온라인쇼핑몰 상품평 문법적 오류 개선을 위한 오피니언 마이닝에 대한 연구

박세정* · 황재승** · 김종배***

*,**,***숭실대학교

Research for the opinion mining for the improvement of online shopping mall review grammatical errors

E-mail : kjb123@ssu.ac.kr

요 약

현대인들은 필요한 물건들을 직접 구매하러 갈 시간이 부족하기 때문에 온라인 쇼핑몰의 이용 빈도가 늘어가고 있으며 이에 따라 온라인 쇼핑몰이 성행하고 있다. 하지만 온라인 쇼핑몰에서 물건을 구매하는 것은 물건을 눈으로 확인할 수 없다는 문제점이 있기 때문에 상품평은 구매를 결정하는데 많은 영향을 준다. 현재 온라인 쇼핑몰에서 고객이 상품평을 통해 상품에 대한 정보를 파악하기 어렵기 때문에 이를 해결하기 위한 연구들이 진행되고 있다. 이러한 연구들로 상품평의 의견을 분석하기 위한 연구로 오피니언 마이닝이 사용되고 있는 추세이다. 그러나 지금까지의 연구는 문법적인 오류, 신조어와 같이 국어사전에 등재되어 있지 않은 단어들을 감성분석기가 올바르게 판단하지 못하기 때문에 분석의 신뢰도가 떨어진다는 문제점이 있다. 그래서 형태소 분석을 실시하기 전에 신조어 사전을 추가하여 Noisy-channel model을 적용하여 더욱 정확한 감성분석이 가능하도록 하였다. 이러한 과정을 통해 가공된 정보를 바탕으로 상품평을 보다 정확하게 분석할 수 있는 시스템을 제안하고자 한다.

키워드

오피니언 마이닝, 상품평, 신조어 사전, 형태소 분석, Noisy-channel model

Ⅰ. 서 론

2010년 25조 2천억 원이었던 온라인 쇼핑 매출액은 2014년 46조9280억 원으로 약 두 배 가까이 증가하여 대형마트 매출액인 46조6364억 원을 넘어섰다. 이처럼 거대한 온라인 쇼핑 시장에서 고객들을 유치하기 위한 경쟁력은 고객들에게 상품에 대한 믿을만한 정보를 한눈에 제공하는 것이다. 고객이 상품에 대한 정보를 찾아 볼 때 가장 많이 고려하는 것은 다른 고객들이 직접 구매하고 사용한 후 남긴 상품평이다. 하지만 고객 스스로가 일일이 상품 평을 분석하는 데에는 많은 시간이 소요되기 때문에 이를 해결하기 위한 연구가 진행되고 있다. 대표적으로, 오피니언 마이닝을 이용하여 상품평의 긍정, 부정 정도를 판단하고, 그 상품에 대한 선호도를 분류하여 평가를 내리는 기법이 있다.

하지만 기존의 연구는 오피니언 마이닝을 통해 형태소를 추출할 때 문법적인 오류, 신조어 및 은어를 포함한 상품 평에 대한 정확도가 떨어지는 문제점이 있었다. 따라서 본 논문에서는 더욱 정확한 오피니언 마이닝을 하기 위해 온라인 쇼핑몰에 올라온 상품 평들을 추출한 다음 맞춤법과 신조어를 검사하여 바꿔줄 것이다. 이를 위해 신조어 사전과 맞춤법 교정을 이용한 문법적 오류 제거 알고리즘을 제시한다. 먼저, 추출된 상품 평에서 가장 많이 사용되는 신조어들을 분류하고, 대체가 가능한 단어로 바꾼다. 그리고 Noisy-channel model을 활용하여 띄어쓰기와 철자를 교정하고 그 효율성을 검증할 것이다.

Ⅱ. 관련 연구

오피니언 마이닝은 다양한 온라인 뉴스와 소셜 미디어의 코멘트, 사용자가 만들어 내는 컨텐츠들에서 표현된 의견을 추출, 분류, 이해하고 자산화하는 컴퓨팅 기술이다. 이를 위한 온라인 텍스트 속의 여러 가지 감정상태를 식별하기 위해 감성 분석이 사용되곤 한다[1, 2, 5]. 오피니언 마이닝은 텍스트 마이닝의 한 분류로서 평판 분석으로도 불리며, 소셜 미디어 등의 정형/비정형 텍스트의 선호도를 판별하는 기술이다[3].

하지만 문법적인 오류로 인해 감성 분석이 제대로 판단되지 않는 경우가 있다. 이를 위해 제시되는 noisy-channel model은 확장하여 띄어 쓰기 오류와 철자 오류를 동시에 교정 가능하여 정확도를 높여준다.[6,7] 대표적인 예로 Noisy-channel model은 '야후!'에서 구축하여 분석한 한글 오타 패턴과 사용자 로그를 기반으로 설계한 질의어 교정 서비스에 대한 모델이 있다.

그리고 교정을 통한 문장은 여러 개의 어절로, 글을 분석하기 위해선 형태소 분석이 필수적이라고 할 수 있다[2]. 형태소 분석기의 성능에 영향을 끼치는 데에는 형태소 분석 알고리즘뿐만 아니라 사용하는 단어 사전의 영향도 매우 크다.[4] 그러므로 사용자의 목적/취향에 맞는 분석기와 사전의 조합을 사용해야 한다.

Ⅲ. 본론

본 논문은 상품평에 있는 문법적 오류나 신조어 및 은어들을 수정할 수 있는 시스템을 구현하였다. 상품평은 인터넷 쇼핑몰에서 자전거 상품평 데이터 1000개를 추출하여 형태소 분석을 해보았다. 시스템 구조도는 그림 1과 같다.

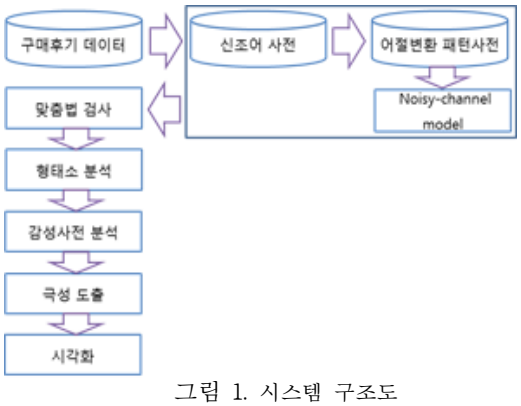

그림 1. 시스템 구조도

데이터를 수집 하여 KoNLP와 데이터베이스를 통해 신조어 사전 비교한다. 상품평 데이터를 수집하여 가장 많이 사용되고 있는 신조어 1000개를 추려내어 저장한다. 그리고 수집된 상품평이 형태소분석기에서 처리가 가능하도록 먼저 신조어에 대한 올바른 대안을 제시한다. 이 검증 과정에서는 미리 준비된 신조어 사전 정보에 따라 상품평 안에서 정확히 일치하는 단어만을 찾아 바꿔주며, 표 1.의 결과와 같다.

표 1. 신조어 사전 비교

원글	수정글
내 뱃살 지못미	내 뱃살 지켜주지 못해서 미안해

그 후, 형태소 분석기를 사용하기 전에 맞춤법 교정을 실시한다. 기존의 전 처리기 알고리즘의 한계를 극복하고, 최대한 사전을 적게 이용하여 효과적으로 교정후보를 생성할 것이다. noisy-channel model을 확장하여 띄어 쓰기 오류와 철자 오류를 동시에 교정 가능한 전처리기를 구현한다.

맞춤법 검사기는 자소 단위의 변환 확률 값을 이용한 교정방법을 이용한다. 그 이유는 철자 오류의 많은 부분은 자소의 변환만으로도 교정이 가능하며, 어절 변환 패턴 사전의 크기를 줄일 수 있게 되기 때문이다. 음절들의 많은 변화를 자소 단위 오류 수정 모델만으로 초,중,종성의 조합을 이용해 교정이 가능하다. 또한, 자소 단위 후보 생성은 말뭉치에 없는 음절이라 하더라도 작은 단위인 자소 조합을 통해 더욱 정밀한 후보생성이 가능하다.

온라인 상품평의 경우 발음상의 실수, 맞춤법 지식의 부재, 고의로 다르게 표기한 경우가 많다. 각각의 예를 표2에 나타내었다.

표 2. 맞춤법 교정

	표기	교정
발음상의 실수	같애요 / 먹어여	같아요 / 먹어요
맞춤법 지식의 부재	희안한	희한한
고의로 다르게 표기한 경우	나와욧	나와요

(7)에서 참조한 수식을 기반으로하여 다음과 같은 변형된 수식을 적용해서 계산할 수 있다.

$$C' = \text{argmax}P(C|T) \qquad (7)$$

이 수식을 통해 띄어쓰기와 철자오류가 모두 포함된 문장인 원문 T(si, bi)로부터 하나의 자소인 si를 초,중,종성 총 3개의 자소로 구성하여 T로

부터의 교정후보 C를 구할 수 있다. 그리고 T의 교정후보 C 중에서 확률이 가장 높은 값을 가진 교정 후보 C가 C'가 된다.

이러한 오류 자소들이 어떤 자소로 교정이 되었는지에 대한경우의 수를 세어서 확률변환 데이터로 구축하여 시스템의 성능 향상이 가능하다. 또한 각각의 자소 변환 확률은 초성, 중성, 종성을 구분하여 계산함으로써 자소 후보의 과생성을 막고자 한다. 그림 2는 자소 단위 후보 생성의 예시를 보여준다.

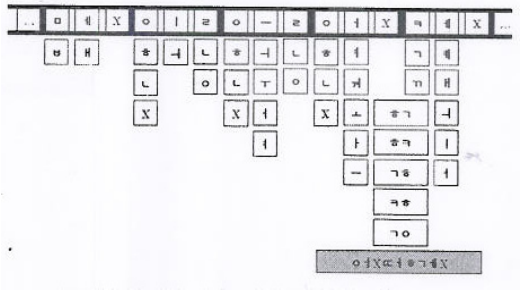

그림 2. 자소 단위 후보 생성의 예시

그림의 맨 윗줄은 입력문장을 자소 단위로 나눈 것이고 둘째 줄 이하부터는 원 자소의 후보 자소들이다. 각 자소노드는 자소변환확률값의 로그 값을 가지고있어 P(C|T)의 계산에 사용된다. 인접한 자소 2개가 후보 생성이 가능할 때 그 후보 또한 생성된다. 그리고 음절 이상 단위의 변환도 어절변환패턴 사전으로부터 후보가 생성되며 실험 결과는 아래 표 3, 표 4와 같다.

표 3. 입력된 띄어쓰기를 무시한 경우

	자소 변환 확률 이용	상수이용
초, 중, 종성 구분	73.57%	65.10%
초, 중, 종성 구분 안함	61.89%	63.71%

표 4. 입력된 띄어쓰기를 유지한 경우

	자소 변환 확률 이용	상수이용
초, 중, 종성 구분	84.25%	80.92%
초, 중, 종성 구분 안함	79.67%	80.61%

실험 결과로 알 수 있는 점은 띄어쓰기 성능이 전체 결과에 상당한 영향을 주고 둘째, 초, 중, 종성을 구분하면서 확률 데이터 값을 적용 하였을 때 성능이 높다는 점을 알 수 있다.

Ⅳ. 결 론

본 논문에서는 오피니언 마이닝을 위해 문법적인 오류, 신조어 및 은어를 수정했다. 이를 위해 각 단계별 프로세스를 간단하게 설명하고 상품 평에 대한 정확도를 향상 시키는 법에 대해 연구해 보았다.

띄어 쓰기와 철자 교정을 동시에 수행한 실험 결과 낮은 성능을 보인 이유로는 첫째, 자소나 음절 단위의 변환 확률 값은 어느 정도의 말뭉치만으로도 신뢰할 만한 수치를 얻을 수 있고 둘째, P(C)를 계산하기 위한 n-gram 데이터는 충분한 양의 말뭉치가 있어야 신뢰성 있는 데이터를 얻을 수 있다는 점 셋째, 말뭉치의 오류 태그가 일관성이 없으므로 성능 측정치가 신뢰성이 높다고 할 수 없고 마지막으로 말뭉치 자체가 전문 용어가 많아 n-gram 데이터의 질에 영향을 준다는 것이다.

기존의 연구들은 문법적 오류 등을 수정하지 않고 오피니언 마이닝을 구현하였다. 본 논문에서 제시하는 방법은 이러한 오류들을 수정한 다음 오피니언 마이닝을 진행하는 방법이다. 좋은 성능을 보였다고 하기는 힘들지만 연구를 통해 더 많은 문법적 오류나 신조어 등의 특징을 찾아 기존의 오피니언 마이닝에 사용한다면 성능 향상에 도움을 줄 것이다.

참고문헌

[1] Hwi Hoon An, Byung Joon Park, "Opinion word extraction after removal of Similar advertising review", KICS, Volume 54, PP.600 - 601, June 2014

[2] Jae-Young Chang, "A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall", The Journal of Society for e-Business Studies, Volume 14, November 2009

[3] Jung-yeon Yang, Jaeseok Myung, Sang-goo Lee, "A Sentiment Classification Method Using Context Information in Product Review Summarization", KIISE, August 2009

[4] G.R. Brindha, B. Santhi, "A Novel Opinion Mining Technique for Product Review Based on Preferences", Research Journal of Applied Sciences, Engineering and Technology, Volume 4, Number 23, December 2012

[5] Daekook Kang, Yongtae Park, “Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach”, Expert Systems with Applications, Volume 41, August 2013

[6] Baoxiang Li, Gang Liu, Jun Guo, Yueming Lu, “An Improved Noisy Channel Model for Speech Recognition Error Correction “International Journal of Advancements in Computing Technology, Volume 4, July 2012

[7] Limin Wang, Chunhong Cao, Xiongfei Li, Haijun Li, “Finding the Optimal Feature Representations for Bayesian Network Learning.“, PAKDD, volume 4426, May 2007