한국지능시스템학회 논문지 2010, Vol. 20, No. 2, pp. 202-207

음절 커널 기반 영화평 감성 분류

A Syllable Kernel based Sentiment Classification for Movie Reviews

김상도* · 박성배* · 박세영* · 이상조* · 김권양**†

Sang-Do Kim*, Seong-Bae Park*, Se-Young Park*, Sang-Jo Lee* and Kweon-Yang Kim**†

*경북대학교 컴퓨터공학과

**경일대학교 컴퓨터공학과

요 약

본 논문에서는 감성 점수가 명시적으로 부여되지 않은 온라인 영화평에 대해 자동으로 감성을 분류하는 방법을 제안한다. 긍정이나 부정과 같은 감성 극성 분류를 위해 문자열 커널의 확장 모델인 음절 커널에 기반한 지지벡터기계를 분류기로 사용한다. 실험을 통하여 띄어쓰기나 철자 오류 같은 문법적인 오류가 빈번한 온라인 영화평에 대한 감성 분류에서 제안한 음절 커널 방법이 효과적임을 보인다.

키워드 : 감성 분류, 음절 커널, 문자열 커널, 지지벡터기계

Abstract

In this paper, we present an automatic sentiment classification method for on-line movie reviews that do not contain explicit sentiment rating scores. For the sentiment polarity classification, positive or negative, we use a Support Vector Machine classifier based on syllable kernel that is an extended model of string kernel. We give some experimental results which show that proposed syllable kernel model can be effectively used in sentiment classification tasks for on-line movie reviews that usually contain a lot of grammatical errors such as spacing or spelling errors.

Key Words : Sentiment classification, Syllable kernel, String kernel, Support Vector Machine

1. 서 론

웹의 급속한 성장과 더불어 컴퓨터 환경에 익숙하지 않은 사용자들도 블로그나 개인 홈페이지를 통하여 자신의 의견을 다른 사용자들과 공유하거나 서로 간에 다양한 정보를 나누는 공간으로 사용하고 있다. 이러한 인터넷 환경의 변화로 인해 사용자들은 상품이나 서비스의 구매 이전에 이들에 대한 다른 사용자의 의견을 분석하여 구매 시에 의사 결정에 참조하려는 경향이 증가되고 있다. 또한 상품이나 서비스 제공자는 소비자의 만족 지수를 높이기 위해 상품에 대한 소비자의 반응을 분석하여 효과적인 마케팅 전략을 수립하고 있다.

감성 분석(sentiment analysis)은 문서 내에서 표현된 특정한 감정 표현의 의미를 추출하는 텍스트 데이터 마이닝(text data mining) 기술의 한 영역이다. 또한 감성 분류(sentiment classification)란 감성 분석에 의해 추출된 감성 표현에 기반하여 문서를 긍정(positive) 의견 또는 부정(negative) 의견과 같은 극성(polarity)으로 분류하는 것을 말한다. 사용자 의견을 공유하는 일부의 개인 블로그나 온라인 리뷰 사이트에서는 사용자가 자신의 의견과 더불어 별점과 같은 형태의 감성 점수를 직접 기입하고 있지만 사용자의 자발적인 참여가 부족한 편이며, 대다수의 경우 해당 주제에 대한 사용자의 의견만 글로 기술할 뿐 명시적인 감성 정보를 기재하지는 않는다. 인터넷 상에 게시된 사용자 의견의 수는 방대할 뿐 아니라 시간이 지남에 따라 계속적으로 증가하고 있기 때문에 이들 의견에 대한 분석 결과를 응용 분야에 즉각적으로 반영하기 위해서는 자동화된 감성 분석 기술이 요구된다.

본 연구에서는 인터넷 상에 게재된 영화평에 대해 긍정과 부정인 극성으로 분류 결과를 구분하는 이진 감성 분류 문제를 다룬다. 영화평 감성 분류를 위해 사용한 기계 학습 방법은 이진 분류 문제에서 비교적 높은 성능을 보이는 지지벡터기계(SVM: Support Vector Machine)를 이용한다.

전형적인 정보 검색 기술에서 문서 표현 방법으로 주로 사용해 온 표준 벡터 모델인 BOW(Bag Of Words) 모델은 단어 간의 순서 정보를 표현하지 않기 때문에 문장 내에서의 위치 정보에 따라 감성 분석의 결과가 달라지는 경우에 정확한 분석 결과를 기대하기 어렵다. 또한 본 논문에서 감성 분석 대상으로 삼은 영화평의 길이가 최대 40자로 비교적 짧으며, 띄어쓰기나 철자의 오류가 빈번히 나타나기 때문에 단어를 기본 비교 단위로 사용하는 BOW 모델은 높은 정확도를 가지는 복잡한 형태소 분석기를 통한 전처리 과정을 사용하지 않는 한 문서에 대한 효과적인 표현을 제공할 수 없다.

접수일자 : 2010년 2월 17일
완료일자 : 2010년 3월 22일
+ 교신저자

BOW 모델과는 달리 문서를 구성하는 요소들 사이의 순서 관계 정보를 표현할 수 있는 문자열 커널(string kernel)은 단어 대신에 문서 내에 포함된 생성 가능한 모든 연속·비연속 부분 순서 문자열(subsequence string)을 이용하여 문서를 표현하고, 공유하는 부분 문자열을 비교함으로써 문서 간의 유사도를 추정하게 한다.

본 연구에서 제안한 감성 분류기는 문자열 커널의 확장 모델인 음절 커널(syllable kernel)을 사용한다. 문자열 커널에서는 기본 비교 단위로 문자를 사용하는 반면에 음절 커널은 음절을 기본 비교 단위로 사용한다. 따라서 음절 커널은 음절들로 결합된 비교적 의미있는 문자열 조합만을 표현함으로써 문자열 커널에 비해 커널 계산에 따른 계산량을 감소시켜 준다.

본 연구에서 제안한 감성 분류기가 띄어쓰기나 철자의 오류가 빈번히 나타나는 문서에서도 안정된 성능을 가짐을 보이기 위해 온라인 인터넷 상의 40자 영화평을 대상으로 실험하였다. 실험 결과 형태소 분석이 어려운 데이터에 대해서도 제안한 음절 커널 모델은 기준 평가 측도(baseline)인 BOW 모델에 비해 훨씬 안정적인 성능을 보였다.

2. 관련 연구

최근, 기업들은 자사의 상품이나 서비스에 대한 고객의 반응에 대해 좀 더 빨리 그리고 적극적으로 대응하기 위해 인터넷 상의 댓글이나 게시글의 감성을 분석하는 연구를 활발히 진행하고 있다. 이러한 감성 분석을 통하여 고객들에게 좀 더 만족스러운 서비스를 제공할 수 있을 뿐만 아니라 기업들의 고객 대응 방안에 대한 전반적인 통찰력을 가지게 된다.

기존 감성 분류 연구는 감성 사전에 기반한 방법과 기계학습 기법을 사용한 방법으로 나눌 수 있으며, 감성 분류기에서 필요시 되는 효과적인 여러 언어적 속성들이 제안되었다. 또한 이들 언어적 속성에 기반하여 여러 기계 학습 알고리즘의 효과성이 검토되고 있다.

영어권에서는 감성 사전을 이용하여 문서 내에 있는 단어의 감성 극성을 판단하고, 각각의 판단 결과에 따라 전체 문서의 감성을 분류하는 방법이 제시되었으며, 특히 Hatzivassiloglou와 McKeown은 품사가 형용사인 단어들에 대한 감성 분류 방법을 제안하였다[1]. 반면에 Turney와 Littman은 형용사 외에 부사를 추가로 감성 분류에 사용하였다[2]. 이들은 특정 패턴을 갖는 의견성 단어에 PMI(Pointwise Mutual Information)방법을 사용하여 "excellent"와 "poor"와 같은 초기 단어들과 공기하는 정보를 이용하여 극성을 할당하고 이를 바탕으로 문서의 극성을 분류하였다. PMI는 비슷한 감성을 가지는 어휘는 가까운 위치에서 함께 나타날 가능성이 높을 것이라는 가정에 근거하여 단어 사이의 관계를 추측하는 방법이다. Chesley 등은 위키피디어(Wikipedia) 온라인 사전인 Wiktionary 사전을 이용하여 블로그의 감성을 분류하는 방법을 제안하였는데, 문서 내에 품사가 형용사인 단어 외에 동사도 이용하였다[3]. Kamp 등은 형용사의 감성 분류를 위해 WordNet[4]에서 정의된 어휘 관계(lexical relations)를 사용하였다[5].

기계 학습 기법에서 자질 선택(feature selection)은 문서의 감성 분류 성능에 가장 큰 영향을 준다. Pang 등은 [6]은 해당 문서에 대한 감성을 긍정과 부정으로 나누는 이진 분류 문제로 보고, 자연어 처리 및 기계 학습 기법을 적용하였다. 사용한 자질은 영화평을 n-그램(n-gram) 단위의 BOW로 표현하여 자질 벡터(feature vector)를 구성하였다. 이 연구에서 다양한 분류기를 실험한 결과 지지벡터기계가 가장 좋은 성능을 보였다.

한국어의 경우 감성 정보를 가진 사전이 아직 제공되지 않고 있는 실정이다. 따라서 황재원 등은 영어권의 감성 사전을 번역하여 한국어 감성 사전을 구축하고 이를 이용하여 문서의 감성 분류를 시도하였지만 기존 방법에 비해 성능 향상에 큰 차이를 보이지 않았다[7]. 명재석 등은 특정 도메인에 대한 감성 사전을 구축하고, 이를 사용하여 상품에 대한 소비자의 선호도를 분석하였다[8]. 남상협 등은 웹에서 쉽게 얻을 수 있는 리뷰에서 n-그램 단위인 의견 어구를 추출하여 각 어구를 긍정과 부정으로 분류하는 방법을 제안하였다[9]. 김묘실 등은 감성 사전 정보를 사용하는 대신에 형태소 분석의 결과를 기반으로 지지벡터기계를 사용하여 댓글의 악성 유무를 판별하는 시도를 하였다[10].

3. 감성 분류기

3.1 지지벡터기계

지지벡터기계는 두 개의 범주를 구분하는 분류 문제 해결을 위해 1995년 Vapnik에 의해 소개된 기계학습 기법으로, 두 클래스의 여백(margin)을 가장 최대화하는 초평면(optimal hyperplane)을 찾는다. 지지벡터기계는 최근 문서 분류에 널리 쓰이고 있으며, 다른 분류기에 비하여 좋은 성능을 보여주고 있다. 특히 지지벡터기계는 많은 양의 데이터와 높은 차원의 자질 집합을 가진 분류 작업에 우수한 성능을 보인다고 알려져 있다. 그림 1은 하나의 초평면에 의해서 결정되는 여백을 보여준다.

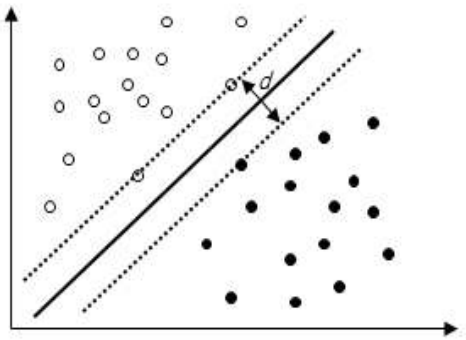

그림 1. 초평면의 여백
Fig. 1. Margin of optimal hyperplane

최적의 초평면을 찾는 것이 지지벡터기계의 역할이고 초평면을 찾음으로써 새로운 데이터가 주어졌을 때, 이를 분류할 수 있다. 학습 데이터가 $D=\{(x_1, y_1),(x_2, y_2) \ldots,(x_n,y_n)\}$와 같이 주어졌을 때, 여기서 x_i는 i번째 데이터를 가리키고 $y_i \in \{-1,+1\}$는 i 번째 데이터의 레이블이다. 이때 초평면의 수식은 다음 (1)과 같다.

$$H=\omega \cdot x+b \tag{1}$$

이때 두 클래스의 지지벡터를 경계로 한 두 영역은 아래와 식 (2)와 같이 정의한다.

$$\omega \cdot x + b \geq +1 \qquad (2)$$
$$\omega \cdot x + b \leq -1$$

위의 수식을 간소화하여 나타내면, 식 (3)

$$y(\omega \cdot x + b) - 1 \geq 0$$

이다. 여기서 위의 식은 지지벡터기계에서 데이터가 선형 분리가능(linearly separable)하다는 가정을 하고 있다. 하지만 실제 데이터는 일반적으로 선형 분리가 불가능(non-linearly separable)하다. 즉, 선형 함수(linear function)를 이용하여 분류할 수 없는 경우가 많다. 이와 같은 경우, 데이터를 선형 함수를 이용하여 분류할 수 있는 고차원 공간에 사상하여 문제를 해결할 수 있다. 지지벡터기계에서는 커널 함수를 사용하여 고차원 선형 공간 상에서의 데이터 유사도를 측정한다. 문서 분류에는 일반적으로 선형 커널 함수가 사용이 간편하고 성능도 우수하다고 알려져 있다.

3.2 문자열 커널

문자열 커널은 Lodhi[11] 등이 제안한 방법으로 문서의 유사성을 비교할 때 단어 대신에 문서 내에 포함된 생성 가능한 모든 연속·비연속 부분 순서 문자열을 비교한다. 부분 순서 문자열은 반드시 연속적일 필요는 없으며, 비연속 문자열의 경우 문자열에 포함된 간격에 따라 서로 다른 가중치를 가진다.

문서에서 나올 수 있는 문자의 집합을 Σ라고 할 때, Σ에 속한 문자를 유한하게 배열한 문자열 s를 자질 공간으로 사상시키는 함수 Φ는 다음과 식 (4)와 같이 정의된다.

$$\Phi_u(s) = \sum_{i:u=s[i]} \lambda^{l(i)} \qquad (4)$$

여기서 u는 문자열 s에 속한 부분 문자열이다. $i=(i_1,\ldots,i_{|u|})$는 $1 \leq i_1 < \ldots < i_{|u|} \leq |s|$을 만족하는, 부분 문자열 u를 구성하고 있는 문자들의 인덱스 집합을 의미한다. λ는 1보다 작거나 같은 조건을 만족하는 상수 값이며, 비연속 순서 문자열에 대해 벌점을 주기 위해 사용된다. 작은 값의 λ는 긴 비연속 부분 문자열에 대해 큰 벌점을 부가하게 되어 연속적인 부분 문자열만 커널 값에 영향을 미치게 된다. 반면에 λ 값이 1이면 비연속 간격이 큰 문자열에 대해서도 벌점을 부가하지 않는 효과를 가지게 된다. $l(i)$는 인덱스 i의 길이를 의미하며, 이는 $i_{|u|} - i_1 + 1$로 계산된다. 이 자질 벡터를 기반으로 두 개의 문자열 s와 t의 내적은 다음과 식 (5)와 같이 계산된다.

$$K_n(s,t) = \sum_{u \in \sum^n} \Phi_u(s)\Phi_u(t) = \sum_{u \in \sum^n} \sum_{i:u=s[i]} \sum_{j:u=s[j]} \lambda^{l(i)+l(j)} \qquad (5)$$

여기서 Φ는 부분 문자열의 길이 n 값이 작은 경우에도 계산량이 너무 많아 자질 벡터의 값을 직접적으로 계산하기가 어렵다[11]. 하지만 문자열 커널에서는 다음과 같은 새로운 함수를 도입하여 이를 효율적으로 계산할 수가 있다.

$$K'_i(s,t) = \sum_{u \in \sum^n} \sum_{i:u=s[i]} \sum_{j:u=s[j]} \lambda^{|s|+|t|-i_1-j_1+2} \qquad (6)$$

위의 함수는 $l(i)$와 $l(j)$ 대신에 특정 문자열의 시작 위치에서 각 문자열 끝까지의 길이를 사용한다. 이 함수는 다음과 같은 동적 프로그래밍 기법에 기반한 재귀적인 규칙을 통하여 계산되며, 이를 바탕으로 K_n을 계산한다.

$$\begin{aligned}
&K'_0(s,t) = 1, \text{for all } s,t,\\
&K'_i(s,t) = 0, \text{if } \min(|s|,|t|) < i,\\
&K_i(s,t) = 0, \text{if } \min(|s|,|t|) < i,\\
&K'_i(sx,t) = \lambda K'_i(s,t) + \sum_{j:t_j=x} K'_{i-1}(s,t[1:j-1])\lambda^{|t|-j+2},\\
&\qquad i = 1,\ldots,n-1,\\
&K_n(sx,t) = K_n(s,t) + \sum_{j:t_j=x} K'_{n-1}(s,t[1:j-1])\lambda^2.
\end{aligned} \qquad (7)$$

문자열 커널은 일반적으로 문서의 길이가 길어질수록 값이 커진다. 이에 따른 치우침을 해결하기 위하여 다음식 (8)과 같이 정규화한다.

$$\tilde{K}(s,t) = \frac{K(s,t)}{\sqrt{K(s,s)K(t,t)}} \qquad (8)$$

3.3 음절 커널

음절 커널은 문자열 커널의 확장된 모델로서 문자열 커널에서 사용된 개별 문자에 비해 보다 큰 구조적 단위인 음절을 기본 비교 단위로 사용한다. 따라서 음절 커널은 의미적으로 불필요한 과도한 문자열 조합을 처리 대상에서 제거함으로써 문서 표현을 간결하게 하는 효과를 제공하여 커널 계산에 따른 계산량을 감소시켜 준다. 비교되는 부분 순서 문자열이 문자 대신에 음절에 기반한다는 점을 제외하고는 문자열 커널에서 정의된 동일한 가중치와 유사성 측도를 사용한다. 문자의 유한 집합인 Σ는 단일 문자 대신에 한글 음절 집합[12]인 Σ={가, 각, 갂, ... , 힢, 힣}로 구성된다.

표 1. 음절 단위 부분 문자열 표현 예제
Table 1. Example of syllable based substring

문장	기대가너무컷나 지루지루 에릭은넘멋쪼
음절 단위 부분 문자열	"기", "대", ..., "멋", "쪼", "기대", "기가", ..., "넘멋", "멋쪼", "기대가", "기대너", ..., "넘멋쪼", "기대가너", "기대가무", "가컷지루", ..., "기대가너무컷나", ..., "기대가너무컷나 지루지루 에릭은넘멋쪼"

음절 커널은 주어진 두 문서가 서로 공유하는 음절 단위의 연속·비연속 부분 문자열이 많을수록 두 문서의 유사도를 높게 추정한다. 표 1은 "기대가너무컷나 지루지루 에릭은넘멋쪼" 문장에 대해 생성 가능한 음절 단위의 부분 문자열을 제시한 예이다. 제시한 문장의 길이가 비교적 짧지만 생성된 음절 단위 부분 문자열은 의미있는 다양한 문자열 조합을 표현할 수 있다.

표 2는 음절 단위 부분 문자열 표현을 사용할 때 두 문장이 서로 공유하는 음절 단위의 부분 문자열을 비교한 예를 보여준다. 예에서 사용된 두 문장은 띄어쓰기나 철자의 오

류가 많이 포함된 이유로 인하여 음절보다 큰 단위인 단어를 사용하는 BOW 모델인 경우에 어근을 추출하기 위해 복잡한 전처리 과정을 사용하지 않은 한 비슷한 의미를 표현한 문장임에도 서로 공유하는 단어가 없음을 알 수 있다.

표 2. 공유 음절 단위 부분 문자열 예
Table 2. Example of shared syllable based substring

문장 A		내 생에 진짜루재밌는영화였어
문장 B		그 영화 진짜로 재밌던 옛날영화였지
공 유 문자열	3음절	“진짜재”, “진짜밌”, “진짜영”, ..., “재영화”, “재영였”, “영화였”
	4음절	“진짜재밌”, “진짜영화”, “재밌영화”, ...

반면에, 음절 커널은 유사한 감성을 가지는 두 문장에 대해 비교 대상이 3 음절인 경우 “진짜재”, “진짜밌”, “진짜영” 등과 같은 공통되는 부분 문자열을 가지며, 4 음절인 경우에 “진짜재밌”, “진짜영화”, “재밌영화” 등과 같은 공통 문자열을 가지게 되어 두 문장의 의미가 유사하다는 분석을 할 수 있게 된다.

4. 실험 및 결과

4.1 실험 자료

본 논문에서 제안한 감성 분류기의 성능을 시험하기 위해 네이버 영화 사이트(http://movie.naver.com)에 게시된 영화 40자평을 실험 데이터로 사용하였다. 2003년 12월부터 2009년 6월 사이의 169만개 40자평 중에서 각각에 포함된 명사 성분의 비율에 따라 5 구간(0%-20%, 21%-40%, 41%-60%, 61%-80%, 81%-100%)으로 나누고 각 구간별로 평점(1점-10점) 각각에 해당하는 800 개의 40자평을 무작위로 선택하여 구간별로 8,000 개씩 총 40,000 개의 40자평을 실험 대상으로 하였다. 이진 분류를 적용하기 위해 평점 1에서 5까지의 40자평은 부정으로, 6에서 10까지는 긍정 의견으로 설정하였다. 감정 분류 실험은 구간별로 8,000 개의 40자평에 대해 시행하였으며, 안정적인 성능 평가를 위해 이중 90%인 7,200 개는 학습용으로 나머지 10%인 800 개는 시험용 자료로 10겹 교차 검증법(10-fold cross validation)을 사용하였다.

실험에서 사용한 형태소 분석기에서는 어휘 사전에 등록되어 있지 않아 분석이 불가능한 단어는 미등록 명사로 분류한다[13]. 미등록 명사에는 고유 명사, 전문 용어, 신조어도 포함되지만, 주로 띄어쓰기나 철자 오류 등으로 인해 형태소 분석에 실패한 단어를 미등록 명사로 분류한다. 따라서, 본 논문에서는 문장에 대한 형태소 분석 결과 미등록 명사 비율이 높으면 그 문장은 문법적인 오류를 많이 포함한다고 가정한다.

표 3은 문서에 포함된 명사 비율에 따라 전체 실험 대상 자료를 5 구간으로 나누고, 각 구간별로 등록 명사와 미등록 명사 비율을 보여준다. 표 3에 따르면 문장 내에 품사 중 명사 비율이 높아질수록 미등록 명사 비율도 증가함을 알 수 있다. 특히, 형태소 분석 결과 명사 비율이 81~100% 구간에서는 미등록 명사로 분류된 어휘의 개수가 등록 명사로 분류된 어휘의 개수보다 4.8배나 더 많은 것을 알 수 있다. 따라서 40자평에 포함된 명사의 비율이 높다는 것은 미등록 명사의 비율을 볼 때 형태소 분석의 오류가 많다는 것을 의미한다.

표 3. 명사 비율에 따른 등록/미등록 명사의 분포
Table 3. Distribution of registered/unregistered nouns according to noun ratio

명사 비율	등록 명사		미등록 명사	
	개수	비율 (%)	개수	비율 (%)
0 ~ 20%	17,472	49.78	17,629	50.22
21 ~ 40%	17,886	43.13	23,586	56.87
41 ~ 60%	12,701	33.04	25,743	67.96
61 ~ 80%	12,651	32.86	25,844	67.14
81 ~ 100%	4,808	17.23	23,102	82.77

표 4는 명사 비율에 따른 전체 169만 40자평 문서의 실제 분포를 보여준다. 이 표에 따르면, 형태소 분석이 비교적 어려운 구간인 61~80% 구간과 형태소 분석이 거의 불가능한 81~100% 구간에 있는 데이터는 전체 데이터 중에 15% 정도 되는 것을 알 수 있다.

표 4. 명사 비율에 따른 실험 데이터 분포
Table 4. Distribution of experiment data according to noun ratio

명사 비율	개수	비율(%)
0 ~ 20%	164,731	9.73
21 ~ 40%	765,525	45.23
41 ~ 60%	506,223	29.91
61 ~ 80%	134,250	7.93
81 ~ 100%	121,968	7.21

4.2 기준 평가 측도

제안한 감정 분류기의 성능을 비교 평가하기 위하여 자연어 처리 분야에서 기준 평가 측도로 일반적으로 사용되는 BOW 모델을 사용하였다. 자질 벡터 생성을 위해 형태소 분석을 수행한 후 품사가 체언이나 용언, 관형사, 부사, 감탄사로 분석된 단어만 자질로 선택하였다. 본 연구에서 실험 대상으로 선정한 영화 40자평은 길이가 40자로 비교적 짧기 때문에 tf · idf(term frequency * inverse document frequency)를 사용하는 대신에 단어의 존재 유무만 구분하는 자질 값을 사용하였다. 따라서 문서에 포함되어 있는 단어일 경우는 자질 값은 1이 되며, 없는 경우는 0이 된다. BOW 모델에서 일반적으로 적용하는 문서의 자질 표현은 출현 단어의 빈도수에 따라 가중치를 부여하는 방법을 사용하지만 최근 Pang 등은[6] 감성 분류 문제에 대해 단어의 빈도수 정보 대신에 단어의 존재 유무만을 자질값으로 사용할 때 더 높은 성능을 얻을 수 있음을 보였다. 기준 평가 측도에서 사용한 기계학습 방법은 선형 커널(linear kernel)에 기반한 지지벡터기계 모델을 사용하였다.

4.3 실험 결과 및 분석

그림 2는 40자평 문서에 포함된 명사 비율에 따라 구분된 구간별 감성 분류기의 정확도 성능을 보여준다. 정확도 성능은 (올바르게 구분한 문서의 수/시험 대상 문서의 수)*100%로 계산한다. BOW 모델을 사용한 기준 평가 측도의 경우는 40자평에 포함된 명사의 비율이 높은 구간일수록 형태소 분석의 오류로 인하여 감성 분류기의 정확도가 떨어짐을 알 수 있다. 특히, 명사가 포함된 비율이 81~100% 구간에서는 다른 구간에 비해 기준 평가 측도의 성능이 크게 하락함을 알 수 있다. 그러나 본 논문에서 제안한 음절 단위의 부분 문자열을 비교하는 감성 분류 모델은 명사 비율이 21~40% 구간을 제외하고는 기준 평가 측도에 비해 높은 정확도 성능을 보여주었다. 특히, 제안한 감정 분류 모델은 40자평에 포함된 명사의 비율이 81~100% 구간에서 기준 평가 측도보다 훨씬 높은 정확도 성능을 보여준다. 또한, 비교적 형태소 분석이 성공적인 구간인 0~20% 구간에서도 제안한 감성 분류 모델은 기준 평가 측도보다 약간 나은 성능을 보여준다.

따라서, 문법적인 오류로 인하여 형태소 분석이 어려운 데이터에서도 제안한 음절 커널 기반 모델은 기준 평가 측도인 BOW 모델에 비해 훨씬 안정적인 성능을 보였고, 형태소 분석이 잘 되는 데이터에서도 제안한 모델은 기준 평가 측도와 비슷한 수준의 성능을 보였다.

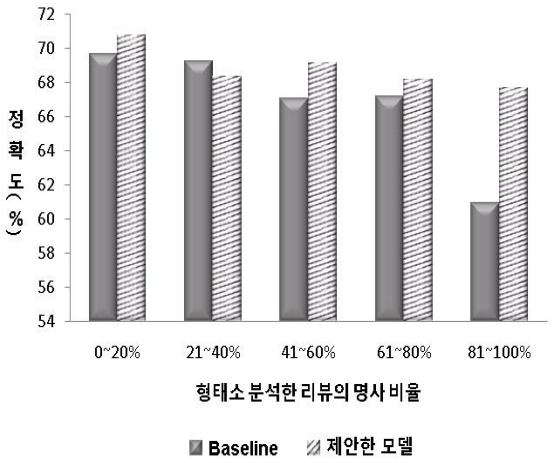

그림 2. 기준 평가 측도와 제안한 모델의 성능 비교
Fig. 2. Performance comparison between baseline and proposed model

정확도 측정을 위한 실험에서 잘못 분류된 결과의 예를 살펴보면 크게 3가지 유형으로 나눌 수 있다. 첫 번째 유형은 사용자가 기재한 평점이 실제 40자평의 내용과 일치하지 않는 경우이다. 예를 들어, “뭔가 진짜 안 웃기구 진짜 재미없다” 라는 40자평은 내용을 볼 때 부정적 감성을 나타내고 있음이 분명하지만, 40자평의 평점을 기재한 사용자는 평점을 6점으로 평가하였다. 이러한 차이는 점수를 평가하는 기준이 개인의 주관적인 견해 차이로 인해 발생하는 것으로 볼 수 있다. 또 다른 예로 “영화이해가잘안갔음”이라는 40자평의 평점으로 9점을 기재한 것은 사용자 개인의 주관적인 견해 차이라고 하기에는 어렵고 기재 오류라 생각된다. 두 번째 유형은 40자평의 내용에 대한 객관적인 평점 결정이 애매한 경우로 40자평의 내용만으로 감성이 어느 쪽인지 알기 어려운 경우이다. 세 번째 유형은 비교급이 사용된 경우로서 “이것보다과속스캔들이훨씬잼있음(평점 4점)”라는 40자평은 평가 대상인 영화보다 다른 영화가 상대적으로 재미있다는 표현이지만 비교급 형태의 자세한 의미를 구분하지 못하여 생긴 오류로 판단된다.

5. 결 론

본 논문에서는 비교적 문서의 길이가 짧고, 띄어쓰기나 철자 오류와 같은 문법적인 오류가 빈번한 영화 40자 평과 같은 문서에 대해 긍정과 부정의 극성으로 분류하는 감성 분류기를 제안하였다. 사용한 기계학습 방법은 이진 분류 문제에서 비교적 높은 성능을 보이는 지지벡터기계를 사용하였으며, 정보 검색 기술 분야에서 일반적으로 사용되고 있는 BOW 모델 대신에 문자열 커널의 확장 모델인 음절 커널을 사용하여 문서 간의 유사성을 측정한다.

실험결과, BOW 모델은 문서에 포함된 명사 비율이 높을수록(문서에 형태소 분석 오류가 많을수록) 형태소 분석기의 오류로 인하여 큰 성능 저하를 보이는데 비해 본 논문에서 제안한 음절 커널에 기반한 감성 분류 방법은 BOW 모델에 비해 훨씬 안정적인 성능을 보였다.

참 고 문 헌

[1] V. Hatzivassiloglou and K.R McKeown, “Predicting the semantic orientation of adjectives,” *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 174-181, 1997.

[2] P.D. Turney and M.L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Trans. Inf. Syst.* 21(4), pp. 315-346, 2003.

[3] P. Chesley, B. Vincent, L. Xu, R.K. Srihari, "Using Verbs and Adjectives to Automatically Classify Blog Sentiment," *Proceedings of American Association for Artificial Intelligence - Spring Symposium Series Technical Reports*, pp. 27-30, 2006.

[4] WordNet (http://wordnet.princeton.edu)

[5] J. Kamps, M. Marx, R.J. Mokken, and M.D. Rijke, "Using WordNet to measure semantic orientation of adjectives," *Proceedings of the 4th International Conference on Language Resources and Evaluation,* pp. 1115-1118, 2004.

[6] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol 10, pp. 79-86, 2002.

[7] 황재원, 고영중, “감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템,” *정보과학회논문지*, 제14권, 제3호, pp. 336-340, 2008.

[8] 명재석, 이동주, 이상구, "반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템," *정보*

과학회논문지. 제35권, 제6호, pp. 392-403, 2008.
[9] 남상협, 나승훈, 이예하, 이용훈, 김준기, 이종혁, "의견 어구 추출을 위한 생성 모델과 분류 모델을 결합한 부분 지도 학습 방법," *한국컴퓨터종합학술대회 논문집*, Vol.35, No.1(C), pp. 268-273, 2008.
[10] 김묘실, 강승식, "SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현," *한글 및 한국어 정보처리 학술대회논문집*, pp. 285-289, 2006.
[11] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification using String Kernels," *Journal of Machine Learning Research*, pp. 419-444, 2002.
[12] http://www.unicode.org/charts/PDF/UAC00.pdf
[13] S. S. Kang, *Korean Morphological Analysis and Information Retrieval*, Hong-Reung publisher, 2002.

저 자 소 개

김상도(Sang-Do Kim)
2008년 : 경일대 컴퓨터공학과(학사)
2010년 : 경북대 컴퓨터공학과(석사)

관심분야 : 오피니언 마이닝, 감정 분석, 시멘틱웹, 온톨로지, 기계학습
Phone : 053-940-8692
Fax : 053-957-4846
E-mail : sdkim@sejong.knu.ac.kr

박성배(Seong-Bae Park)
1994년 : 한국과학원 전산학과(학사)
1996년 : 서울대 컴퓨터공학과(석사)
2002년 : 서울대 컴퓨터공학과(박사)
2004년～현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 자연어처리, 기계학습, 정보검색, 바이오인포매틱스
Phone : 053-950-7574
Fax : 053-957-4846
E-mail : seongbae@knu.ac.kr

박세영(Se Young Park)
1980년 : 경북대 전자공학과(학사)
1992년 : KAIST(석사)
1999년 : 프랑스 파리 7대학(박사)
1982～2000년 : ETRI 책임연구원
2003년～2005년 : 정보통신연구진흥원 전문위원

2005년～현재 : 경북대학교 컴퓨터공학부 교수

관심분야 : 시멘틱웹, 자연어 처리, 정보 검색, 소셜네트워크 분석
Phone : 053-950-6551
Fax : 053-950-2122
E-mail : seyong@knu.ac.kr


이 상 조

이상조(Sang-Jo Lee)
1974년 : 경북대학교 수학교육과(학사)
1976년 : 한국과학원 전산학과(석사)
1993년 : 서울대 컴퓨터공학과(박사)
1976년～현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 자연언어처리, 기계번역, 정보검색, 시멘틱 웹
Phone : 053-950-5552
Fax : 053-957-4846
E-mail : sjlee@knu.ac.kr

김권양(Kweon Yang Kim)
1983년 : 경북대학교 전자공학과(학사)
1990년 : 경북대학교 전자공학과(석사)
1998년 : 경북대학교 컴퓨터공학과(박사)
1983～1988년 : ETRI 연구원
1999년～2000년 : University of Central Florida 방문교수

1991년～현재 : 경일대학교 컴퓨터공학과 교수

관심분야 : 시멘틱웹, 한글공학
Phone : 053-850-7287
Fax : 053-850-7609
E-mail : kykim@kiu.ac.kr