Machine Translation Technology for Korean/Chinese/English Spoken Language and Business Documents" funded by the Ministry of Knowledge Economy (MKE, Korea)

† 주저자. 서울특별시 종로구 명륜동 3가 53 성균관대학교 독어독문학과 110-745. Email: skkhmp@skku.edu

‡ 교신저자. 대구광역시 북구 산격동 1370 경북대학교 IT대학 전자공학부 701-702. Email: shinmy@knu.ac.kr

Pattern

1. 서론

본 연구는 패턴매칭 (Pattern Matching)과 기계학습 (Machine Learning) 기술을 활용한 텍스트 논조 자동분석 (Automatic Sentiment Analysis)에 관한 것이다. 텍스트 논조 (Text Sentiment)란 특정 주제 (예를 들어, 특정 회사의 휴대폰)에 대한 글쓴이의 호감/비호감 등과 같은 감정을 말하며, 텍스트 논조 자동분석은 텍스트에 드러난 이러한 감정을 컴퓨터가 자동으로 분석하여 그 결과를 출력하는 것을 뜻한다. 자연언어처리 분야에서는 이러한 연구를 '감성분석' 또는 '의견마이닝' (Opinion Mining) 등의 용어로 부르기도 하지만, 본 연구에서는 '논조분석'이라는 용어를 사용하도록 한다.

단순히 단어의 의미파악이나 문장의 명제적 의미를 파악하는 것을 넘어 주제에 대한 화자 혹은 저자의 긍정적/부정적 감정까지 파악하는 논조분석은 특히 최근 들어 자연언어처리 분야에서 많은 연구와 관심의 대상이 되고 있다. 이는 블로그나 인터넷 게시판을 중심으로 작성되는 특정 제품 또는 기업 등에 대한 글들이 기업의 브랜드 가치를 결정하거나 다른 소비자들의 구매의사를 결정하는데 큰 역할을 하기 때문이다. 따라서 각 기업에서는 이러한 인터넷상의 여론 동향을 신속히 파악하여 대처할 필요가 있다.

본 연구에서는 기존의 논조분석 연구에서 주로 제시되었던 순수 기계학습 방식의 접근법과 순수 언어분석 방식 접근법의 문제점을 지적하고, 이러한 문제를 극복할 수 있는 새로운 방법론을 제시하고자 한다. 새로운 접근법은 순수 기계학습 접근법에서 나타나는 학습데이터 (training data) 구축의 어려움과 낮은 정확율 (precision) 문제 등을 해결하고, 순수 언어분석 접근법의 낮은 재현율 (recall) 문제를 해결할 수 있는 기계학습 방식의 장점과 언어분석 방식의 장점을 모두 취하는 하이브리드 (hybrid) 방식의 접근법이다.

새로운 접근법은 입력문장에 대해 형태소분석과 구조분석을 수행한 후, 문장논조패턴을 적용하여 패턴에 매칭되는 문장에 대해서는 우선적으로 긍정/부정 논조의 분류를 시도한다. 패턴을 사용하여 논조분석을 시도할 경우, 보다 정확한 논조분석이 가능할 뿐만 아니라, 그러한 논조가 어떤 세부 속성에 대한 것인지도 파악할 수 있다는 장점이 있다. 순수 기계학습 기반의 방법론에서는 텍스트나 문장이 긍정 논조를 전달하느냐 혹은 부정 논조를 전달하느냐를 분석하는 것에만 집중한다. 그러나 패턴을 사용할 경우에는 예를 들어, 사용자가 휴대폰의 어떤 속성(통화품질, 배터리수명, 액정해상도 등)에 대해 긍정적인가 혹은 부정적인가를 파악할 수 있는 장점이 있다.

패턴에 매칭되지 않는 문장에 대해서는 'SVM' (Support Vector Machine)을

사용한 기계학습 기반의 분류기 (classifier)를 통한 재분류를 시도한다. 기계학습을 통한 분류는 주제의 특정 속성에 대한 긍/부정 논조를 정확하게 파악하기는 어렵다는 단점이 있으나, 기계학습을 통해 초벌 분류된 문장에 대해 감수 작업을 수행하면 완전 수작업을 통해 분류를 수행하는 것 보다는 약 25% 이상의 시간, 비용 절감효과가 있다.[1] 따라서 본 논문에서 제안하는 접근법은 언어구조분석 기반 방법론이 갖는 장점인 높은 분류의 정확율을 그대로 유지하면서, 이 방법의 문제점인 낮은 재현율은 기계학습을 통해 어느 정도 해결할 수 있다는 장점이 있다.

본 연구의 2장에서는 논조분석과 관련된 기존 연구들을 살펴보도록 한다. 제 3장에서는 논조분석의 관점에서 본 한국어 텍스트의 언어학적 속성들을 다루며, 기계학습 등의 주요 특징 (feature)[2]으로 사용될 수 있는 속성들을 살펴보도록 한다. 또한 본 연구를 위해 구축된 휴대폰 분야 한국어 논조태그부착 말뭉치에 대해서도 언급한다. 4장에서는 언어구조분석에 기반한 패턴기반 논조분석 방법론에 대해 다룬다. 5장에서는 'SVM'을 활용한 기계학습 기반 논조분석 방법론을 소개한다. 6장에서는 본 연구에서 제안하는 하이브리드 접근법을 소개하며 새로운 분류방법의 성능 평가결과를 다룬다. 끝으로 7장에서는 본 연구의 요약 및 향후 연구과제 등에 대해 다루게 된다.

2. 관련연구

논조분석에 관련된 주요 연구들은 크게 두 가지의 주제로 분류될 수 있다. 하나는 논조 분석 연구에 기반이 되는 논조속성이 부착된 어휘의 의미 데이터베이스에 관한 것이다. 대표적인 것으로는 기존의 워드넷 (WordNet)에 감정관련 정보를 추가적으로 부착한 센티워드넷 (SentiWordNet)[3]과 워드넷어펙트 (WordNet Affect)가 있다. (cf. Esuli & Sebastiani (2006), Strapparava & Valitutti (2004)) 센티워드넷은 워드넷 각각의 동의어 집합 신셋 (Synset)에 긍정/부정/객관 3가지로 분류되는 추가 감정레이블을 부착한 것이다. '긍정', '부정', '객관적'과 같은 각 레이블이 갖는 점수의 범위는 0.0~1.0이며, 신셋별로 점수의 총 합은 1.0이다. 워드넷어펙트의 경우도 이와 유사하게 각 신셋에 감정정보를 추가적으로 부착하였다.

자동 논조분석에 있어서의 또 다른 연구는 자동논조분석 방법론에 관한 것이다. 자동논조분석 방법론은 다시 기계학습 기반의 분석방법론과 언어분석 기반의 분석방법론으로 나뉠 수 있다. 기계학습 기반의 분석방법론에 관한 가장 대표적인 연구는 Pang & Vaithyanathan(2002)이다.[4] 이 연구에서는 'Naïve Bayes', 'Maximum

[1] 논조분석시스템을 업무에 활용하고 있는 국내 업체담당자와의 사신(personal communication)

[2] 언어학 분야에서는 'feature'가 주로 '자질' 또는 '속성'으로 번역되나, 기계학습 분야에서는 '특징'으로 번역되어 널리 사용되고 있으므로, 본 논문에서는 기계학습 분야에서 사용되는 용어 'feature'를 '특징'으로 번역하도록 한다.

[3] http://sentiwordnet.isti.cnr.it/

[4] 기계학습을 기반으로 한 자동논조분석에 대한 주요 연구로는 Pang & Vaithyanathan(2002) 이외에

Entropy Classification', 'Support Vector Machines'와 같은 세가지의 대표적 기계학습 알고리즘을 적용하여 문서단위의 논조분석을 수행한다. 이 연구는 기계학습 알고리즘을 적용한 대표적인 연구이지만, 논조분석을 문장단위가 아닌 문서단위로 수행하였다는 점에서 본 연구와는 다른 성격을 갖는다.

Wilson et al. (2005)과 Fahrni & Klenner(2008) 등의 연구에서는 어휘가 갖는 의미의 긍/부정과 같은 극성 (Polarity)을 절대적, 상대적 두 가지로 분류하였다. 이들은 특정 어휘들은 문맥에 상관없이 항상 동일한 논조를 갖는다는 대표 극성 (prior polarity)의 개념과, 동일한 어휘가 갖는 논조 혹은 감정정보가 대상에 따라 달라질 수 있다는 타겟의존적 극성 (Target-specific polarity)[5] 개념의 차이를 논조분석에 사용하였다. 타겟의존적 극성이란 예를 들어 영어의 형용사 'warm'이 'pizza'를 수식할 경우에는 긍정논조의 어휘이지만 'beer'를 수식할 경우에는 부정논조의 어휘로 기능을 하는 것과 같이, 하나의 어휘가 수식하고 있는 타겟이 무엇이냐에 따라 상이한 논조를 나타내는 현상을 말한다.

Turney(2002)의 연구는 비지도학습 방식에 기반한 대표적인 논조자동분석 연구이다. 이 연구에서는 어떤 단어의 의미 성향 혹은 논조가, 이 단어와 'excellent' 사이의 상호 정보 (Mutual Information)에서, 주어진 단어와 'poor'사이의 상호 정보를 뺀 값으로 계산될 수 있음을 보였다.

명재석 et al.(2008)은 한국어 문장의 논조를 언어학적 분석에 기반하여 분류하는 방법론을 소개하였다. 이 연구는 의미분석을 통해 문장의 논조를 분석하려는 시도를 하고 있다. 의미분석결과로 도출된 의미표상에 대해 논조분석용 패턴을 적용함으로써, 비교적 높은 정확도의 논조분석을 수행할 수 있다고 주장한다. 그러나 낮은 재현율 문제가 그대로 드러나고 있으며, 이를 높이기 위한 해결책을 제시하지 못하고 있다.

3. 한국어 텍스트 논조 결정 요인 분석

Fries(2009)에 따르면 감정이 언어수단을 통해 표출되는 양상은 크게 네 가지로 나눌 수 있다. 첫째는 감정이 문장으로 직접 표출되는 형태이다. (1), (2)예문에서는 '기쁘다', '뿌듯하다', '짜증나다'와 같은 감정을 직접적으로 나타내는 어휘에 의해 화자 혹은 저자의 감정이 표출되고 있다.

(1) 울만에 스카이를 대하니 넘 기쁘고~ 뿌듯하다~

(2) 진짜 짜증난다 ...

Alm et al.(2005), Gamon(2004), Hatzivassiloglou & Mackeown (1997), Pang & Lee(2004) 등의 연구를 들 수 있다.

[5] 이는 Fahrni & Klenner (2008)에서 사용되는 용어이며, Wilson et al. (2005)의 논문에서는 같은 개념이 문맥 의존 극성 (Contextual polarity)이라는 용어로 표현된다.

두 번째 양상은 서법 mood이나 이모티콘 등과 같은 비명제적인 수단을 통해 감정이 간접적으로 표출되는 형태이다.

(3) 참신하긴 하지만 주얼함 빼고 액정을 넣었으면 진짜 딱! 좋았을텐데.

(4) 하다못해 스크린의 반절을 키패드로 보여주는 기능이라도 있었으면 400배는 편했을텐데 그딴건 없습니다.

(5) 뽑기를 잘못한건지..ㅠㅠㅠㅠ

(6) 실제로 보면 너무 귀엽고 예쁘다.^^

(3), (4)의 예문에서는 저자의 감정이 '좋다', '편하다'와 같은 어휘로 직접 표출되는 것이 아니라 '~ㄹ텐데'와 같은 양상어미를 통해 '아쉬움' 등과 같은 부정적 논조가 표출되고 있다. (5), (6)에서는 최근 온라인 텍스트에서 특히 많이 사용되는 이모티콘이 저자의 감정을 나타내는 경우이다.

세 번째 양상은 문장부사나 감탄사 등을 통해 감정이 표현되는 경우이다.

(7) 퀄컴사의 스냅드래곤을 가지고도 아이팟 보다 터치감이 느리다니 ...헐..

(8) 쳇... 노키아 역시 충전을 할 때 우리나라의 다른 핸드폰 충전기는 이용할 수 없는 듯 합니다.

(9) 조금 부끄럽네요 호홋

(7)~(9)에서는 문장의 논조가 감탄사 등을 통해 표현되었다. 마지막으로 논조가 표현되는 양상은 완전한 문장의 형태가 아니라 명사구 등과 같은 문장의 일부(sentence fragment)로 나타나는 경우이다.

(10) 아이폰의 최대 단점!! 배터리 문제...

(11) 스타일을 원하시면 초코렛

(12) 빠르고 안정적인 텍스트 입력

본 연구에서는 한국어 휴대폰 리뷰 등에 등장하는 논조의 표현 양상을 파악하기 위해 논조정보부착 말뭉치를 구축하였다. 이 말뭉치를 기반으로 한국어 문장의 논조 결정 요인에 대한 언어학적 분석을 수행하였으며, 또한 이는 기계학습을 위한 학습데이터로 활용되었다.

3.1 논조정보부착 말뭉치

한국어로 작성된 휴대폰 관련 리뷰 중 긍정 혹은 부정을 표현하는 어휘들의 언어학적 특성을 살펴보기 위해, 인터넷 상의 블로그 및 휴대폰 관련 사이트의 게시판 등에서 문장을 추출하여 코퍼스를 구축하였다. 추출된 문장에 대해서는 2명의 작업자가 긍정/부정/혼합/객관의 태그를 부착하였으며, 2명의 작업자가 동일한 태그를 부착한 문장만이 코퍼스로 활용되었다. 이와 같이 구축된 코퍼스는 총 99,242 어절 규모이며, 이는 다시 32,461 어절 규모의 긍정논조 코퍼스와 21,253 어절 규모의 부정논조 코퍼스, 6,029 어절 규모의 혼합논조 코퍼스, 그리고 39,499 어절 규모의 객관논조코퍼스로 나뉜다. 각 문장은 다음의 예와 같이 논조에 따라 <positive>, <negative>, <mixed>, <objective>의 태그가 부착되어 있다.

(13) <positive> 영화/음악/OZ를 이렇게 편하고 넓고 선명한 화면에서 즐길 수 있다는 것과 기존 햅틱시리즈에 비하여 더 편리해진 위젯 및 곳곳에 보이는 사용자 편의 기능 등은 강력추천할만 합니다.</positive>

(14) <negative> 액정이 키패드에 닿아 자국이 남음 </negative>

(15) <mixed> '발열량이 심하고 스피커음이 찢어진다'에서부터 '배터리가 빨리 닳는다' 까지 많은 불평불만을 쉽게 접할 수 있음에도 그 슬림함과 가벼움, 심플한 디자인에 끌리는 유저가 꽤 있는 듯 하다. </mixed>

(16) <objective> 카메라 오른쪽으로 좀 보시면 핸드스트랩 고리 부분이 보이구요 아랫부분은 배터리 커버에요. </objective>

3.2 긍/부정 논조 어휘

문장의 논조를 결정하는데 가장 큰 역할을 하는 것은 어휘의미이다.

(17) 3.5인치의 넓은 화면으로 전체 일정을 확인할 수 있어 캘린더를 이용하는 사용자는 *유용하게* 사용할 수 있습니다.

우리는 휴대폰 분야 한국어 텍스트의 논조를 결정하는 대표적인 어휘를 파악하기 위해 앞서 소개한 말뭉치에 대한 분석을 수행하였다. 먼저 32,461 어절 규모의 긍정논조 코퍼스를 분석한 결과 13,918개의 어휘 타입이 조사되었다. 이 중 문장 논조에 결정적인 영향을 미친다고 판단되는 368개의 어휘타입을 수작업을 통해 골라내었다. 이 어휘타입을 다시 문맥에 상관없이 항상 긍정적인 논조를 갖는 어휘와 문맥에 따라 긍정 혹은 부정이 될 수 있는 어휘로 분류하였다. 이를 이후 '절대긍정어휘', '문맥의존 긍정어휘'로 부르기로 한다.

빈도수에 따라 소팅된 상위 10위 안에 드는 '절대긍정어휘' 및 '문맥의존 긍정어휘'의 목록은 표1과 같다. '절대긍정어휘'로 분류되는 어휘는 문맥에 상관없이

긍정논조를 전달하는 역할을 하였다. 반면에 '문맥의존긍정어휘'로 분류된 어휘는 주변에 등장하는 어휘의 종류 및 문맥에 따라 긍정논조를 전달하기도 하고 때로는 부정적인 논조를 전달하기도 한다

표1에서 볼 수 있는 바와 같이 절대긍정어휘들은 문맥에 상관없이 항상 긍정적인 논조를 전달한다. 예를 들어 '좋다', '예쁘다', '다양하다' 등과 같은 어휘는 주변에 공기하는 단어와 무관하게 늘 긍정논조를 전달한다. 그러나 문맥의존 긍정어휘는 공기하는 단어에 따라 긍정 혹은 부정논조를 전달할 수 있다. 예를 들어 '넓다'의 경우 '화면이 넓고 화질도 좋은 편이에요'의 경우에는 긍정논조를 전달하지만, '옆으로 넘 넓어서 그립감이 별로던데.. '의 경우에는 부정논조를 전달한다.

[표 1] 휴대폰분야 긍정논조 어휘목록

	절대긍정어휘	문맥의존긍정어휘
1	좋다	많다
2	예쁘다	잘
3	다양하다	쉽다
4	가능하다	빠르다
5	편리하다	크다
6	장점	화려하다
7	재미있다	얇다
8	만족하다	높다
9	깔끔하다	넓다
10	괜찮다	저렴하다

부정 논조의 어휘도 긍정논조의 어휘와 마찬가지로 절대부정 어휘와 문맥의존 어휘로 나누었다. 표 2는 코퍼스에 등장하는 부정논조 전달 어휘를, 문맥에 상관없이 무조건 부정논조를 전달하는 '절대부정어휘'와 문맥에 따라 논조가 상이할 수 있는 '문맥의존부정어휘'로 나누고, 이를 출현빈도순에 따라 제시한다.

'절대 긍/부정'과 '문맥의존 긍/부정'을 분류한 이유는 Wilson et al. (2005)에서 지적한 바와 같이 이러한 분류가 문맥기반 극성 (contextual polarity)을 파악하는데 적용될 수 있기 때문이다. 기계학습의 과정에서 '절대 긍/부정' 어휘와 '문맥의존 긍/부정'어휘에 대한 가중치를 달리하면 좀 더 정확한 분류가 가능할 것이다. 그러나 본 연구에서 수행된 기계학습에서는 이러한 분류를 따로 수행하지는 않고 향후 연구에서 이 분류를 적용할 예정이다.

[표 2] 휴대폰분야 부정논조 어휘목록

	절대부정어휘	문맥의존부정어휘
1	아쉽다	별로
2	불편하다	작다
3	단점	비교하다
4	문제	많다
5	아니올시다	닳다
6	느리다	낮다
7	힘들다	역시
8	떨어지다	높다
9	버그	줄다
10	비싸다	두껍다

3.3 부정어 / 양상어미

문장의 논조를 파악할 때 고려해야할 문법범주는 부정어들과 양상어미이다. 부정부사, 부정형용사, 부정어미 등과 같은 부정어들은 문장의 논조 혹은 극성 (polarity)을 바꿔주는 역할을 하기 때문에 분석단계에서 반드시 고려되어야 한다.

(18) 그런데...터치감이..생각보다 좋지 않습니다.

(19) 화질이 너무나도 안 좋은 폰이라는 점과 상당히 비싸다는 점

(20) 화질이 좋지 못하구요.

(21) 저 경우는 불편해서라도 못 쓰겠더라구요

(18)~(21)의 예문은 '~ 지 않', '안', '~ 지 못', '못' 등과 같은 부정어의 사용으로 문장의 극성이 '긍정'에서 '부정'으로 바뀐 경우이다. 어휘출현만을 특징 (feature)으로 기계학습을 시도하는 소위 'Bag of Words' 방식에 의하면 (18)~(20) 문장은 '좋다'라는 어휘에 의해 긍정으로 분류될 수도 있다. 물론 Pang et al.(2002) 등의 연구에서는 부정어휘를 극성을 변환시키는 특징으로 취급하여, 문장 내에 'not' 등과 같은 부정어휘가 등장할 경우 문장 전체의 극성이 바뀌는 것으로 판단한다. 그러나 문장 내에 한 개 이상의 용언이 출현하여 부정어의 수식영역 (scope)을 올바르게 결정해야 하는 경우, 이러한 순수 'Bag of Words' 방식은 한계에 부딪힌다.

부정어와 마찬가지로 문장의 논조 또는 극성을 변경하는 역할을 하는 어휘는 가정과 같은 양상정보를 나타내는 어휘들이다.

(22) 사용자를 생각해서 만들어줬으면~ㅠ

(23) 본체에 안테나가 내장되어 있으면 참 편할텐데...

(24) 아무리 스마트폰이라 하더라도 주기능이 모바일인 이상 좀 더 신경을 써서 만들어야 했다.

(22)~(24)의 예문은 모두 현실과 반대되는 가정의 사태를 양상어미('~으면', '~ㄹ텐데') 또는 보조용언('~어야 하')을 통해 표현하고 있다. 일반적으로 순수 기계학습 기반의 논조분석에서는 형태소 분석을 수행한 후, 기능어를 제외한 내용어에 대해서만 유니그램 정보를 추출하게 되는데, 이 경우 위 문장들의 극성을 잘못 계산하게 될 가능성이 높아진다.

4. 패턴기반 논조분석

명재석 et al.(2008)의 연구는 패턴을 활용한 논조분석 방법론을 다루고 있다. 이들의 연구에서는 반자동으로 구축된 패턴을 문장분석결과로 도출한 의미표상에 적용함으로써 문장의 긍정/부정과 같은 극성정보를 파악한다. 문장의 논조분석을 위해 패턴을 적용할 경우, 패턴매칭이 이루어지면 비교적 높은 정확율로 문장의 극성정보를 알아낼 수 있다는 장점이 있다. 그러나 다양한 도메인에 대한 논조분석 패턴을 구축하는 것은 많은 시간과 비용을 요구하는 작업이다. 이미 3장에서 언급한 바와 같이 하나의 어휘라도 적용하는 도메인에 따라 극성정보가 달라질 수가 있고 하나의 도메인 안에서도 주제가 무엇인가에 따라 극성정보가 달라질 수 있기 때문이다. 예를 들어 휴대폰 분야에서의 형용사 '길다'는 '배터리 지속시간'을 주제로 할 경우에는 긍정논조를 전달하는 반면, '배터리 충전시간'을 대상으로 할 경우에는 부정적인 논조를 전달할 수 있기 때문이다.

따라서 패턴을 이용한 논조분석방법에서는 각 도메인별로 수많은 패턴의 구축이 필요하다. 그러나 현실적으로 이러한 패턴을 구축한다는 것은 매우 어려운 문제이기 때문에 이 방법론의 실현가능성을 낮게 하는 요인이 된다. 우리는 이러한 문제를 해결하기 위해 기존에 다른 목적으로 구축된 패턴을 재활용하는 방안을 제안한다. 우리는 이러한 문제를 해결하기 위해 패턴기반 기계번역 시스템을 위해 구축된 패턴을 논조분석에 재활용하는 방안을 제안한다.

Hong et al.(2005)에서 소개된 한영 기계번역시스템의 대역패턴은 한국어 분석을 위한 한국어 술어-논항구조와 영어생성을 위한 영어 술어-논항구조로 이루어져 있다.

한영 대역패턴:

한국어 술어-논항구조 > 영어 술어-논항 구조

예) A=사람!가 B=음료!를 마시다 > A drink:v B

한국어 술어-논항 구조는 용언을 중심으로 해당 용언이 취할 수 있는 논항의 슬롯갯수 및 격조사 정보, 그리고 논항에 대한 의미제약조건 정보로 구성되어 있다. 그리고 '>'를 기준으로 오른쪽에 영어 문장의 생성을 위한 영어 술어-논항 구조가 표기되어 있다. 이 패턴의 한국어 분석부분은 의존문법 기반의 파서를 위한 분석지식으로 사용될 수 있을 뿐만 아니라 논조정보가 부착될 수도 있다. 이에 본 논문에서는 위의 한영 대역패턴을 수정한 아래와 같은 패턴포맷을 패턴기반 논조분석을 위한 방법으로 제안한다.

논조분석패턴:

한국어 술어-논항구조: 논조정보(타겟)
논조정보: 긍정/부정
타겟 = {디자인, 통화품질, 배터리, 가격, 애프터서비스 ...}

위에서 '타겟'이란 휴대폰 분야의 리뷰 등에 주제로 등장할 수 있는 '디자인', '통화품질', '배터리' 등과 같은 키워드들이다. 아래의 예에서 소개하는 패턴은 각각 '긍정'과 '부정'의 논조를 결정할 수 있는 패턴들이고, 각 패턴이 적용될 수 있는 주제가 '타겟' 정보로 나타나있다.[6]

예1) A=인공물!가 B=모양!가 예쁘다 : 긍정(target: 디자인)
ex) 아이폰은 겉모습이 참 예뻐서 맘에 들어요

예2) A=속도!가 느리다 : 부정(target: 속도)
ex) 키패드의 반응속도가 너무 느려서 답답합니다

이 방법론은 위와 같은 논조분석패턴을 구조분석 결과에 적용하면 문장의 논조를 정확하게 파악할 수 있는 장점이 있다. 본 연구에서 활용한 ETRI 한국어 의존구조 분석파서는 입력문장에 대해 의존구조 분석을 수행할 때 한영 대역패턴을 적용한다. 아래의 예문 '또한, 홈 화면에 위젯으로 배치하고 사용할 수 있어 편리하다'의 의존구조를 분석하기 위해 적용한 한영 대역패턴들은 다음과 같다.

(25) 또한, 홈 화면에 위젯으로 배치하고 사용할 수 있어 편리하다

<VP1> 편리하!다 { A=*!가 편리하!다: 긍정(target:*) > A be:v_convenient [:DEFAULT] [편리하다1] ==> [단일용언] }
VP-1::be:vconvenient$VERB$[]

<VP2> 사용하!다 { A=*!가 사용하!다 > A use:v [:DEFAULT]$$$ {GENERATED

[6] 예 1, 2에서 알파벳 대문자 A, B는 논항의 위치를 표시하는 변수이며, '인공물', '속도', '모양'은 논항을 의미적으로 제약하는 의미코드이고, !는 논항명사와 조사를 구분해주는 구분자이다.

BY ENGINE} [사용하다1] ==> [단일용언] }
VP-1::use:v[]

<VP3> 홈화면[home_screen^구체물의부분]!에 위젯[Widget^ETC]!로 배치하!다 { A=*!가 B=*!로 배치하!다 > A arrange:v in:p B [:DEFAULT] [배치하다177] ==> [불완전매칭-조사일부매칭] }
VP-1::arrange:vin$PREP$Widgetin$PREP$thehomescreen[]

의존구조분석 파서는 문장을 구성하는 단문의 영역(scope)을 결정한 후, 각 단문내의 용언을 중심으로 의존구조를 생성한다. 본 연구를 위해 구축한 논조분석 패턴은 이러한 의존구조에 매칭될 경우, 해당 문장의 논조를 결정하게 된다. 위 예문은 3개의 단문으로 구성되어 있고, 각 단문의 의존구조가 논항의 의미정보와 함께 표기되어 있다. 이 예문에서는 첫 번째 단문이 'A=*!가 편리하다:긍정(target:*)'의 패턴에 매칭되어 '긍정'의 논조를 지니는 것으로 분석된다.

물론 문장의 논조결정이 패턴매칭으로만 완전히 끝나지는 않는다. 앞 절에서 살펴본 극성변화 요인들, 예를 들어, 부정어미, 양상어미 등의 출현여부에 따라 패턴매칭에 의해 결정된 논조는 변경될 수 있다. 예를 들어, '통화음질이 좋지 않다'의 경우 'A=*!가 좋다: 긍정(target:*)' 패턴이 적용되어 '긍정'의 논조로 분석되나, 그 다음 단계에서 문장의 논조, 혹은 극성 polarity을 변경시키는 요인인 부정어미 '~지않'의 영향으로 긍정논조가 부정논조로 변경되게 된다. 마찬가지로 예를 들어 '좋을텐데'의 경우도 '좋다' 패턴의 적용으로 우선적으로는 긍정으로 결정되지만 양상어미 'ㄹ텐데'의 적용으로 최종적으로는 부정으로 결정된다.

본 연구에서는 '긍정' 학습코퍼스에서 2회 이상 출현한 162개의 용언을 키워드로 갖는 한-영 대역패턴으로부터 한국어 부분을 추출하여 논조분석패턴을 수작업으로 구축하였다. '부정'코퍼스에서도 마찬가지로 2회 이상 출현한 161개의 용언을 추출하여 '긍정'패턴과 같은 방법으로 논조분석패턴을 구축하였다. 휴대폰 분야에 대하여 '긍정'과 '부정'패턴으로 구축된 용언은 다음과 같다.

<u>긍정패턴 용언</u>:

가능하다, 가볍다, 간결하다, 간단하다, 간지나다, 간편하다, 감동하다, 강력하다, 경쾌하다, 고급스럽다, 괜찮다, 굉장하다, 구현되다, 권하다, 귀엽다, 근사하다, 기대하다, 기쁘다, 기특하다, 깔끔하다, 깜찍하다, 깨끗하다, 꽂히다, 끌리다, 날렵하다, 날씬하다, 내뿜다, 너끈하다, 넉넉하다, 노련하다, 눈여겨보다, 능가하다, 다양하다, 다재다능하다, 다채롭다, 다행이다, 단단하다, 달콤하다, 대단하다, 덕분이다, 독창적이다, 돋보이다, 땡기다, 똑똑하다, 뛰어나다, 러블리하다, 막강하다, 막대하다, 만족하다, 맛깔스럽다, 매끄럽다, 매료하다, 먹어주다, 멋지다, 모던하다, 반갑다, 반하다, 발휘하다, 배려하다, 부드럽다, 부럽다, 빛나다, 빠져들다, 빨려들어가다, 뽐

내다, 뿌듯하다, 사랑스럽다, 사로잡다, 상큼하다, 새롭다, 색다르다, 선호하다, 세련되다, 세심하다, 섹시하다, 소중하다, 수월하다, 순조롭다, 스마트하다, 스타일리시하다, 시간절약하다, 신기하다, 신비롭다, 신선하다, 신통하다, 심플하다, 쏠쏠하다, 쓸만하다, 아기자기하다, 아름답다, 안전하다, 안정적이다, 알맞다, 알차다, 앙증맞다, 앞서다, 야무지다, 야심차다, 어울리다, 어필하다, 영리하다, 영특하다, 예쁘다, 완벽하다, 용이하다, 우수하다, 우월하다, 유용하다, 유쾌하다, 윤택하다, 은은하다, 이색적이다, 익숙하다, 인상깊다, 인상적이다, 자랑하다, 자유롭다, 재미있다, 젊다, 정교하다, 조화롭다, 좋다, 죽이다, 즐겁다, 즐기다, 지르다, 짜릿하다, 쩔다, 착하다, 참신하다, 청순하다, 추천하다, 충실하다, 친절하다, 칭송하다, 쾌적하다, 큐트하다, 탁월하다, 탐나다, 특별하다, 특출나다, 튼실하다, 편리하다, 편안하다, 편하다, 푸짐하다, 풍성하다, 혁신적이다, 현명하다, 활발하다, 활용하다, 황홀하다, 효과적이다, 훈훈하다, 훌륭하다, 흐뭇하다, 흡족하다, 흥미롭다

부정패턴 용언:

거슬리다, 걸리다, 걸리적거리다, 겁나다, 과하다, 괴로워하다, 교체하다, 구리다, 굴러다니다, 귀찮다, 급급하다, 깨지다, 꺼림직하다, 꺼지다, 꽉막히다, 끊기다, 난감하다, 너무하다, 농락하다, 느끼하다, 느리다, 늙다, 답답하다, 당황하다, 더러워지다, 둔탁하다, 따끔하다, 따지다, 딸리다, 떨어지다, 막막하다, 만무하다, 망가뜨리다, 망하다, 못되다, 무겁다, 무색하다, 무시무시하다, 무시하다, 무안하다, 뭉게지다, 미려하다, 미약하다, 밋밋하다, 버겁다, 버리다, 버벅거리다, 번잡하다, 벗겨지다, 베끼다, 변변하다, 복잡하다, 부끄럽다, 부실하다, 부자연스럽다, 부정확하다, 부족하다, 불가능하다, 불쌍하다, 불안하다, 불편하다, 비싸다, 빈곤하다, 빈약하다, 빡빡하다, 빡치다, 빽빽하다, 뻔하다, 뻣뻣하다, 뿌옇다, 생뚱맞다, 생소하다, 석연치않다, 성질내다, 속터지다, 슬프다, 시끄럽다, 식상하다, 실망하다, 싫다, 심심하다, 쏟아붓다, 쓸데없다, 쓸모없다, 아니올시다, 아쉽다, 안타깝다, 애매하다, 약하다, 양산하다, 어둡다, 어렵다, 어색하다, 어이없다, 어지럽다, 억지스럽다, 엉성하다, 엉키다, 역행하다, 열받다, 열악하다, 우려먹다, 우려하다, 우울하다, 울다, 울렁거리다, 위태롭다, 위험하다, 의심하다, 저하되다, 전무하다, 죄송하다, 주의하다, 주저하다, 죽다, 지다, 지겹다, 지나치다, 지루하다, 지저분하다, 지지직거리다, 지체하다, 질리다, 쪽팔리다, 찌그러지다, 찜찜하다, 찢어지다, 체념하다, 초라하다, 촌스럽다, 추궁하다, 추락하다, 타박하다, 탁하다, 토로하다, 퇴보하다, 투박하다, 퉁명스럽다, 튕기다, 팔아먹다, 폐기하다, 폐쇄적이다, 포기하다, 한심하다, 할말없다, 해봤자다, 허술하다, 허전하다, 허접하다, 험하다, 헤매다, 형편없다, 환불하다, 황당하다, 후퇴하다, 후회하다, 흉하다, 흉흉하다, 흐리다, 흔하다, 힘들다,

이 외에도 학습코퍼스에서 단 1회 출현한 용언의 경우도 문장의 논조분석에는 큰 역할을 할 수 있다. 그러나 본 연구에서는 단 1회만 출현한 용언은 시간과 비용의

문제 때문에 별도로 패턴을 구축하지는 않았다. 이러한 패턴을 추가적으로 구축할 경우 성능의 월등한 향상이 기대되는데, 전체 출현용언의 대다수를 차지하는 저빈도출현 용언의 패턴을 구축하는 방안으로는 영어 대역어 정보를 활용하는 방안을 생각해볼 수 있다.

이러한 점에서 본 연구에서 제안하는 기계번역 시스템의 패턴을 논조분석 패턴으로 재활용하는 방안은 영어 대역어정보를 활용할 수 있다는 장점이 있다. 한국어 분석패턴에 대응하는 영어 단어에 대해 센티워드넷 등의 논조 정보를 참조하여 역으로 한국어 단어의 논조를 반자동으로 파악할 수 있기 때문에, 향후 논조분석패턴을 확장하기에 용이하다. 예를 들어 'A=*!가 만족스럽다 > A be:v satisfied'와 같은 패턴에서 'satisfied'의 논조정보를 센티워드넷을 참조할 경우, '객관논조=0.625, 긍정논조=0.375, 부정논조=0'과 같은 결과를 얻게 된다. 즉, 영어 단어 'satisfied'는 주로 긍정논조로 사용됨을 알 수 있다. 이를 통해 우리는 거꾸로 이 단어에 대응하는 한국어 표현 '만족스럽다'가 긍정논조일 것임을 추측할 수 있다. 물론 이러한 접근법에서는 어휘의 의미가 센티워드넷에서 어떤 신셋에 해당하는지를 정확하게 파악해야 하지만, 논조패턴을 반자동으로 구축하는데 큰 도움이 될 수 있음을 알 수 있다. 이 방안에 대한 연구는 본 논문에서는 다루지 않기로 하고 향후의 연구로 남기도록 한다.

5. 기계학습기반 논조분석

순수 패턴기반의 논조자동분석은 패턴을 대용량으로 구축하는 것이 어렵기 때문에 데이터부족 현상을 보인다는 문제 뿐 아니라, 문장의 논조가 용언을 중심으로 한 패턴으로 표현되지 않고, 명사나 부사 등과 같은 품사만으로 표현될 경우에 그 논조를 파악하기 어렵다는 문제도 있다. 다음의 예문을 보자.

(26) 재질은 '럭셔리'를 표방하는 제품답게 메탈로 처리되어 있다.

(27) 포스트에 언급된 버그들이 한둘이 아닌데 여기서 내가 겪은 것들을 한 번 말해본다.

예문 (26)에서의 긍정논조는 '럭셔리'라는 긍정논조의 명사를 통해 표현되고, (27)에서의 부정논조는 '버그'라는 명사를 통해 표현된다. 순수 패턴기반 논조분석 방식에서는 문장의 논조를 결정하는 요인을 동사나 형용사와 같은 용언으로 보고 이를 중심으로 한 패턴의 적용여부를 통해 문장의 논조를 계산한다. 그러나 위의 예문들과 같이 문장의 논조가 용언이 아닌 명사나 부사 등과 같은 기타 품사를 통해 표현될 경우, 문장의 논조를 파악하기가 어렵다. 이에 본 연구에서는 유니그램(unigram)을 사용한 기계학습 기반의 논조자동분류 방안을 패턴 기반논조 분석의

백업장치로 활용하였다. 본 연구에서 적용한 'Support Vector Machine'은 자연언어처리 분야에서는 이미 수년전부터 널리 사용된 방법론이지만 언어학 분야에서는 비교적 생소한 테마이므로, 기계학습기반 논조분석 방법론을 다루기에 앞서 다음 장에서 'Support Vector Machine'에 대한 내용을 먼저 다루기로 한다.

5.1 'Support Vector Machine'을 포함한 기계학습 개념

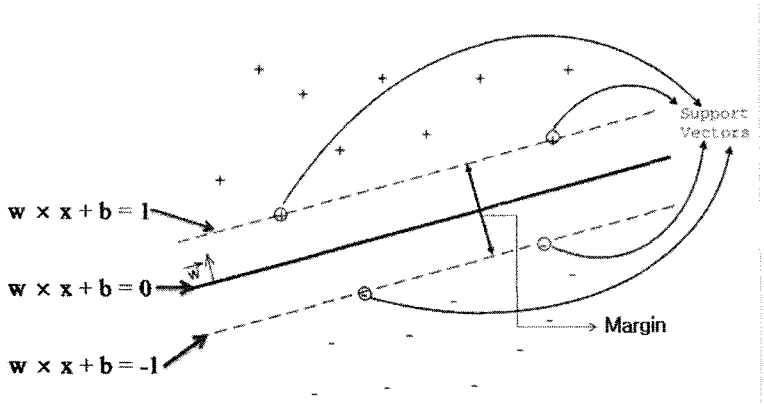
기계학습 (Machine Learning)이란 인간의 학습 능력을 컴퓨터를 통해 구현하기 위한 노력의 하나로, 경험적 데이터 (empirical data)를 기반으로 하여, 점진적인 학습 과정을 통해 주어진 데이터를 분류 (classification)하거나 그룹화 (clustering)하기 위한 모델을 찾아내는 과정을 말한다. 이를 위한 방법은 크게 데이터 분류를 위한 전사적 학습 (supervised learning) 알고리즘과 그룹화를 위한 비전사적 학습 (unsupervised learning) 알고리즘으로 나눌 수 있다. 전사적 학습 알고리즘은 분류하고자 하는 각 예제 (example)의 특징 값들과 그에 대응하는 클래스 정보로 이루어진 학습 데이터, 즉 (예제, 클래스) 형태로 구성된 학습 데이터를 기반으로 예제의 특징값들과 대응하는 클래스 간에 매핑관계를 표현하는 함수 혹은 규칙을 컴퓨터를 통해 자동으로 찾아내며, 이를 기반으로 새로운 예제가 주어졌을 때에 그에 대응하는 클래스를 예측한다.

반면에 비전사적 학습 알고리즘은 각 예제의 특징 값들만으로 이루어진 학습 데이터를 기반으로 유사한 예제들의 그룹을 자동으로 찾아냄으로써 학습데이터에 내재된 그룹의 수나 특징을 파악하는 것을 목적으로 한다.

본 연구의 텍스트 논조 자동분석은 전사적 학습 알고리즘을 통해 주어진 문서 혹은 문장들의 논조를 결정하는 함수 혹은 규칙을 생성하는 문제로 형식화될 수 있다. 전사적 학습 알고리즘으로는 'Support Vector Machine' (이하 'SVM'으로 칭함), 신경망 (Neural Network), 결정 트리 (Decision Tree) 등의 다양한 방법들이 있다. 특히 이 중 'SVM'은 최근 여러 응용 분야에 적용되고 있으며, 좋은 성능을 보여주고 있어 많은 주목을 받고 있다. 그러므로 본 연구에서는 텍스트 논조 자동분석을 위해 학습 코퍼스에 포함된 각 문장들의 논조를 결정하는 함수를 'SVM'을 이용하여 학습하고 이를 기반으로 새로이 주어진 문장의 논조를 예측하고자 한다.

Cortes&Vapnik(1995)에 의해 제안된 방법인 'SVM'은 두 개의 클래스로 이루어진 학습 데이터에 대해 두 클래스간의 구분을 위한 결정 경계면 (decision boundary) 주위에 최대한의 마진 (margin)을 갖도록 결정 함수 (decision function)를 학습하는 방법으로, 최대마진분류기 (maximal margin classifier)라고도 한다. 아래 그림은 'SVM' 중에서도 선형적 결정 경계면 (linear decision boundary)을 가지는 선형 모델을 이용한 분류의 예를 보여주는 것으로, 두 종류의 개체들인 + 클래스와 -클래스를 잘 구분하면서 최대의 마진을 갖는 결정 경계면의 예를 보여주고 있다. (그림

1)



[그림 1] SVM의 선형모델

위와 같이 결정 경계면이 선형적 (linear)인 모델의 경우, 두 클래스를 구분 짓는 결정 함수는 'wx + b = 0'의 형태로 표현되며, 여기서 w는 결정경계면의 기울기를 나타내는 가중치 벡터 (weight vector)이고 b는 바이어스 (bias)를 의미한다. 그리하여 주어진 학습 데이터로부터 궁극적으로 결정함수의 파라메터인 w와 b를 학습과정을 통해 결정한다. 본 연구에서는 논조분석을 위해 자바 'Java'로 구현된 'SVM' 라이브러리인 'LIBSVM'의 2.89버젼을 사용하여 실험을 수행하였다.[7]

5.2 언어학적 특성을 반영한 분류 알고리즘

5.2.1 입력 특징 단어 선택. 논조자동분석 실험을 위해 학습 코퍼스에 속한 각 문장들로부터 추출한 유니그램 (unigram)을 'SVM' 기반 학습을 위한 입력특징 (input feature)으로 사용한다. 다만, 이러한 특징들을 모두 고려할 경우 결정 모델이 매우 복잡해지고 학습 시간은 많이 걸리는 반면 학습 성능은 만족스럽지 못한 문제가 있기 때문에 유니그램 특징들 중 'CPD (Categorical Proportional Difference)' 값이 상위랭크에 위치한 것들만을 추출하여 학습과정에 사용하였다. 주어진 단어 w를 포함한 문장이 클래스 c에 해당될 때, 단어 w의 클래스 c에 대한 'CPD' 값은 아래와 같이 계산된다.

$$\mathrm{CPD}(w,c) = \frac{A-B}{A+B} \tag{8}$$

A: 클래스 c에 속한 문장에서 단어 w가 나타나는 횟수

B: 클래스 c에 속하지 않은 문장에서 단어 w가 나타나는 횟수

[7] LIBSVM에 대해서는 Chang & Lin(2001) 참조

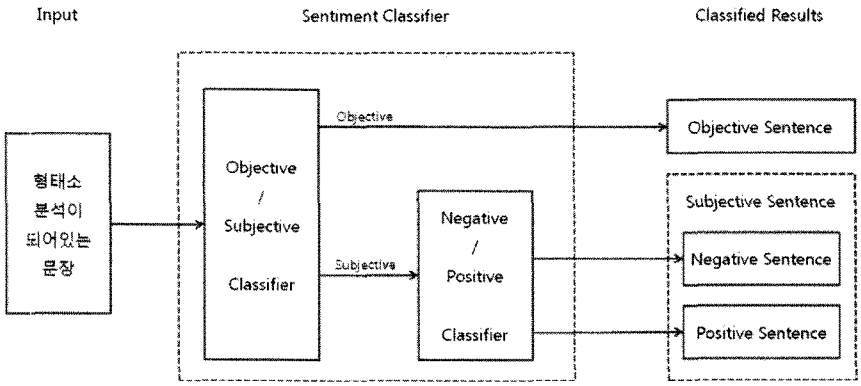
위의 수식을 이용하여, 유니그램 특징으로 추출된 각 단어에 대하여 클래스별로 'CPD' 값을 계산한 후 아래와 같이 최대값을 그 단어의 'CPD' 값으로 최종 결정한다. 즉, 주어진 단어 w에 대한 'CPD' 값은 아래와 같이 계산된다.

$$\mathrm{CPD(w)} \quad = \quad \mathrm{maxi}\{\mathrm{CPD(w, ci)}\} \tag{9}$$

이렇게 얻어진 'CPD' 값은 이진 클래스 (binary class)의 경우 0과 1사이의 값을 가진다. 계산된 'CPD' 값을 가장 큰 값인 1에서부터 내림차순으로 정렬한 후, 10회에 걸친 교차검증 (10-fold cross validation)에 의해 실행된 결과가 최적의 분류정확도를 가지는 'CPD' 값을 임계값 (threshold)으로 사용하여 그 값 이상의 'CPD'를 갖는 유니그램 특징들을 추출하여 'SVM' 기반 분류기 학습에 사용하였다.

실제 실험에 사용한 학습 코퍼스로부터 추출된 전체 유니그램 특징 단어 각각에 대해 'CPD' 값을 계산해본 결과, 그 값이 1인 경우가 많이 발생하였고 이 중 많은 경우에 전체 문장들에 나타난 빈도수가 1인 경우가 대부분이었다. 그리하여, 본 실험에서는 전체 문장들 중 최소 2번 이상이 나타나는 특징 단어만을 추출한 후, 추출된 특징 워드에 대해 'CPD' 값을 계산하여 'SVM' 기반 분류기의 최종 입력 특징 단어를 결정하였다. 일단 분류기의 입력 특징 단어가 결정되면, 각 입력 특징 단어의 빈도수가 아닌 존재여부 (presence)에 따라 1 (presence) 또는 0 (not presence)으로 구성된 값을 가지는 특징 벡터를 입력 데이터로 사용하게 된다. (cf. Bo Pang et al.(2002))

5.2.2 문장 단위의 논조분석을 위한 'SVM' 기반 분류기. 본 연구에서는 문장들의 논조자동분석을 위해 아래와 같은 형태로 'SVM' 기반 분류기를 설계하였다. 즉, 형태소 분석이 되어있는 문장을 논조 분류기를 통해 '객관/긍정/부정'의 3가지 클래스로 분류하기 위해, 2단계의 이진 분류기를 사용한다. 먼저 주어진 문장에 대해 형태소 분석을 수행하고 이로부터 추출된 유니그램 특징을 기반으로 입력 특징 단어에 대한 특징 벡터를 구성한 후, 아래 그림에서와 같이 '주관/객관' (Subjective/Objective) 분류기를 통해 문장의 논조를 '주관/객관'으로 분류하고, '주관' 문장에 대해서는 다시 'Positive/Negative' (P/N) 분류기를 통해 긍정 혹은 부정의 논조를 최종 판단하는 2단계의 분류를 시도한다. 이를 통해, 각 문장은 최종적으로 '객관/긍정/부정' 중의 하나의 논조로 예측된다. (그림 2)



[그림 2] 2단계 'SVM' 기반 문장 논조 분류기에 의한 논조 자동분석 과정

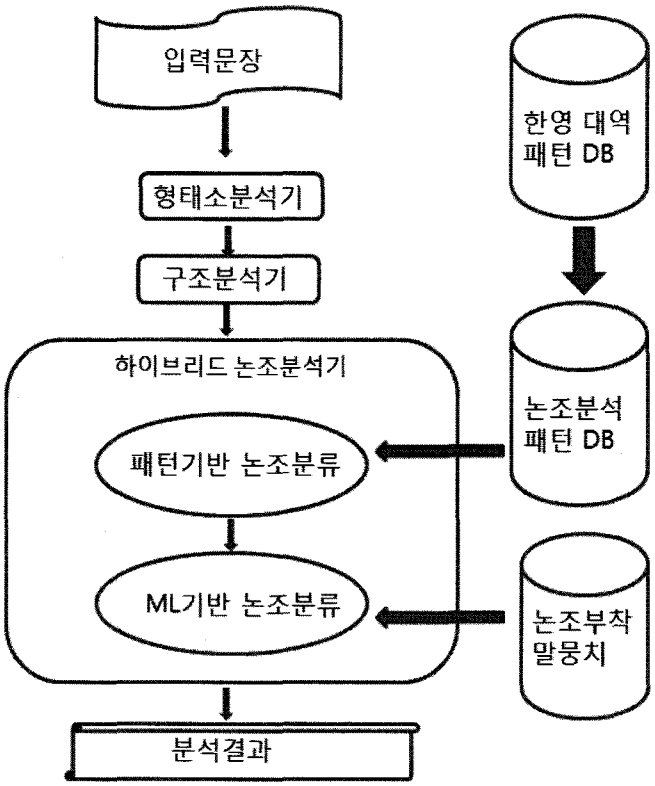
6. 하이브리드 텍스트 논조 자동분석

6.1 하이브리드 분석 방법

본 연구에서 제안하는 텍스트 논조 자동분석 시스템은 크게 두 가지의 하위 모듈로 구성되어 있으며, 각 모듈은 순차적으로 적용된다. 입력문장이 형태소 분석과 구조분석을 거치게 되면 우선 패턴기반 추출기가 적용된다. 패턴기반 추출기는 논조분석패턴 DB를 활용하여 논조분석 패턴의 매칭을 시도하며, 패턴이 매칭된 경우 문장의 논조를 변경하는 부정어미나 양상어미의 존재여부를 확인한 후, 만약 이러한 어미 등이 존재하지 않는다면 패턴매칭 결과가 바로 최종분석결과로 출력된다. 그러나 '~지않' 등과 같은 부정어미나 '~ㄹ텐데' 등과 같은 양상어미가 존재하면 패턴매칭으로 도출된 분석결과가 극성이 바뀌어 출력되게 된다.

그러나 패턴매칭에 실패하면 이 문장은 5.2.2에서 소개하였던 기계학습 기반의 논조분석 추출기에 입력된다. 기계학습 기반의 추출기는 입력문에 대해 'SVM' 알고리즘을 적용하여 우선 문장의 논조가 객관적인지 주관적인지를 판단한 후, 만약 논조가 주관적이라고 판단되면 다시 긍정논조인지 부정논조인지를 판단하게 된다. (그림 3)

이와 같은 방법론을 적용할 경우 논조분석패턴을 우선적으로 적용하여 문장의 논조에 대한 정확한 분석을 할 수 있을 뿐만 아니라, 패턴이 갖고 있는 장점, 즉, 문장의 논조 뿐만 아니라 논조의 토픽이 되는 '타겟'까지도 추가적으로 파악할 수 있다는 장점이 있다. 또한 패턴기반 방법론으로 파악하기 쉽지 않는 문장의 논조를 추가적으로 기계학습 기반의 분류방안을 적용함으로써 분석의 정확도 뿐만 아니라 재현율을 높일 수 있다는 장점이 있다. 다음 절에서는 이러한 방법론의 성능을 평가하기 위한 실험의 과정과 결과를 소개한다.



[그림 3] 하이브리드 문장논조 자동분석

6.2 성능평가 및 결과분석

본 논문에서 제안하는 방법론의 타당성을 검증하기 위해 다음과 같은 실험을 수행하였다. 첫째, 본 연구를 위해 구축된 코퍼스에서 추출된 고빈도 논조분석패턴만을 활용한 순수 패턴기반 방법론의 정확율과 재현율을 조사하였다. 둘째, 'SVM' 기반 기계학습 방법론의 정확율과 재현율을 조사하였다. 셋째, 본 논문에서 제안하는 패턴기반 방법론과 기계학습 방법론을 혼합한 하이브리드 방법론의 정확율과 재현율을 조사하였다. 앞선 첫 번째, 두 번째 방법론을 통한 결과는 본 논문에서 제안하는 방법론과 비교하기 위한 베이스라인으로 활용되었다.

실험에 사용된 테스트셋은 총 401문장으로 구성되었다. 학습코퍼스와 동일한 출처에서 추출한 문장이지만 학습코퍼스의 문장들과 중복되지는 않았다. 총 401문장에 대해 학습코퍼스를 구축한 경험이 있는 2인의 작업자가 논조태깅을 수행하였다.[8] 401문장에는 제품의 사용방법 등과 같은 객관적 논조의 문장, 긍정논조 및

[8] 실제로는 총 450문장에 대한 논조태깅을 수행하였으나 2명의 작업자의 태깅결과가 서로 다르고 논

부정논조의 문장 등이 골고루 포함되어 있다. 그러나 테스트셋에서 긍정과 부정의 논조가 혼합되어 있는 문장은 배제하였다. (표 3 참조)

[표 3] 테스트셋 논조별 구성

	긍정논조	부정논조	객관논조
문장수	124	124	153

먼저 순수 패턴기반의 방법론만을 적용한 결과 총 67개의 문장에 대해 논조 분석패턴이 적용되어 이 중 53개의 문장이 정확하게 분석되었다. 이는 약 79.11%의 정확율이지만, 정확율과 재현율을 모두 반영한 'F-measure' 값은 아래의 수식에 의해 약 0.28이다.

F-measure = 2*Precision*Recall/(Precision+Recall)

F-measure = 2*0.7911*0.1670/(0.7911+0.1670) = 0.28

'SVM' 기반의 순수 기계학습 분류방법을 적용해본 결과 총 401문장에 대해 모두 결과를 출력했으며, 258문장을 올바르게 분류하였다. 따라서 64.34%의 정확율을 보였으며, 'F-measure' 값은 약 0.78이다.

F-measure = 2*0.6434*1/(0.6434+1) = 0.7830

최종적으로 본 연구에서 제안한 하이브리드 방법론을 적용한 결과 총 401문장 중 262문장을 올바르게 분류하였다. 이 결과 65.34%의 정확율을 보였고, 'F-measure' 값은 약 0.79이다.

F-measure = 2*0.6534*1/(0.6534+1) = 0.7903

[표 4] 논조분석 성능평가 결과

	순수패턴기반	순수기계학습기반	하이브리드 방법론
정확율	0.79	0.6434	**0.6534**
F-measure	0.28	0.7830	**0.7903**

본 연구에서 제안한 하이브리드 문장논조 자동분석 방법론은 기존의 순수패턴기반 방식이 갖는 분석의 정확율은 크게 해치지 않으면서 정확율과 재현율을 모두 고려한 'F-measure' 측면에서는 많은 향상을 가져올 수 있음을 보였다. 그러나 순수 기계학습 기반의 방법론과 비교했을 때는 정확율과 'F-measure'의 측면에서 약간의 성능향상을 보였지만 그 차이가 크지는 않았다. 그럼에도 불구하고 본 연구에서

란의 여지가 있는 문장들을 제거한 후, 최종적으로 401개의 문장만으로 실험을 수행하였다.

제안하는 방법론은 순수 기계학습 방법론에서 수행하는 것과 같은 문장의 단순한 긍정/부정 논조분석 뿐만 아니라, 대상의 어떠한 자질에 대해 긍정/부정 논조가 전달되는지를 추가적으로 파악할 수 있는 질적인 향상을 가져올 수 있다. 예를 들어 '디자인이 너무 지루한 듯..'과 같은 문장을 단순히 '부정'논조로만 파악하는 것이 아니라, 이것이 '디자인'이라는 휴대폰의 자질에 대한 '부정'논조임을 패턴적용을 통해 알아낼 수 있었다.

(28)은 패턴이 잘못 적용되어 논조가 오분류된 예이다. 이 예문에는 긍정논조패턴이 구축되어 있는 '만족하다'가 등장하기 때문에, '만족하다'패턴에 매칭되어 긍정논조로 분류되었다. 이 경우는 어미 '~ㄹ지'가 문장의 논조를 긍정에서 부정으로 바꾸는 역할을 하기 때문에 이 문장의 논조는 '부정'으로 분석되는 것이 맞다. 그러나 어미 '~ㄹ지'는 항상 양상어미로 기능하는 것이 아니기 때문에 현 시스템에서는 논조를 바꾸는 어미로 등록되어 있지 않다.

(28) 진짜...T는 언제까지 뒤에다가 로고를 붙여야 만족할지...

예문 (29)는 이와 유사한 예이지만, 논조를 바꾸는 역할을 양상어미 등이 하는 것이 아니라 동사 '해결하다'가 하고 있다. 아래 예문에 등장하는 '어렵다'가 부정논조를 전달하는 어휘이므로, 부정논조패턴에 적용되어 전체 문장의 논조가 '부정'으로 분류된다. 그러나 의미적으로 보면 '어렵다는 인식을 해결하다'의 의미이므로 '긍정'으로 분류하는 것이 타당하다.

(29) 특히, 햅틱 UI를 탑재하여 삼성전자 풀 터치 휴대폰 사용자들도 쉽게 적응하여 사용할 수 있도록 한 점은 스마트폰이 어렵다는 인식을 해결하고 조금 더 쉽게 갤럭시A를 사용할 수 있게 해준다.

이 두 예문의 경우는 모두 기계학습기반 모듈은 올바르게 분류한 것이어서, 하이브리드 방법론의 성능을 더 높이기 위해서는 우선적으로 패턴기반 분석모듈의 성능을 높여야 할 것으로 보인다.

다음 예문 (30)은 패턴기반 방법론이 기계학습 기반 방법론와 비교하여 올바른 분석결과를 출력하는 경우이다. 이 문장에는 '좋다'라는 긍정어휘가 사용되었으나 '~으련만'과 같은 양상어휘가 결합된 경우이기 때문에, 패턴기반 모듈에서는 '부정'논조로 분석된다. 그러나 유니그램만을 고려하는 기계학습 기반 모듈에서는 이 문장을 '긍정'으로 분류하였다.

(30) WM이 속도가 조금만 빨라도 참 좋으련만

아래 예문 (31)은 패턴적용에 실패하고 기계학습 기반의 분석모듈도 잘못 분류한 예문이다. 기계학습 기반 분석모듈은 아래 문장을 '객관적'으로 분류하였다. 그

이유는 아마도 이 문장에 긍정이나 부정 논조를 추측할 수 있는 어떤 어휘도 등장하지 않기 때문일 것이다.

(31) 연구진들 아이폰 한번 만져나 보고 아이폰 대항마로 이런 폰을 출시하는건지 모르겠습니다.

이러한 문제는 향후 학습데이터의 확충으로 유니그램 뿐만 아니라 바이그램 등을 학습특징으로 고려할 경우 어느정도 해결될 수 있을 것으로 보인다.

7. 결론

본 연구에서는 텍스트 논조의 자동분석을 위해 언어분석 기술을 활용한 패턴기반 논조분석과 기계학습기반의 논조분석 방법론의 장점만을 결합한 하이브리드 방법론을 제안하였다. 하이브리드 방법론은 패턴기반 방법론의 높은 정확율과 유용한 분석결과 정보라는 장점을 거의 유지하면서 패턴기반의 문제점으로 지적되었던 낮은 패턴매칭률의 문제를 기계학습기반 방법론을 적용하여 해결하였다. 이 결과 순수 패턴기반의 경우 0.28에 불과하였던 'F-measure' 스코어가 0.79로 상승함을 볼 수 있었다.

본 논문에서 제안한 방법론의 성능을 더 높이기 위한 향후 연구방향으로는 크게 두 가지의 연구를 들 수 있다. 첫째는 문장의 논조정보가 부정어나 양상어미 등으로 변경되는 것이 아니라 일반 어휘의 의미에 의해 변경되는 경우에 대한 처리 방안이다. 아래의 예문 (32)에 등장하는 '단순함'이라는 명사는 일반적으로 부정적인 논조를 전달하는 단어이지만, 이 명사의 술어 '탈피하다'의 어휘의미로 인해 부정논조가 긍정논조로 변경되었다.

(32) 하단은 검정색으로 처리하여 단순함을 탈피하였다.

현재는 극성을 바꾸는 요인으로는 부정 선어말 어미와 양상어미 등만을 고려하였기 때문에 위와 같은 문장의 극성을 올바로 파악하기가 어렵다. 또한 기계학습기반의 분석방법에서도 유니그램만을 고려하였기 때문에 이 문장의 극성을 올바로 파악하기는 쉽지 않다. 이러한 문제를 해결하기 위한 방안으로 바이그램 정보를 기계학습의 특징으로 사용할 수 있겠지만, 이 역시 데이터 부족문제에 부딪힐 가능성이 높기 때문에 어휘의미를 고려한 방법론의 고안이 필요하다고 할 수 있다.

둘째는 앞서 4장에서 언급한 바와 같이 한영 대역패턴의 영어대역어 정보를 활용하여 한국어 용언의 논조정보를 반자동으로 구축하는 방안이다. 센티워드넷에 부착된 영어어휘의 논조정보와 한영 대역패턴의 정보를 활용하면 한국어 논조자동분석 패턴 뿐만 아니라 한국어 감정어휘망도 반자동으로 구축할 수 있을 것으로 보인다.

< 참고문헌>

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

Aman, S. and S. Szpakowicz. 2008. Using Roget's Thesaurus for Fine-grained Emotion Recognition. In *Proceedings of IJCNLP*, pp. 312–318.

Chang, C. and C. Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://csie.ntu.edu.tw/~cjlin/libsvm.

Cortes, C. and V. Vapnik. 1995. Support vector network. *Machine Learning* 20, 273–297.

Ding, X. and B. Liu. 2007. The Utility of Linguistic Rules in Opinion Mining. *SIGIR 2007*, pp. 812–812.

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

Fries, N. 2009. "Die Kodierung von Emotionen in Texten". *Journal of Literary Theory*, pp. 19–71.

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)*, pp. 841–847.

Hatzivassiloglou, V. and K. Mackeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 174–181.

Hong, M., Y. Kim, C. Kim, S. Yang, Y. Seo, and S. Park. 2005. Customizing a Korean-English MT System for Patent Translation. In *Proceedings of MT-Summit 2005.*

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

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

Pang, B. and 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)*, pp. 271–278.

Pang, B., L. Lee, and S. Vaithyanathan. 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, 2002.*

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

Turney, P.D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 417–424, Philadelphia, Pennsylvania.

Valitutti, A., C. Strapparava, and O. Stock. 2004. Developing Affective Lexical Resources. *PsychNology Journal* 2.1.

Wang, B. and H. Wang. 2008. Bootstrapping Both Product Features and Opinion Words from Chinese Customer Reviews with Cross-Inducing. In *Proceedings of IJCNLP 2008*, pp. 289–295.

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

명재석 · 이동주 · 이상구. 2008. 반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템. *정보과학회 논문지 2008 년 6 월*, 392–403 쪽.

접수 일자: 2010 년 11 월 7 일
게재 결정: 2010 년 12 월 3 일