강건한 한국어 상품평의 감정 분류를 위한 패턴 기반 자질 추출 방법

(A Robust Pattern-based Feature Extraction Method for Sentiment Categorization of Korean Customer Reviews)

신 준 수 † 김 학 수 ††
(Junsoo Shin) (Harksoo Kim)

요 약 기계 학습 기반의 많은 감정 분류 시스템들은 문장으로부터 언어적 자질을 추출하기 위하여 형태소 분석기를 사용한다. 그러나 온라인 상품평에는 많은 띄어쓰기 오류 및 철자 오류가 포함되어 있어서 일반적으로 형태소 분석기가 좋은 성능을 내기 어려우며, 기반 시스템의 낮은 성능은 감정 분류 시스템의 성능하락을 초래한다. 이러한 문제를 해결하기 위하여 본 논문에서는 어절 패턴과 음운 패턴의 최장 일치 매칭(matching)에 기반한 자질 추출 방법을 제안한다. 두 종류의 패턴은 대용량의 품사 부착 말뭉치로부터 자동으로 구축된다. 어절 패턴은 명사, 동사와 같은 내용어를 포함하는 어절들로 구성되며, 음운 패턴은 동사나 형용사와 같은 용언의 초성과 중성의 쌍으로 구성된다.

음운 패턴에 초성과 중성만을 사용한 이유는 철자 오류에 영향을 덜 받기 때문이다. 제안 방법을 평가하기 위하여 SVM(Support Vector Machine)을 기계 학습기로 사용하는 감정 분류 시스템을 구현하였다. 한국어 상품평에 대한 실험에서 제안 방법을 자질 추출 모듈로 사용하는 감정 분류 시스템이 형태소 분석기를 사용하는 것보다 우수한 성능을 보였다.

키워드 : 상품평 감정 분류, 최장일치 기반 자질 추출, 어절 패턴, 음운 패턴

Abstract Many sentiment categorization systems based on machine learning methods use morphological analyzers in order to extract linguistic features from sentences. However, the morphological analyzers do not generally perform well in a customer review domain because online customer reviews include many spacing errors and spelling errors. These low performances of the underlying systems lead to performance decreases of the sentiment categorization systems. To resolve this problem, we propose a feature extraction method based on simple longest matching of Eojeol (a Korean spacing unit) and phoneme patterns. The two kinds of patterns are automatically constructed from a large amount of POS (part-of-speech) tagged corpus. Eojeol patterns consist of Eojeols including content words such as nouns and verbs. Phoneme patterns consist of leading consonant and vowel pairs of predicate words such as verbs and adjectives because spelling errors seldom occur in leading consonants and vowels. To evaluate the proposed method, we implemented a sentiment categorization system using a SVM (Support Vector Machine) as a machine learner. In the experiment with Korean customer reviews, the sentiment categorization system using the proposed method outperformed that using a morphological analyzer as a feature extractor.

· 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2010-0009875).
또한 이 논문의 일부는 한국전자통신연구원 위탁연구과제의 지원을 받아 수행되었음

† 학생회원 : 강원대학교 컴퓨터정보통신공학과
nlpsjs@kangwon.ac.kr
†† 종신회원 : 강원대학교 컴퓨터정보통신공학과 교수
nlpdrkim@kangwon.ac.kr

논문접수 : 2010년 7월 19일
심사완료 : 2010년 9월 27일

정보과학회논문지: 소프트웨어 및 응용 제37권 제12호(2010.12)

Key words : Sentiment categorization of customer reviews, Feature extraction based on longest matching, Eojeol pattern, Phoneme pattern

1. 서 론

최근 인터넷 쇼핑몰이 대중화됨에 따라 상품 구매 후기와 같은 상품평의 양이 급격하게 증가하고 있다. 이러한 상품평은 물건에 대한 사용자의 다양한 의견이 포함되어 있기 때문에 해당 제품의 구매에 큰 영향력을 미치는 중요한 정보이다. 그러나 방대한 양의 상품평을 확인하는 데는 많은 시간이 소요된다는 문제점을 가지고 있다. 이러한 문제를 해결하기 위하여 감정 분류에 대한 연구가 활발히 진행되고 있다[1-4]. 감정 분류에 대한 기존의 시스템들은 대부분 형태소 분석기와 같은 언어 분석기를 활용하여 자질을 추출하고 이것을 기반으로 기계학습을 하여 감정을 분류한다. 그러나 온라인(online) 상품평에는 자의적이든 그렇지 않든 다수의 띄어쓰기 오류와 철자 오류가 포함되어 있으며, 이러한 오류들은 형태소 오분석에 따른 감정 분류 시스템의 성능 저하를 초래한다. 이러한 문제를 해결하기 위하여 형태소 분석기를 사용하지 않고 어절 및 초성 패턴의 최장 일치법을 기반으로 하는 자질 추출 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 감정 분류와 관련된 기존 연구를 기술하고, 3장에서는 제안 시스템에 대해서 설명한다. 4장에서는 시스템의 성능을 평가하며, 5장에서 결론을 맺는다.

2. 관련 연구

감정 분류에 대한 연구는 자질 추출 방법에 대한 관점과 기계 학습 방법에 대한 관점에서 나누어 볼 수 있다[1-4]. 자질 추출 방법으로는 사전, 스니핏(snippet), 전문가와 같은 외부자원을 이용하여 자질에 가중치를 부여하거나[1-3] 슬라이딩 윈도우(sliding window) 기반의 형태소 바이그램(bigram)을 이용하여 새로운 자질을 생성하는 연구가 있다[4]. 이러한 연구들은 모두 형태소 분석기를 이용하고 있기 때문에 철자 오류에 민감하다는 단점이 있다. 형태소 분석기를 이용하지 않는 자질 추출 방법으로는 음절 n-그램(n-gram)을 이용하는 연구와 음절 커널을 이용하는 연구가 있다[5,6]. 음절 n-그램 방법은 형태소 분석기를 이용하는 방법에 비해 철자 오류에 덜 민감하지만 '좋아요', '조아용'과 같이 같은 의미지만 형태가 다른 자의적인 철자 오류가 발생하면 동일 의미 자질의 정보량이 분산되는 문제가 발생한다. 음절 커널을 이용하는 연구는 동일 의미의 자질들을 추출할 수 있지만 계산량이 많다는 단점이 있다. 기계 학습 방법으로는 CRFs(Conditional Random Fields)나 SVM(Support Vector Machine)를 이용한 연구가 있다[1,4]. 다양한 기계 학습 방법론을 감정 분류에 활용해 보는 것은 좋은 시도라고 생각된다. 그러나 기계 학습은 근본적으로 입력 자질에 따라 성능이 좌우되기 때문에 자질 추출에 대한 연구가 어떤 기계 학습 방법론을 사용했는가 보다는 더 중요할 것으로 생각된다. 외국에서도 상품평 분류에 대한 연구[7,8]가 활발히 진행되고 있지만 초, 중, 종성이 모여서 하나의 음절을 구성하는 한글의 독특한 특성 때문에 외국의 사례를 한국어에 그대로 적용하기는 쉽지 않다.

3. 강건한 감정 분류를 위한 자질 추출 방법

대부분의 온라인 상품평은 그림 1과 같이 다양한 띄어쓰기 및 철자 오류를 포함하고 있다. 이렇게 철자 오류가 포함된 문장으로부터 기계 학습에 필요한 자질을 추출하기 위하여 본 논문에서는 2가지 종류의 사전을 자동 구축한다. 첫 번째 사전은 21세기 세종 계획[9] 형태소 말뭉치에서 체언과 용언을 포함하는 어절들을 추출하여 구축한 어절 패턴 사전이다. 어절 패턴 사전은 그림 1의 두 번째 예 "어머님이마이조아하시네영"과 같이 띄어쓰기가 잘못된 문장에서 체언과 용언 정보를 자질로 추출하는데 사용된다. 두 번째 사전은 21세기 세종 계획 형태소 말뭉치에 포함된 용언 어절들의 초/중성 패턴 사전이다. 상품평에 포함된 자의적 철자 오류는 '이뻐용', '조아하시네요'와 같이 주로 용언에서 발생하며, 이러한 철자 오류는 대부분 종성이 변형되어 나타난다는 특징이 있다. 그러므로 초/중성 패턴 사전은 변형이 적은 초/중성 패턴을 이용하여 철자 오류가 포함된 문장에서 용언 정보를 자질로 추출하는데 사용된다.

깔끔하고 이뻐용
어머님이마이조아하시네영
잘샀다는 생각에 기분조아요

그림 1 철자 오류가 포함된 상품평의 예

3.1 어절 패턴 사전 구축

21세기 세종 계획 형태소 말뭉치에서 체언과 용언이 포함되어 있는 어절만을 대상으로 어절 패턴 사전을 구축한다. 예를 들어 그림 2와 같은 형태소 말뭉치에서 체언이나 용언을 포함하는 어절인 '어머님이'와 '울었어'가 어절 패턴 사전의 엔트리(entry) 후보가 된다. 모든 어절 패턴 사전 후보를 추출한 후, 빈도수가 2 이상인 어

어머님이	어머님/NNG+이/JKS
많이	많이/MAG
울었어	울/VV+었/EP+어/EC

그림 2 형태소 말뭉치의 예

절만을 대상으로 어절 패턴 사전을 구축한다.

그림 3은 최종적으로 구축된 어절 패턴 사전의 일부를 보여준다. 그림 3에서 보는 것과 같이 어절 패턴 사전은 어절을 키(key)로 하고, 해당 어절의 형태소 분석 결과를 바탕으로 추출한 체언과 용언의 원형을 데이터(data)로 한다.

키(key)	데이터(data)
어머님이	어머님
생각에	생각
기분	기분
좋아하시네	좋아하
샀다는	사
좋아요	좋
지워야	지우
이뻐요	이쁘

그림 3 어절 패턴 사전의 예

3.2 초/중성 패턴 사전 구축

자의적인 철자 오류는 초성과 중성이 크게 변하지 않는다는 특징이 있다. 예를 들어, 상품평에서 '좋아요'는 '조아요', '조아용', '좋아여'와 같이 자주 표현되며, 이것들을 그림 4와 같이 초성, 중성, 종성으로 나누어 보면 초성이 모두 일치하는 것을 확인할 수 있다. 또한 중성의 경우에도 '좋아여'를 제외하고 모두 일치하는 것을 알 수 있다. 이러한 특징을 이용하여 21세기 세종 계획 형태소 말뭉치에 포함된 용언 어절들의 초성과 중성을 이용하여 초/중성 패턴 사전을 구축한다.

그림 5는 초/중성 패턴 사전의 일부를 보여준다. 그림 5에서 보는 것과 같이 초/중성 패턴 사전은 원본 어절의

어절	초성	중성	종성
좋아요	ㅈ	ㅗ	ㅎ
	ㅇ	ㅏ	
	ㅇ	ㅛ	
조아요	ㅈ	ㅗ	
	ㅇ	ㅏ	
	ㅇ	ㅛ	
조아용	ㅈ	ㅗ	
	ㅇ	ㅏ	
	ㅇ	ㅛ	ㅇ
좋아여	ㅈ	ㅗ	ㅎ
	ㅇ	ㅏ	
	ㅇ	ㅕ	

그림 4 철자 오류 용언의 초, 중, 종성 예

원본 어절	키		데이터
	초성	중성	
좋아하시네	ㅈㅇㅎㅅㄴ	ㅗㅏㅏㅣㅖ	좋아하
좋아하시는	ㅈㅇㅎㅅㄴ	ㅗㅏㅏㅣㅡ	좋아하
샀다는	ㅅㄷㄴ	ㅏㅏㅡ	사
싫다는	ㅅㄷㄴ	ㅣㅏㅡ	싫
좋아요	ㅈㅇㅇ	ㅗㅏㅛ	좋
지워야	ㅈㅇㅇ	ㅣㅝㅑ	지우
예뻐요	ㅇㅃㅇ	ㅖㅓㅛ	예쁘
이뻐요	ㅇㅃㅇ	ㅣㅓㅛ	이쁘

그림 5 초/중성 패턴 사전의 예

초성과 중성을 키로하고 해당 어절의 형태소 분석 결과를 바탕으로 추출한 용언의 원형을 데이터로 한다. 만약 '좁아요', '좋아요'처럼 초성과 중성이 모두 일치하는 어절의 경우에는 학습 데이터 내 원본의 빈도수가 높은 어절을 선택한다. 예를 들어 수집된 상품평 데이터 내 '좋아요'의 빈도수가 '좁아요'의 빈도수보다 높다면 '좋아요'를 선택하여 사전을 구성한다. 사전 구성 시 초성과 중성이 모두 일치하는 어절의 경우는 11%였으며, 실제 실험에서 사용된 자질은 추출된 전체 자질 중 4%였다.

3.3 최장 일치 기반 자질 추출

기계 학습을 위한 자질은 어절 패턴 사전과 초/중성 패턴 사전을 입력 문장과 최장 일치법으로 매칭(matching)하여 추출한다. 상품평에는 띄어쓰기 오류가 다수 포함되어 있는데 대부분이 '어머님이 마이조아하시네요'와 같이 띄어 쓸 것을 붙여 쓰는 오류이며, '어 머님이 마 이 조아하시 네요'와 같이 붙여 쓸 것을 불필요하게 띄어 쓰는 오류는 거의 존재하지 않는다. 그러므로 본 논문에서는 띄어 쓴 부분은 올바른 어절 경계라는 것을 가정하고 띄어쓰기 단위로 최장 일치를 수행한다. 최장 일치 시에는 어절 패턴 사전의 용언, 어절 패턴 사전의 체언, 초/중성 패턴 사전 순으로 우선순위를 부여하여 매칭한다. 그림 6은 그림 4, 5의 사전을 이용하여 그림 1의 상품평으로부터 자질을 추출한 결과이다.

입력 문장	어절 패턴 사전		초/중성 패턴 사전
	용언	체언	
깔끔하고 이뻐용	x	x	이쁘
어머님이마이조아하시네영	x	어머님	좋아하
잘샀다는 생각에 기분조아요	x	생각, 기분	좋

그림 6 최장 일치법을 이용한 자질 추출의 예

4. 실험 및 평가

실험을 위해서 가격 비교 사이트의 세탁기 카테고리에서 12,291 문장을 수집하였다[10]. 수집된 문장의 감정은 4명의 자연어처리 전공 석사과정 학생들에 의해서 부착되었으며, 이 중 2명 이상이 같은 감정을 부착한 문

표 1 실험 결과

자질 추출 방법	재현율(%)	정확률(%)	F1-척도(%)
형태소 분석 기반 체언, 용언 추출	71.36	83.45	76.94
어절 패턴 사전 최장 일치	74.80	82.99	78.68
어절 패턴 사전 + 초/중성 패턴 사전 최장일치	75.26	84.51	79.62

장만을 실험 대상으로 삼았다. 결론적으로 긍정 4,743 문장과 부정 1,492 문장이 실험에 사용되었다. 기계 학습은 이진 감정 분류 문제에서 높은 성능을 보인 SVM을 이용하였다[11]. 실험은 10배 교차검증(10-fold cross validation)으로 진행되었으며, 평균 F1-척도(F1-measure)를 이용하여 성능을 측정하였다. 상대 평가를 위해서는 제안 방법과 형태소 분석기를 이용하여 체언과 용언을 자질로 추출하는 방법을 비교하였다.

표 1은 자질 추출 방법에 따른 감정 분류 시스템의 성능을 보여준다.

표 1에서 보듯이 형태소 분석기를 이용한 용언과 체언을 추출하는 방법보다 어절 패턴 사전만을 이용하여 단순 최장 일치하는 방법이 온라인 상품평의 감정 분류에서 1.74% 높은 F1-척도를 나타냈다. 제안한 것과 같이 초/중성 패턴 사전까지 이용한 경우에는 2.68%나 높은 F1-척도를 나타냈다. 이러한 실험 결과를 토대로 온라인 상품평에는 다수의 철자 오류가 포함되어 있기 때문에 형태소 분석기를 이용하게 되면 형태소 오분석으로 인해 성능이 높게 나타나지 않는 것을 확인할 수 있었다. 또한 자의적인 철자 오류는 초성과 중성이 크게 변하지 않는다는 특징을 이용하여 자질을 추출할 수 있었으며, 시스템의 성능 향상에 영향을 미친다는 것을 확인할 수 있었다. 결론적으로 띄어쓰기 및 철자 오류를 많이 포함하는 온라인 상품평의 감정 분류에서는 형태소 분석기를 사용하는 것보다 패턴 기반의 최장 일치법을 사용하는 것이 보다 효과적임을 알 수 있었다. 추가 실험으로 기존 띄어쓰기 및 철자 오류에 대비한 김상도[6]의 연구와 비교하였다. 실험은 [6]과 같은 방법으로 진행되었으며 정확률을 평가 측도로 사용하였다. 표 2는 김상도의 방법 및 본 논문에서 제안한 방법의 정확률을 보여준다.

표 2에서 볼 수 있듯이 본 논문에서 제안하는 어절 패턴 사전과 초/중성 패턴 사전을 이용하는 방법이 더 높은 성능을 보였다. 비교 실험을 통하여 본 논문에서 제안하는 방법이 상품평 뿐만 아니라 다른 도메인의 감정 분류에도 적용되어 좋은 성능을 보임을 알 수 있었다.

감정 분류 측정 실험에서 시스템의 성능을 저하시키는 원인은 크게 두 가지 유형으로 나눌 수 있다. 저, "수평이자꾸트러지네요"와 같이 띄어쓰기 오류를 포함하는 경우이다. 말뭉치 내에서 "수평이"라는 어절은 존재하지 않으며 "수평", "이자"만을 포함하고 있기 때문에 "수평", "이자"를 사전으로 구축하게 되어 "수평"과 "이자"가 자질로 추출되어 나타나는 오류가 있다. 이러한 문제를 해결하기 위해서는 명사 자질 뒤에 조합될 수 있는 조사를 부착하여 사전을 확장하는 방법이 필요할 것으로 보인다. 두 번째 오류 유형은 초/중성 패턴 사전에 따른 오류이다. 패턴 사전 구축 시 초성과 중성이 일치하는 경우 학습 데이터에 따라 패턴을 대표하는 어절 하나만을 사전으로 이용하였다. 이 때 위의 예에서 "좀아요"가 실험 데이터로 사용되는 경우 "좋아요"의 자질을 추출함에 따라 오류가 발생한다. 이러한 문제를 해결하기 위해서는 패턴 추출 시 데이터 내 빈도수 뿐만 아니라 현재 패턴의 앞뒤 어절을 추가 자질로 이용하여 입력 문장에 따른 패턴 자질 추출 방법이 필요할 것으로 보인다.

표 2 기존 연구 비교 실험 결과

자질 추출 방법	정확률(%)
어절 패턴 사전 + 초/중성 패턴 사전 최장일치	77.61
음절 커널 기반 영화평 감성 분류[6]	71.00

5. 결론 및 향후연구

본 논문에서는 기계 학습 기반의 온라인 상품평에 대한 감정 분류 시에 형태소 분석기를 이용하여 자질을 추출하게 되면 띄어쓰기 및 철자 오류로 인하여 성능 저하가 일어난다는 것을 확인하였다. 이러한 문제를 해결하기 위하여 본 논문에서는 어절 패턴 사전과 초/중성 패턴 사전을 구축한 후, 단순 최장 일치법에 기반하여 자질을 추출하는 방법을 제안하였다. 실험 결과에 따르면 띄어쓰기 오류와 철자 오류가 많이 포함된 온라인 상품평의 감정 분류에 제안한 자질 추출 방법이 더 적합함을 알 수 있었다.

향후 연구과제는 다음과 같다. 제안 방법으로 추출된 자질 중에서 중요한 자질을 선택하는 통계적인 방법을 적용하여 감정 분류 시스템의 성능을 높이는 연구를 진행 할 예정이다. 또한 영화평과 같이 자의적인 철자 오류가 빈번히 발생하는 다른 영역에 대해서도 다양한 실험을 진행할 예정이다.

참 고 문 헌

[1] J. Hwang and Y. Ko, "A Korean Document Sentiment Classification System based on Seman-

tic Properties of Sentiment Words," *Journal of KIISE : Software and Applications*, vol.37, no.4, pp.317-322, Apr. 2010. (in Korean)
[2] H. Yune, H. Kim and J. Chang, "An Efficient Search Method of Product Reviews using Opinion Mining Techniques," *Journal of KIISE : Computing Practices and Letters*, vol.16, no.2, pp.222-226, Feb. 2010. (in Korean)
[3] J. Myung, D. Lee and S. Lee, "A Korean Product Review Analysis System using a Semi-Automatically Constructed Semantic Dictionary," *Journal of KIISE : Software and Applications*, vol.35, no.6, pp.392-403, Jun. 2008. (in Korean)
[4] J. Shin, J. Lee and H. Kim, "Sentiment Categorization of Korean Customer Reviews using CRFs," *Proc. HCLT(Human & Cognitive Language Technology)*, vol.20, no.1(C), pp.58-62, 2008.(in Korean)
[5] M. Bae and J. Cha "Comments Classification System using Topic Signature," *Journal of KIISE : Sofrware and Applications*, vol.35, no.12, pp.774-779, Dec. 2008. (in Korean)
[6] S. Kim, S. Park, S. Park, S. Lee and K. Kim, "A Syllable Kernel based Sentiment Classification for Movie Reviews," *Journal of KIISS*, vol.20, no.2, pp.202-207, Jun. 2010. (in Korean)
[7] A. Esuli, F. Sebastiani, "PageRanking WordNet Synsets: An Application to Opinion Mining," *In Proceedings of the ACL*, pp.424-431, 2007.
[8] S.M. Kim and E. Hovy, "Determining the Sentiment of Opinions," *In Proceedings of the COLING conference*, pp.1367-1373, 2004.
[9] http://www.sejong.or.kr
[10] http://shopping.naver.com
[11] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," *In Proceedings od the EMNLP*, pp.79-86, 2002.

신 준 수
정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 10 호

김 학 수
정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 10 호