감성과학, Vol. 13, No. 3, pp.449-458, September 2010

한국어 특성을 고려한 감성 분류*

Sentiment Classification considering Korean Features

김정호**† · 김명규** · 차명훈** · 인주호** · 채수환***

JungHo Kim**† · MyungKyu Kim** · MyungHoon Cha** · JooHo In** · Soo-Hoan Chae***

한국항공대학교 컴퓨터공학과**

Department of Computer Engineering, Korea Aerospace University**

한국항공대학교 항공전자정보통신공학부***

School of Electronics, Telecommunication and Computer Engineering, Korea Aerospace University***

Abstract

As occasion demands to obtain efficient information from many documents and reviews on the Internet in many kinds of fields, automatic classification of opinion or thought is required. These automatic classification is called sentiment classification, which can be divided into three steps, such as subjective expression classification to extract subjective sentences from documents, sentiment classification to classify whether the polarity of documents is positive or negative, and strength classification to classify whether the documents have weak polarity or strong polarity. The latest studies in Opinion Mining have used N-gram words, lexical phrase pattern, and syntactic phrase pattern, etc. They have not used single word as feature for classification. Especially, patterns have been used frequently as feature because they are more flexible than N-gram words and are also more deterministic than single word. Theses studies are mainly concerned with English, other studies using patterns for Korean are still at an early stage. Although Korean has a slight difference in the meaning between predicates by the change of endings, which is 'Eomi' in Korean, of declinable words, the earlier studies about Korean opinion classification removed endings from predicates only to extract stems.

Finally, this study introduces the earlier studies and methods using pattern for English, uses extracted sentimental patterns from Korean documents, and classifies polarities of these documents. In this paper, it also analyses the influence of the change of endings on performances of opinion classification.

Keywords : sentiment, classification, opinion, pattern, endings

요 약

다양한 분야에서 인터넷 상의 방대한 양의 문서 혹은 리뷰로부터 유용한 정보를 얻고자 하는 노력이 높아짐에 따라 문서 혹은 리뷰 상의 생각 및 의견에 대한 자동 분류 연구의 필요성이 대두되었다. 이러한 자동 분류를 감성 분류라 하며, 감성 분류 연구는 크게 세 가지 단계를 가지는데, 첫 번째로 주관적인 생각이나 느낌을 표현하는 문장을 추출하기 위한 주관성 분류 연구, 두 번째로 문서 또는 문장을 긍정, 부정으로 나누는 극성 분류 연구, 그리고 세 번째로 문서 또는 문장이 어느 정도의 주관성 및 극성을 갖는지 그 강도를 구하

* 이 논문은 2010년도 (주)이엠넷의 지원을 받아 연구되었음.

† 교신저자 : 김정호 (한국한공대학교 컴퓨터공학과)

E-mail : natul2@kau.ac.kr

TEL : 02-300-0146

는 강도 분류 연구이다. 최근 의견 분류에 대한 연구들을 살펴보면, 분류를 위해 자질(Feature)로서 단일어(Single word)가 아닌 2개 이상의 N-gram 단어, 어휘 구문 패턴 및 통사 구문 패턴 등을 사용하는 것을 확인할 수 있다. 특히, 패턴은 단일어나 N-gram 단어에 비해 유연하고, 언어학적으로 풍부한 정보를 표현할 수 있기 때문에 이를 이용한 많은 연구가 이루어져 왔다. 그럼에도 불구하고, 이러한 연구들은 주로 영어에 대한 연구들이었으며, 한국어에 패턴을 적용하여 주관성을 갖는 문장을 분류하거나, 극성을 분류하는 연구들은 아직 미비하다. 한편, 한국어는 용언의 활용이 발달되어 있어, 어미의 변화가 다양하며, 그 변화에 따라 의미가 미묘하게 변화한다. 그러나 기존 한국어에 대한 의견 분류 연구들은 단어의 핵심 의미만을 파악하기 위해 어미 부분을 제거하고 어간만을 취해서 처리하여 어미에 대한 의미변화를 고려하지 못하였다.

그래서 본 연구는 영어에 적용된 패턴을 이용한 기존 방법들을 정리하고, 그 방법들 중에서 극성을 지닌 문장성분 패턴을 한국어에 적용하였다. 그리고 어미의 변화에 대한 패턴을 추출하여 이 변화가 의견 분류의 성능에 미치는 영향을 분석하였다.

주제어 : 감성, 분류, 의견, 패턴, 어미

1. 서론

각종 매체와 인터넷의 확산으로 개인들은 방대한 양의 정보를 손쉽게 제공 받을 수 있게 되었다. 또한 Web 2.0 시대를 맞아 작성자 및 그것을 읽는 사용자들이 각 기업 사이트, 블로그(blog), 포탈 게시판 등에 제품에 대한 평가 및 의견 등을 올려 서로의 생각을 공유할 수 있게 되었다. 이러한 이유로 기업에서는 자사의 이미지를 파악하고, 타 회사의 제품 및 서비스에 대한 벤치마킹을 위해 인터넷 상에 있는 의견 데이터들을 빨리 파악하길 원한다. 또한 개인들은 자신들의 관심 있는 분야에 대해 다른 사람들의 견해를 알고 싶어 하며, 자신이 물건을 구매하기 전에 다른 사람들의 사용 후기를 통해서 해당 제품에 대한 정보를 구하길 원한다. 이러한 요구로 인해, 정보 검색의 한 하위 분야로 문서 및 문장의 의견 분류(감성 분류)에 대한 연구가 활발히 진행되고 있다.

의견 분류 연구는 기본적으로 의견 표현에 사용된 단어나 절, 구문 등을 기반으로 하여 패턴을 추출하거나(정유철 등, 2008)(Ellen Riloff & Janyce Wiebe, 2003)(Ellen Riloff & Janyce Wiebe, 2004)(Kobayashi et al., 2006)(Peter D. Turney, 2002)(Zhongchao et al., 2004), 공기 관계를 계산하고(양정연 등, 2009)(Peter D. Turney, 2002), 또한 벡터 모델을 이용한 기계학습 방법(서형원 등, 2009)(황재원과 고영중, 2007)(황재원과 고영중, 2008)(Bo Pang & Lillian Lee, 2002)(Chui et al., 2006)을 이용한다. 그렇기 때문에 분류의 단서(Clue)가 되는 자질(Feature)의 선택이 매우 중요하다. 초기 연구들은 자질로서 단일어(Single word)를 사용하고, 그 위에 여러 분류기법 등을 적용하면서 성능을 증진시키기 위해 노력하였다. 그러다 점차 단일어의 한계를 느끼고, N-gram 단어, 절(Clause) 및 구문(Phrase) 패턴 등을 이용하기 시작하였다. 이러한 자질들은 단일어보다 세부적으로 의미를 결정하기 때문에 모호성을 제거할 수 있다. 그러나 현재 이러한 연구들은 주로 영어에 대해 이루어진 것이며, 한국어에 대한 의견 분류 연구는 아직 감성 단일어만을 자질로 선택하여 이용하는 연구들이 주를 이루고 있다(황재원과 고영중, 2007)(황재원과 고영중, 2008)(서형원 등, 2009).

한편, 이러한 단어 기반 시스템(Keyword based system)을 이용한 의견 분류 연구에서 말뭉치의 역할은 매우 중요하다. 이러한 인식으로 영어권에는 WordNet(Alina Andreevskaia & Sabine Bergler, 2006), SentiWordNet (Esuli A & Sebastiani F, 2006), LKB(Lexical Knowledge-Based of near-synonym differences)(Diana inkpen & Graeme Hirst, 2004) 등 다수의 프로젝트들이 존재하고 있다. 국내 경우에도 “21세기 세종 계획”과 같은 국립국어원에서 대표적으로 한글 정보화 작업을 주도하고 있으나(문화관광부와 국립국어원, 2007), 영어권과 같은 단어의 속성, 특히 감성 속성은 제공하지 않는다. 그래서 한국어의 많은 의견 분류 연구들이 영어권의 말뭉치를 한영 번역하여 이용하거나, ePinion.com, LiveJounal.com 등과 같이 인터넷 사용자의 의견 정보 및 감성 정보를 제공하는 사이트로부터 말뭉치를 구축한다(서형원 등, 2009)(양정연 등, 2009)(정유철 등, 2008)(황재원과 고영중, 2007)(황재원과 고영중, 2008).

하지만 효율적인 한영 번역 및 처리, 그리고 단어의 핵심 의미만을 취하기 위해 어간 추출(Stemming)을 하는 과정에서 한국어의 특징인 어미의 변화에 대한 미묘한 극성 변화를 고려하지 못하게 되었다.

이러한 이유들로 본 연구는 자질 선택에 있어 언어학적으로 풍부한 의미를 반영할 수 있고 모호성을 방지할 수 있는 통사적 구문 패턴을 이용하며, 어미의 변화가 극성에 미치는 영향을 고려하기 위해, 이러한 극성 변화를 시키는 어미를 따로 추출하여 패턴에서 구별할 수 있도록 하였다. 그렇게 함으로써 본 연구는 자질로서 패턴을 사용한 경우와 어미의 변화를 고려한 경우가 한국어에 대한 의견 분류의 성능에 미치는 영향을 보여주는데 그 목적이 있다.

이 논문의 구성은 다음과 같다. 2장에서 의견 분류의 각 세부연구에서 패턴을 이용한 경우들을 소개하고, 3장에서 본 연구의 시스템 구성 및 어미의 변화를 고려한 패턴에 대해서 설명하고, 4장에서는 자질로서 단일어를 사용한 경우, 패턴을 사용한 경우와 어미의 변화를 고려한 패턴을 사용한 경우의 극성 분류 성능을 비교하며 마지막으로 5장에서 결론 및 향후 연구를 제시한다.

2. 의견 분류와 패턴

의견 분류 연구는 세 가지 세부 연구 단계로 나뉠 수 있다(Andrea Esuli & Fabrisio Sebastiani, 2006). 첫 번째는 문서 및 문장이 개인의 생각이나 느낌을 표현하는 주관적 정보를 지니는지 아니면 어떤 현상이나 사실을 표현하는 객관적 정보를 지니는지 분류하는 연구이고, 두 번째는 문서 및 문장이 긍정적인 의미를 지니는지 아니면 부정적인 의미를 지니는지 그 극성을 분류하는 연구이다. 마지막으로 세 번째는 주관성 및 극성이 분류된 문서 및 문장에 대해 그 강도를 분류하는 연구이다. 그 동안 이러한 세부 연구 단계들에 대해서 여러 분류 방법들이 제안되었으며, 이 분류 방법들은 크게 학습 말뭉치를 이용하는 기계학습 방법과 Pointwise Mutual Information(PMI) 등과 같이 통계적 확률 계산을 이용하는 방법으로 나뉜다(Jonathon Read, 2004).

한편, 분류 기법에서 사용하는 자질의 선택에 있어서는 단일어, N-gram 내용어, 패턴 등을 사용해 왔으나 의견 데이터들의 특성상 단일어로는 충분히 그 의미를 표현할 수 없기 때문에 N-gram 내용어나 구문 패턴을 자질로 사용하는 연구들이 활발하게 이루어졌다. 그래서 본 연구는 구문 및 문장 패턴에 중점을 두고, 다음으로 이러한 패턴을 사용한 관련 연구들을 소개하고자 한다.

2.1. 패턴을 이용한 의견 분류

기존 단일어를 사용한 의견 분류의 연구의 한계가 드러나면서 의견 데이터의 의미를 더 정확히 표현하기 위한 자질 선택의 필요성이 대두되었다. 이에 단일어뿐만 아니라 의미 표현의 모호성을 줄이기 위해 N-gram 단어 및 패턴을 사용한 연구가 이루어졌다.

의견 분류를 위해, Turney(2002)는, 첫 번째 단어는 형용사 또는 부사이고, 두 번째 단어는 내용어인 두 개의 연속적인 단어를 한 쌍으로 패턴을 구성하여 이를 추출하고, 이 구문 패턴에 대해서 PMI(Pointwise Mutual Information)를 계산하여 추출한 각 패턴의 극성 강도를 계산하였다.

아래 표 1은 인터넷 사용자들이 긍정적으로 분류한 리뷰에 대해서 분석한 결과이다. 여기서 각 품사 태그를 살펴보면, JJ는 형용사, RB는 부사, NN은 명사, VB는 동사를 나타낸다. 이렇게 품사 태그를 통해 형용사와 부사를 포함한 구문 패턴을 추출하고, 이 패턴과, 긍정 문서에서 많이 나타나는 단어에 대한 PMI에서 부정 문서에서 많이 나타나는 단어에 대한 PMI를 빼서 극성 강도를 계산한다. 극성 강도가 양의 값을 가지면 구문 패턴은 긍정의 의미를 지니고, 음의 값을 가지면 부정의 의미를 지닌다.

Turney(2002)가 제안한 이 방법은 긍정문서 170개, 부정문서 240개로 이루어진 총 410개의 문서에 대해서 최대 84%의 정확도를 보여주었다.

한편, Fei 등(2004)은 Turney(2002)의 방법을 좀 더 확장하여, 형용사와 부사를 포함한 구문 패턴뿐만 아니라, 접속사, 전치사 등의 품사를 포함하는 통사적 패턴을 사용하였다. 아래 표 2는 Fei 등(2004)이 사용한 품사 태그이며, 각 품사 태그들은 긍정과 부정을 구분할 수 있도록 정의되었다.

이러한 극성 품사 태그를 이용하여, 개체를 주어로 하고, 극성 품사 태그를 적어도 하나 반드시 포함하는 패턴을 추출하여, 문장의 극성을 판별하였다. 아래 표 3는 Fei 등(2004)이 50개의 긍정 문서와 50개의 부정 문서에서 추출한 패턴을 나타내고 있으며, 이 패턴들의 긍정 문서에서 나타나는 빈도수와 부정 문서에서

나타나는 빈도수를 이용하여 계산된 극성 강도를 보여주고 있다.

표 1. 긍정 리뷰(Review)에 대해서 Turney가 제안한 방법으로 극성을 분류한 결과(Turney, 2002)

Extracted Phrase	POS Tags	Semantic Orientation
online experience	JJ NN	+2.253
low fees	JJ NNS	+0.333
local branch	JJ NN	+0.421
small part	JJ NN	+0.053
online service	JJ NN	+2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	+1.288
well other	RB JJ	+0.237
inconveniently	RB VBN	-1.541
located		
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
Average Semantic Orientation		+0.322

표 2. Fei, Liu, Wu가 사용한 극성 품사 기호(Zhongchao Fei et al., 2004)

Type	Tag	Example
Subject to review	n	team
Positive adjective	aj	good
Negative adjective	dj	bad
Positive adverb	ad	perfectly
Negative adverb	dd	poorly
Positive noun	an	star
Negative noun	dn	garbage
Positive verb	av	achieve
Negative verb	dv	frustrate
Special Prep	np	without
Special Conj	nc	however

Fei 등(2004)이 제안한 이 방법은 긍정문서 170개, 부정문서 150개로 이루어진 총 320개의 스포츠 리뷰(Review)에 대해서, 86%의 정확도를 보여주었다.

또한, Cui 등(2006)은 순수 N-gram 어절(N≥3)을 이용하여 디지털 카메라, 노트북, PDA, MP3 플레이어 등의 전자제품에 대한 온라인 상품평의 극성을 분류하였다. 이들은 대용량의 문서로부터 1-gram부터 6-gram까지의 데이터를 추출하고 χ^2 가중치 기법을 적용하여 정제된 N-gram 어절을 자질로 사용하였다. 그리고 여러 분류 기법을 사용하여 그 성능들을 비교하였으며, 최대 90%의 분류 성능을 보여 주었다.

표 3. 극성을 지닌 구문 패턴(Zhongchao Fei et al., 2004)

구문 패턴(40)	긍정 빈도수	부정 빈도수	극성 강도
n + aj	108	24	+1.0540
n + dj	37	119	-1.1680
...	...	...	...
n + dd + av +aj	1	11	-2.3979
n + ad + dv + aj	12	6	+0.6931
n + ad + dv + aj	8	0	+4.1972
n + ad + dv + an	2	20	-2.3026
...	...	...	...

2.2. 패턴의 종류 별 특성

지금까지 살펴 본 의견 분류를 위한 패턴들은 그 구성에 따라 세 가지 종류로 나뉠 수 있다. 패턴 전체가 어휘로만 구성된 N-gram 어절 패턴(N-gram Word), 어휘와 품사로 구성된 어휘-통사적 패턴(Lexical-Syntactic Pattern), 그리고 극성을 표시하되 순수 품사로만 구성된 극성 통사적 패턴(Sentiment Syntactic Pattern)이다.

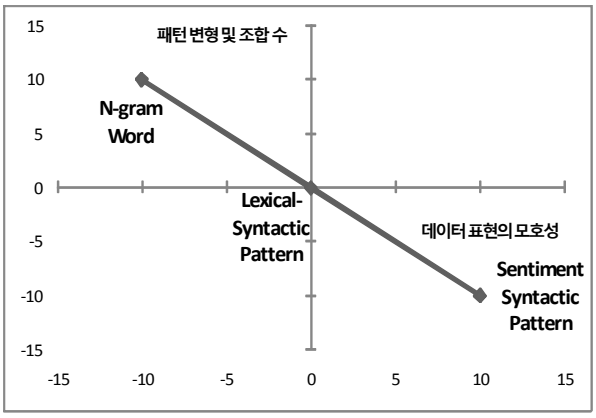

그림 1. 패턴의 종류 별 특성

이러한 패턴들은 각각의 특징이 있는데, 위 그림 1에서 보이는 바와 같이 N-gram 어절 패턴은 단어들이 모여서 이룬 전체 의미를 명확히 표현하지만 변형 및 조합 수가 너무 많다는 것이고, 반대로, 극성 통사적 패턴은 그 구성 요소가 품사들로만 이루어져 있기 때

문에 패턴 조합의 수는 적지만 극성을 나타내지 않는 감성 단어들로 이루어진 극성 구문에 대해서 명확히 표현하지 못한다는 것이다. 마지막으로 어휘-통사적 패턴은 이 두 패턴을 절충하여 데이터 표현을 명확하게 하되 그 패턴의 변형 및 조합 수도 줄였다.

3. 어미의 변화를 고려한 통사적 구문 패턴

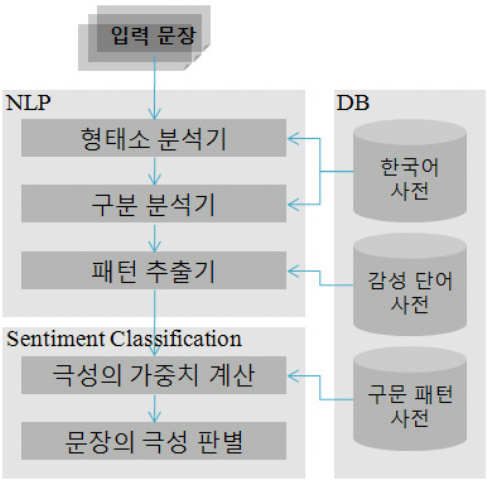

그림 2. 전체 시스템 구성도

3.1. 전체 시스템 구성도

본 연구에서 실험을 위해 구성한 시스템은 문장의 극성을 판별하는 것을 목표로 하며, 그 구성은 아래 그림 2와 같다. 이 시스템은 우선 학습 문장에 대해 형태소 분석, 구문 분석 등의 전처리 작업을 하고, 패턴 추출기와 감성사전을 통해 의견을 표현하는 구문 패턴을 추출한 후, 패턴 사전에 저장한다. 그 후 임의의 입력 문장에 대해 똑같이 전처리 작업 후, 추출되는 패턴 후보의 극성을 사전에서 검색하여, 문장의 극성을 판별한다.

3.2. 패턴의 구성

본 연구는 분류에 사용할 자질로서, 의견을 표현하는 통사적 구문 패턴(Syntactic Phrase Pattern)을 사용한다. 이 구문 패턴은 문장의 주관성을 판단하는데 유용하고, 극성을 판별하는데 좋은 성능을 보여주기 때문이다. 이 때 고려해야 할 점은 한국어는 표 4에 나타난 바와 같이 하나의 어절이 여러 형태소들로 이루어져 있기 때문에 어절을 이루는 모든 형태소들의 품사를 패턴에 적용하면, 생성될 수 있는 패턴 조합의 수가 너무 많아져 계산 양이 증가하고 정확률이 떨어질 수 있다는 것이다.

표 4. 한국어와 영어의 한 어절을 이루는 형태소들의 품사 비교

한국어	영어	문장성분
명사+조사 대명사+조사 수사+조사 동사+명사형 전성어미 형용사+명사형 전성어미	명사 대명사 수사 동명사	주어, 목적어
동사+어미 형용사+어미 명사+동사파생접미사+어미 명사+형용사파생접미사+어미	동사 be 동사+형용사	서술어
관형사 명사+관형격 조사 동사+관형사형 전성어미 형용사+관형사형 전성어미	형용사	관형어
부사 명사+부사격 조사 동사+부사형 전성어미 형용사+부사형 전성어미	부사	부사어

그래서 본 연구는 어절을 구성하는 형태소들의 품사가 아니라 문장성분을 패턴의 구성 요소로 사용한다. 문장 구문의 패턴을 표현함에 있어서 한 어절의 세부적인 구성 성분까지 고려할 필요는 없고, 오히려 문장 내에서 어절의 역할을 나타내는 요소가 필요한데, 이러한 점에 있어서 문장성분의 사용은 합당하다. 게다가 한국어는 격조사가 발달되어 있어 영어와 달리 문장성분을 분석하기 용이하다.

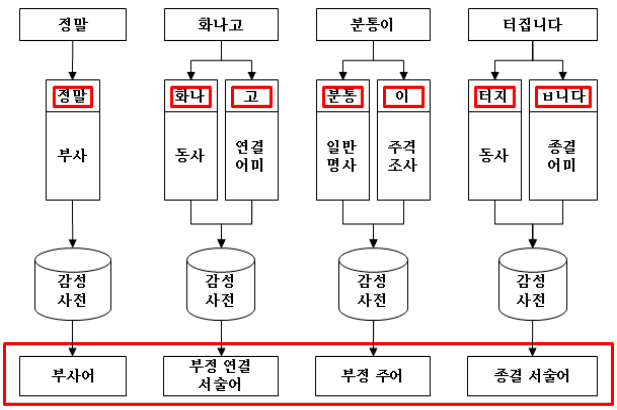

그림 3. 문장성분으로 구성된 구문 패턴 추출 예제

앞 그림 3은 패턴을 추출하는 예를 나타낸다. 문장에 대해서 형태소 분석 및 감성사전 검색을 통해 극성을 나타내는 문장성분 패턴을 추출한다.

구체적으로 설명하면, 부사는 부사어, 연결어미는 연결 서술어, 주격조사는 주어, 종결어미는 종결 서술어를 나타내고, 동사 '화나-'와 명사 '분통'은 감성 사전에서 극성을 나타내는 단어이기 때문에 이 단어를 포함하는 문장성분에 극성을 표시하여 각각 '부정 연결 서술어', '부정 주어'로 나타내었다.

3.3. 어미에 대한 고찰

의견 분류 연구들 중, 많은 논문들이 영어권의 WordNet, SentiWordNet과 같은 유의어 사전(Thesaurus) 및 관계 형성 사전에서 명사, 형용사, 동사, 부사의 단어를 가지고 와 한글에 적용하였는데 이 과정에서 어미에 대한 처리가 간과되어 왔다. 한국어는 용언의 활용이 빈번하기 때문에, 모든 활용을 데이터로 처리하기 어려워 형태소 분석을 통해 용언을 기본형으로 복원시킨다. 즉, 활용된 용언에서 어미 부분을 제거하고, 어간만을 추출(Stemming)하는 것이다.

표 5. 용언의 활용 예제

종류	한국어	영어
기본형	좋아하다	like
현재	좋아한다	like/likes
과거	좋아했다	liked
현재완료	좋아했었다	have liked
현재완료	좋아해왔다	have liked
명사형	좋아함	liking
관형사형	좋아하는	N/A
부사형	좋아하게	N/A
명령형	좋아해라	like
청유형	좋아하자	N/A
감탄형	좋아하구나	N/A
의문형(현재)	좋아하니	N/A
의문형(과거)	좋아했니	N/A
역접	좋아하지만	N/A
역접	좋아하나	N/A
바램	좋아했으면	N/A
이유	좋아해서	N/A

위 표 5를 보면, 영어의 용언(동사) 변형은 단수, 복수 및 시제를 구분하기 위한 문법적 기능 때문에 사용되지만, 한국어의 용언(동사, 형용사) 변형은 문법적 기능 이상을 표현하기 위해 사용된다. 그렇기 때문에 이와 같은 어미를 간과한 채, 어간만을 추출하여 의미 분석을 하게 되면, 많은 언어 정보를 놓치게 될 것이다.

본 연구에서는 어미의 변화가 일으키는 여러 영향 중 특히 역접의 의미를 갖는 어미로 인해 발생하는 극성의 변화에 대해 중점을 둔다. 극성 분류 측면에서만 볼 때, 단어의 형태나 의미의 변화보다 긍정, 부정의 극성의 변화가 중요한 관심사이기 때문이다. 표 6는 이러한 어미들로 인해, 감성 단어들의 고유의 극성과 반대되는 극성을 갖는 구문들을 나타낸다.

표 6. 어미에 따른 극성의 변화

감성 단어	어미	예 문	문장 극성
망설이-(부정)	-는데	엠피쓰리는 처음 사용해보는 거라서 구입하기까지 많이 망설였는데 저렴한 가격에도 제품 성능이 좋은 것 같네요.	부정
부족하-(부정)	-지만	동영상 보기에는 화소수가 좀 부족하지만 뮤직비디오 정도 담아 보기에는 딱 좋군요.	긍정
좋-(긍정)	-면	단지 하나 안 좋다면 폴더를 이전 폴더로 이동하고 싶을 때 버튼을 하나하나 다 눌러서 찾아야 한다는 것이에요.	부정
만족하-(긍정)	-나	색상과 가격은 만족하나 이어폰 삽입구가 MP3의 옆 부분에 있어 불편해요.	부정
망하-(부정)	-는지	그 회사는 이미 망했는지 없어졌더군요.	부정

가령, 감성 단어 '망설이-', '부족하-'는 각각 부정, 긍정의 극성을 지니지만, 이러한 단어들과 어미 '-는데', '-지만'과 함께 이루어진 문장의 극성은 반대의 극성을 나타냄을 확인할 수 있다. 그 밖에 역접의 의미를 지니지 않는 어미들에 대한 예도 나타내었다.

이렇듯 극성 분류 연구에서 어미의 변화는 고려해야 할 요소이다. 그래서 본 연구에서는 앞에서 언급했던 문장성분으로 구성된 통사적 구문 패턴에 이러한 어미들을 표시한다. 그림 4를 보면 문장성분으로 구성된 통사적 구문 패턴에 감성사전에 등록된 극성을 변화시키는 어미 '-지만'을 같이 표기한 것을 확인할 수 있다. 이렇게 함으로써, 어미 별로 패턴의 극성에 미치는 영향을 확인할 수 있다.

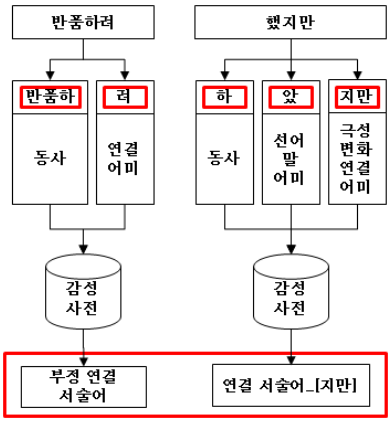

그림 4. 어미의 변화를 고려한 구문 패턴 추출 예제

3.4. 패턴의 극성 강도

패턴의 극성 강도를 직접 사람이 부여할 수도 있지만 이는 그 기준이 주관적이어서 같은 패턴에 대해서 다르게 해석하여 극성을 부여할 수 있기 때문에 본 연구에서는 Fei 등(2004)이 사용한 기계학습 방법을 이용한다. 이 방법은 패턴의 긍정문서와 부정문서에 나타나는 각 빈도수를 구하고, 이를 아래 공식을 통해 계산하여 구한다. 여기서 구한 극성 강도 W_i가 양수이면 긍정을, 음수이면 부정을 나타낸다.

T_{p_i}는 패턴 i의 긍정문서에 나타난 빈도수, T_{n_i}는 패턴 i의 부정문서에 나타난 빈도수라 할 때,

극성 강도

$$W_i = \begin{cases} \log(\frac{T_{p_i}}{T_{n_i}}) & T_{p_i} \neq 0 \quad \text{and} \quad T_{n_i} \neq 0 \\ C + \log(\frac{T_{p_i}+1}{T_{n_i}+1}) & T_{p_i} = 0 \quad \text{or} \quad T_{n_i} = 0 \end{cases}$$

이다.

3.5. 기호 및 실제 패턴

본 연구에서 사용하는 패턴의 구성 요소는 문장성분이며, 이러한 문장성분을 표현하기 위한 기호는 표 7과 같다. 그리고 극성 문장성분을 표시하기 위해 문자 'P' 및 'N'을 기호 앞에 붙인다. 가령, 긍정의 극성을 갖는 주어인 경우 'PS', 부정의 극성을 갖는 주어인 경우 'NS'라고 표시한다. 한편, '아니-', '않-', '못하-', '말-' 등과 같은 부정어를 표시하기 위해, 문자 'I'를 서술어 기호 앞에 붙인다. 또한 어미의 변화를 고려하기 위해, 서술어 기호 뒤에 고려해야 할 어미를 붙여 표기한다. 예를 들어, '-지만'이라는 어미를 고려하기 위한 연결 서술어인 경우, 'CP_지만'이라고 표시한다. 여기서 연결 서술어와 종결 서술어란 것을 정의하였는데, 연결 서술어란 연결 어미로 끝나는 문장성분이며, 종결 서술어란 종결 어미로 끝나는 문장성분을 말한다.

표 7. 문장성분 기호

서술어	기호
주어	S
연결 서술어	CP
종결 서술어	FP
관형어	Adn
부사어	Adv

표 8. 어미를 고려한 극성 구문 패턴 및 예제

극성 구문 패턴	긍정 빈도수	부정 빈도수	극성	구문 예제
PCP CP_는데	1	5	−1.6094	기대하고 샀는데
Adv NS CP	0	3	−1.3862	약간 흠집이 있어서
NCP FP	1	9	−2.1972	반품하고 싶네요
Adv NFP	3	35	−2.4567	조금 아쉽네요
Adv PCP	113	25	+1.5085	왕 친절하고
S PCP IFP	1	0	+0.6931	소리가 좋지 않아요
NCP CP	4	9	−0.8109	불편해 전화드리고
PCP PFP	62	2	+3.4339	이쁘고 좋은데요
S PCP	87	25	+1.2470	디자인은 좋은데
NCP Adv FP	0	1	−0.6931	반품하고 다시 받았구요

위 표 8은 지금까지 언급한 패턴의 추출 방법 및 극성 강도 계산 방법에 의해 구해진 패턴의 일부와 이러한 패턴에 의해 추출된 구문 예제를 나타낸다. MP3P 관련 댓글 3000개의 문서로부터 학습된 패턴의 수는 총 1429개이며, 이 중 긍정 극성을 갖는 패턴의 수는 753개이고, 부정의 극성을 갖는 패턴의 수는 676개이다.

4. 실험 및 평가

4.1. 데이터 구성 및 실험 방법

데이터를 얻기 위해 Yahoo 쇼핑 사이트에서 MP3P/PMP에 관한 문서를 Crawler를 통해 추출하였다. Yahoo 쇼핑 사이트는 다른 여러 쇼핑몰 사이트에서 제공하는 댓글들을 한 곳에 모아 제공하기 때문에 댓글 데이터를 수집하기 용이하다. 실험을 위해 3000개의 문서에 있는 문장들을 학습 데이터(Training Data)로 사용하였고, 520개의 문장을 실험 데이터(Test Data)로 사용하였다. 그 후, 감성 사전에 저장되어 있는 단일어를 자질로 사용하는 경우, 극성 구문 패턴을 자질로 사용하는 경우, 어미를 고려한 극성 구문 패턴을 자질로 사용하는 경우에 대해, 각각 같은 학습 데이터와 실험 데이터를 적용하여 분류 성능을 비교하였다.

표 9. 데이터 구성

구 분	긍정 데이터	부정 데이터
학습 데이터 (Training Data)	3000 문서	3000 문서
실험 데이터 (Test Data)	260 문장	260 문장

4.2. 성능 평가 방법

성능의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확률(Precision)과 재현율(Recall)을 사용하였다. 이 때, 분류 결과 패턴이 발견되지 않아 분석이 안 된 문장들은 제외하고, 패턴이 발견된 문장들에 대해서 분류의 성능을 계산하였다. 이는 본 연구가 의견 데이터의 인식률 향상에 목적을 두는 것이 아니라, 자질 선택에 따른 극성 분류 성능의 변화에 관심을 두기 때문이다.

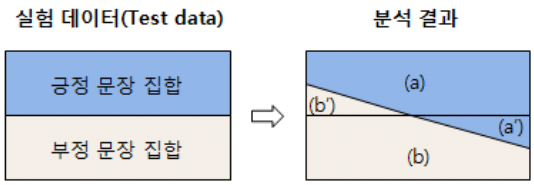

그림 5. 정확도 계산 방법

$$\text{긍정 } precision = \frac{a}{a+a'} \qquad \text{부정 } Precision = \frac{b}{b+b'}$$

$$\text{긍정 } Recall = \frac{a}{a+b'} \qquad \text{부정 } Recall = \frac{b}{a'+b}$$

$$Accuracy = \frac{a+b}{a+a'+b+b'}$$

(a) 긍정으로 올바로 분류된 긍정 문장 수
(b) 부정으로 올바로 분류된 부정 문장 수
(a′) 부정으로 잘못 분류된 긍정 문장 수
(b′) 긍정으로 잘못 분류된 부정 문장 수

4.3. 실험 결과

다음 표 10과 그림 6을 보면, 자질의 선택에 따라 문장의 극성 분류 정확도가 달라지며, 감성 단일어를 사용한 경우보다 극성 구문 패턴을 사용한 경우가 정확도가 높고, 이 극성 구문 패턴에서 어미를 고려한 경우 더욱 분류 정확도가 높다는 것을 알 수 있다. 실제 실험 데이터를 살펴보았을 때, 문장 "한 가지 재생할 때 소리가 좋지 않아요."에 대해서 감성 단일어를 자질로 사용한 경우 긍정 단어 '좋-'로 인해 이 문장을 긍정으로 분류하였다. 그러나 구문 패턴을 사용한 경우, 패턴 [PCP IFP]에 의해 구문 '좋지 않아요'를 인식하고 문장을 부정으로 분류하였다. 한편, 문장 "물건은 좋으나..."에 대해서 일반 구문 패턴을 사용한 경우, 패턴 [S PFP]에 문장을 긍정으로 인식하였지만, 어미를 고려한 구문 패턴을 사용한 경우, 패턴 [S PFP_지만]에 의해 문장을 부정으로 인식하였다. 이 결과로 한국어에 대해 자질로서 단일어보다 구문 패턴이 더욱 풍부한 언어 정보를 반영하여 모호성을 줄일 수 있다는 것과 어미 또한 무시해서는 안 될 중요한 요소라는 것을 알 수 있다.

표 10. 실험 결과

자질 선택 / 성능 평가 항목	감성 단일어	극성 구문 패턴	극성구문 패턴 (어미고려)
긍정 문장 정확률(Precision)	72.08%	77.78%	79.41%
부정 문장 정확률(Precision)	86.67%	89.25%	90.32%
긍정 문장 재현률(Recall)	92.27%	94.15%	94.74%
부정 문장 재현률(Recall)	58.43%	64.34%	66.66%
총 정확도(Accuracy)	76.62%	81.33%	82.83%

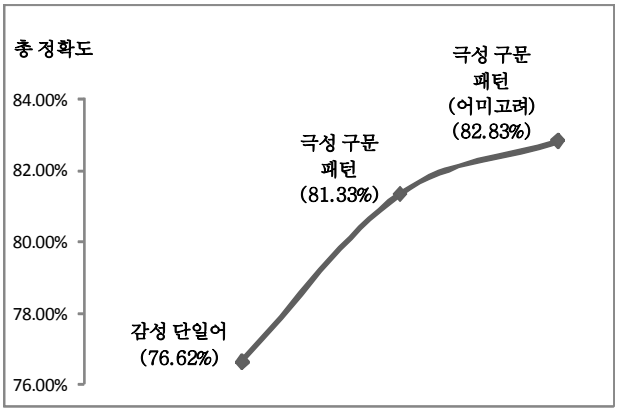

그림 6. 자질에 따른 문장의 극성 분류 정확도

5. 결론 및 향후 연구

본 논문에서는 문장의 극성 분류를 위해 자질로서 통사적 품사패턴을 사용하였으며, 어미의 변화가 극성 분류에 어떠한 영향을 미치는지 분석하였다. 그 결과 어미의 변화를 고려한 경우, 분류 정확도의 성능이 향상됨을 확인할 수 있었다. 비록, 성능 향상이 두드러지게 나타나지는 않았지만, 이는 본 연구에서 극성을 반전시키는 어미들인 '-는데', '-지만', '-나' 등의 몇몇 어미들만을 가지고 실험을 하였기 때문이며, 이러한 극성을 반전시키는 어미들의 데이터 수를 늘리고 의미를 강조시키거나 변화시키는 어미들도 고려한다면 좀 더 좋은 성능 향상을 기대할 수 있을 것이다.

한편, 본 실험에서 극성 분류의 분석 단위를 문장으로 하였는데, 의견 분류 연구에서 보통 문장 단위로 실험하는 경우 그 목적은 주관성 분류이며, 단어 및 문서 단위로 실험하는 경우 그 목적은 극성 분류이다. 즉, 문장 단위에 대해서 주관성 분류를 한 후, 이러한 문장들을 기반으로 문서의 극성을 분류하는 것이 일반적인 흐름이라는 것이다. 현재 본 실험에서는 문장의 극성을 분류하는 것으로 끝냈지만, 앞으로 더 나아가 각 문장의 극성들을 어떻게 조합하여 문서의 극성 분류할지 고려해야 할 것이다. 이를 위해서 문장 간의 순접, 역접 등의 관계를 고려하는 연구가 필요하다.

또한 본 실험에서 사용한 패턴은 어미 표시를 제외하고는, 극성을 지니는 문장성분들로만 이루어진 통사적 패턴(Syntactic Pattern)이었다. 그러나 이러한 순수 통사적 패턴은 극성을 지니지 않는 단어들로 구성된 구문이나 문장에 대해서는 그 의미를 파악하는데 모호성이 존재한다. 가령, "그 물건이 마음에 든다."라는 문장의 경우 긍정의 극성을 갖지만, 그 문장을 구성하는 단어들 중 감성 단어는 존재하지 않기 때문에 순수 통사적 패턴으로는 극성을 분류하기 어렵다. 그래서 추후 연구에서는 어휘-통사적 패턴(Lexical-Syntactic Pattern)을 이용하고자 한다. 이러한 패턴을 이용한다면, 부정극어 및 '않-', '못하-' 등과 같은 보조용언과 상대적 극성을 갖는 단어들, 그리고 여러 어미의 변화를 좀 더 명확히 표현할 수 있을 것이다.

마지막으로, 온라인 상품평 및 댓글에는 띄어쓰기 오류 및 철자 오류가 많고, 이모티콘, 줄임말, 신조어 등이 빈번하게 사용된다. 이러한 이유로 규칙 기반의 정형화된 패턴이 잘 매칭(Matching)되지 않는 문제점이 존재한다. 그래서 앞으로 이러한 온라인 문서에 패턴을 적용하기 위해서 오류 데이터(Noisy Data)에 대한 처리가 요구된다.

참고문헌

권혁철, 최준영 (1992). 단일화 기반 의존 문법을 이용한 한국어 분석기. *정보과학회논문지, 19(5),* 467-476.

김진동, 임희석, 임해창 (1997). Twoply HMM : 한국어 특성을 고려한 형태소 단위의 품사 태깅 모델. *정보과학회논문지, 24(12),* 1502-1512.

문화관광부, 국립국어원 (2007). 21세기 세종계획 : 최종 성과물 안내서.

서형원, 김형철, 김재훈 (2009). 기계학습 방법을 이용한 댓글의 감정 인식. *한국마린엔지니어링 학회 학술대회 논문집,* 373-374.

이용훈, 이종혁 (2008). 기계학습 기법을 이용한 한국어 구문분석. *한국정보과학회, 35(1),* 285-288.

양정연, 명재석, 이상구 (2009). 상품 리뷰 요약에서의 문맥정보를 이용한 의견 분류 방법. *한국인지과학회, 36(4),* 254-262.

정유철, 최윤정, 맹성현 (2008). 감정 기반 블로그 문서 분류를 위한 부정어 처리 및 단어 가중치 적용 기법의 효과에 대한 연구. *한국인지과학회, 19(4),* 477-497.

최선화, 박혁로 (2003). 한국어 확률 의존문법 학습. *한국정보과학회, 30(1),* 513-515.

황명진, 강미영, 권혁철 (2006). 규칙과 어절 확률을 이용한 혼합 품사 태깅 모델. *한국정보과학학회, 33(2),* 11-15.

황재원, 고영중 (2007). 효과적인 감정 자질을 이용한

한국어 문서 감정 분류 시스템. *한국정보과학회, 34(2),* 60-61.

황재원, 고영중 (2008). 감정 자질을 이용한 한국어 문장 및 문서 분류 시스템. *한국정보과학회 논문지, 14(3),* 336-340.

황재원, 고영중 (2008). 문장 감정 강도를 반영한 개선된 자질 가중치 기법 기반의 문서 감정 분류 시스템. *한국정보과학회, 36(6),* 491-497

Alina, A. & Sabine, B. (2006). Mining WordNet for fuzzy sentiment Sentiment tag extraction from WordNet glosses. *11th Conference of the European Chapter of the Association for Computational Linguistics,* 209-216.

Andrea, E. & Fabrisio, S. (2006). Determining Term Subjectivity and Term Orientation for Opinion Mining. the 11rd Conference of the European Chapter of the Association for Computational Linguistics, 193-200.

Bo, P. & Lillian, L. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. The Conference on Empirical Methods on Natural Language Processing, 79-86.

Cui, H., Mittal, V. & Datar, M. (2006). Comparative Experiments on Sentiment Classification for Online Product Reviews. proceedings of the national conference on artificial intelligence, 21(2), 1265-1270.

Diana, I. & Graeme, H. (2006). Building and Using a Lexical Knowledge Base of Near-Synonym Differences. Computational linguistics - Association for Computational Linguistics, 32(2), 223-262.

Ellen, R. & Janyce, W. (2003). Learning Extraction Patterns for subjective Expressions.

Ellen, R. & Janyce, W. (2004). Exploiting Subjectivity Classification to Improve Information Extraction. In Proceedings of the 20th Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 22-24.

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

Hellwin, P. (1995). Dependency Unification Grammar. Computational linguistics, 21(1), 95-102.

Jonathon, R. (2004). Recognising Affect in Text using Pointwise-Mutual Information. University of Sussex.

Jung, H. S. (1987). Korean Phrase Structure Grammar. Proceeding of the First Natural Language Processing Workshop, SIGAI of Korean Information System Society, 3-37.

Kobayashi, N., Unui, K., Matsumoto, Y., Tateishi, K. & Fukusmia. (2004). Collecting Evaluative Expressions for Opinion Extraction. In Proceedings of the First International Joint Conference on Natural Language Processing, 584-589.

Peter, D. T. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), 417-424.

Theresa, W., Janyce, W. & Rebecca, H. (2004). Just how mad are you? finding strong and weak opinion clauses. Proceeding of the 19th national conference on Artifical intelligence, 761-767.

Yahoo 쇼핑 사이트, http://kr.product.shopping.yahoo.com/.

Zhongchao, F., Jian, L. & Gengfeng, W. (2004). Sentiment Classification Using Phrase Patterns. In The Fourth International Conference on Computer and Information Technology (CIT'04), 1147-1152.

원고접수 : 10.06.19
수정접수 : 10.09.07
게재확정 : 10.09.10