모바일 텍스트의 감성분류를 위한 SVM 기반 음운 커널 기법

(The Phoneme Kernel Technique based on Support Vector Machine for Emotion Classification of Mobile Texts)

김 현 우 † 이 승 룡 ††
(Hyunwoo Kim) (Sungyoung Lee)

요 약 감성분류는 미래 모바일 분야에서 필요로 하는 기술이며, 인간 중심 인터페이스가 중요해지는 현재 핵심적인 기술이다. 특히 모바일 기술의 발전과 함께 SNS, 모바일 메신저, 문자메시지와 같은 단문 메시지는 사용자의 감성을 매우 잘 나타내는 매체라 할 수 있다. 이러한 텍스트로부터 사용자의 감정을 분류해 내기 위해, 문장으로부터 적절한 특징 값을 추출하여야 하며, 이는 감정인지의 신뢰성과 정확성을 높이는데 중요한 역할을 하게 된다. 본 논문은 SVM을 이용하여 140자 이내의 단문 메시지를 긍정 및 부정적 의미로 분류하기 위해 자질을 추출하는 음운 커널 방법를 제안한다. 스트링 커널 함수를 한글에 적용한 음절 커널을 음운 커널로 수정하고, 모바일 텍스트에 적용하기 위한 적절한 소멸계수 λ을 구한다. 이를 통해 음운 커널이 중성/종성의 변화가 많은 모바일 텍스트에 적합한 분류기임을 증명한다.

키워드: 감성분류, 지지벡터기계, 문자열커널, 음운커널, 모바일텍스트

Abstract Emotion Recognition is the technique that is needed in future mobile field, and the key technique in these days when human central interface is important. Specially, the short message(like SNS, mobile messenger and instant message) is the media what is well expressed because of the growth of the mobile techniques. In order to classify the emotion of user from these texts, it's important role that extracts the appropriate feature values from sentences, and that improves reliability and accuracy. This paper proposes a phoneme kernel method that extracts the features in order to classify the short messages within 140 letters as a positive or negative signification. In addition, We change from the syllable kernel adapted the string kernel for Korean and find the optimized decay factor λ for the mobile texts. We prove the classifier using the phoneme kernel that is adapted in the mobile texts.

Keywords: emotion classification, support vector machine, string kernel, phoneme kernel, mobile text

1. 서 론

최근 스마트폰이 급속도로 보급됨에 따라 언제, 어디서나, 쉽게 자신의 상황을 글로써 SNS에 등록을 하거나 모바일 메신저를 통해 단문 메시지를 주고받고 있다. 스마트폰을 통해 생산되는 모바일 텍스트(SNS, 모바일 메신저 등) 일상생활에서 빠르고 편리하게 자신을 표현하는 글로써, 사용자의 감성을 가장 잘 표현하는 매체 중 하나이다. 이러한 텍스트를 기반으로 감성을 분류하는 것은 사용자의 성향 및 스트레스 지수 등으로 재 표현이 가능하며, 점차 인간 중심의 UX가 강조되고 있는 스마트폰 시장에서 중요한 정보로 활용이 가능할 것이다.

스마트폰에서 입력되는 모바일 텍스트는 빠르게 자신의 상황을 전달하고자하는 사용자의 입력 환경으로 인

· 본 연구성과는 중소기업청에서 지원하는 2011년도 산학연공동기술개발사업(No. 00048272)의 연구수행으로 인한 결과물임을 밝힙니다.

† 학생회원 : 경희대학교 컴퓨터공학과
khw@oslab.khu.ac.kr
†† 종신회원 : 경희대학교 컴퓨터공학과 교수
sylee@oslab.khu.ac.kr
(Corresponding author임)

논문접수 : 2012년 12월 6일
심사완료 : 2013년 3월 13일

정보과학회논문지: 소프트웨어 및 응용 제40권 제6호(2013.6)

해 오·탈자가 많고 띄어쓰기 생략 및 오류, 약어나 통신 용어 등의 자주 발생되며, 이러한 모바일 텍스트의 특성으로 인해 형태소 분석에 따른 감성 분류 시스템의 성능을 저하시키게 된다.

본 연구에서는 스마트폰에서 입력된 모바일 텍스트를 기반으로 사용자의 감성을 긍정과 부정으로 이진 분류하는 문제를 다룬다. 모바일 텍스트를 분류하기 위한 알고리즘으로는 이진분류 문제에서 비교적 높은 성능을 보여주는 지지벡터기계(SVM: Support Vector Machine)를 이용한다. 또한, 텍스트를 지지벡터기계에 적용하기 위해 영문에서 사용되는 분류커널인 문자열 커널(String Kernel)을 한글에 적용한 음절 커널(Syllable Kernel)과 본 논문에서 제안하는 음운 커널(Phoneme Kernel)을 비교하고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서 감성 분류 관련 기존연구를 영문과 한글로 나누어 서술하고, 3장에서는 본 연구의 기초가 되는 SVM과 문자열 커널, 음절 커널에 대해 서술하고, 4장에서 본 논문에서 제안하는 음운커널에 대해 설명한다. 5장에서는 제안한 분류기의 성능을 비교, 평가하고, 6장에서 결론을 맺는다.

2. 관련 연구

텍스트로부터 감성을 분류하기 위한 연구는 인터넷 상의 글로부터 대중의 의견을 파악하는 오피니언 마이닝(Opinion Mining)으로부터 시작 되며, 문학작품이나 기사, 영화평 등 장르 분류, 또는 감성 분류를 중심으로 활발히 연구되고 있다. 텍스트에서의 감성분류 연구는 감성 사전 등을 이용한 의미 분석기반의 방법과 기계학습에 의한 방법으로 나눌 수 있다. 의미 분석 기반의 방법은 형태소 분석을 통해 이루어지며, 문서 내 감성키워드와 분석된 단어의 동시 빈출도에 따라 단어별 가중치를 줌으로써 감성 분류를 하는 PMI(Pointwise Mutual Information) 방법[1]과 추출된 형용사를 WordNet에 정의된 어휘 관계(lexical relations)를 사용하여 감성을 분류하는 방법 등이 제안 되었다[2]. 형태소 분석기를 사용할 경우 철자오류에 민감하다는 단점이 있어 이를 개선하기 위한 방법으로 음운 패턴 사전을 통한 원형 복원 방법이 제안되었다[3]. 기계 학습에 의한 방법은 n-gram 단위의 BOW(Bag-Of-Word)로 표현함으로써 단어의 의미 벡터를 구성하는 방법[4]와 String Kernel을 이용한 방법[5]로 나누어진다. BOW를 통한 벡터 추출방법은 형태소 분석기를 이용한 의미분석 기법에 비해 철자 오류에 덜 민감하지만, 통신용어나 두음법칙에 의한 발음 표기 형태의 단어에서 정보량이 분산되어 동일 의미로 추론하는데 문제가 발생한다. String Kernel의 경우 BOW의 문제점을 보완한 방법으로 BOW가 단어를 중심으로 연속부분문자열(Contiguous Substring)만을 활용하는 것에 반해 String Kernel에서는 비연속 부분문자열(Non -contiguous Substring)과 연속문자열을 모두 고려하여 벡터 값을 찾아내므로 동일 의미의 특징 값을 추론하는데 BOW 보다 유리하다.

한국어의 경우 알파벳 대신 음절 단위를 사용하여 String Kernel을 적용하는 시도가 이루어졌으나, 40자 내외의 짧은 영화 감상평을 기준으로 이진 분류하는 문제를 다루고 있다. 하지만 영화 감상평의 경우 주로 문어체이므로 이모티콘이나 통신용어의 사용이 드물다[6].

본 논문에서는 모바일 텍스트에서 보다 높은 정확성을 보이는 감성 분류 알고리즘을 구현하기 위해 음절 보다 작은 단위인 음운(초성, 중성, 종성과 같은 자/모음의 집합)으로 나누어 음운 변형으로 발생된 단어의 의미 비교가 가능한 음운 커널을 제안하고 그에 대해 기술한다.

3. SVM 기반 감성 분류기

이 장에서는 텍스트 감성 분류에서 높은 성능을 보이는 SVM의 구조와 SVM에 문자열을 매핑(mapping)시키기 위한 Kernel 기법에 대해 기술하고, 영어권을 위한 String Kernel과 이를 한국어에 적용한 음절커널에 대해 서술한다.

3.1 Support Vector Machine

SVM은 1992년 Boser 등에 의해 제안된 기계학습 알고리즘으로, 커널 매핑 개념과 최적화 기술을 통계적 학습의 원리에 통합한 알고리즘이다. 이 알고리즘의 단순한 형태로는 두 집합을 분리하기 위해 Margin(평면과 가장 가까운 좌표 간의 거리)을 가장 크게 나타내는 최적화된 Hyperplane을 찾음으로써 두 개의 클래스로 나누는 것이다. SVM은 가장 최적화된 Hyperplane을 찾는 것이 목표이며, 학습 데이터를 통해 Hyperplane을 찾고 이후 입력된 벡터 값의 좌표의 위치에 따라 해당 클래스를 반환한다. SVM은 데이터가 선형 분리가 가능한 상태이어야 한다. 하지만 현실에서 수집되는 대부분의 데이터는 선형 분리가 불가능한 경우가 많다. 이 경우 슬랙변수(Slack Variable)를 도입하거나 커널법을 적용함으로써 극복이 가능하다.

3.2 String Kernel

String Kernel은 Lodhi[5] 등이 제안한 방법으로 문자열 간의 유사성을 단어 중심의 n-gram 대신에 문자열이 포함하고 있는 모든 연속·비연속부분문자열을 비교하여 벡터를 추론하는 방법이다. SVM을 통해 문자열의 최적화된 margin을 찾기 위해서는 선형 분류가 가능하도록 고차원 공간에 데이터 좌표를 매핑시킬 필요가 있다. 이를 위해 두 문자열의 부분문자열(Substring)을 비교하게 된다. 이 방법은 부분문자열간의 연속성이 있는

지에 대한 부분은 중요하지 않으며, 단지 부분문자열의 가중치가 얼마인지가 중요하다. 예를 들어, 부분문자열 'c-a-r'는 'card'와 'custard'에서 모두 존재하는 부분 문자열이다. 하지만 각각의 가중치는 다르다. 이러한 부분 문자열들이 문자열 내에 얼마나 가득 차 있는지, 얼마나 자주 나타나는지에 따라 특징 공간(Feature Space)에 표현되는 값이 결정된다.

유한한 알파벳(공백 포함)의 집합을 Σ라고 할 때, 하나의 문자열은 Σ의 원소(알파벳 문자)들로 구성된 순열이라 할 수 있다. 이 때, 문자열 s를 특징 공간에 매핑하는 함수 Φ는 아래 식과 같이 정의된다.

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

u는 문자열 s에 포함된 부분 문자열이며, $i=(i_1,...,i_{|u|})$는 문자열 s에 존재하는 모든 부분문자열 u에 대한 인덱스 집합이다. 소멸계수(Decay Factor)$\lambda(0 \le \lambda \le 1)$은 비연속부분문자열에 대한 가중치로써 1에 가까울수록 길이가 긴 비연속부분문자열이 커널 값에 미치는 영향이 커지게 되며, 반대로 0에 가까울수록 영향력이 작아지게 된다. 글의 종류에 따라 최대 정확도를 나타내는 소멸계수는 달라지므로 적절한 소멸계수를 설정하는 것이 중요하다.

두 문자열 s와 t의 특징 벡터(Feature Vector)의 내적은 각 문자열의 공통된 부분문자열로부터 얻어진 발생 빈도와 길이에 따라 구하며, 다음 식과 같다.

$$K_n(s,t) = \sum_{u \in \Sigma^n} \langle \Phi_u(s) \cdot \Phi_u(t) \rangle = \sum_{u \in \Sigma^n} \sum_{i : u = s[i]} \lambda^{l(i)} \sum_{j : u = t[j]} \lambda^{l(j)}$$

$$= \sum_{u \in \Sigma^n} \sum_{i : u = s[i]} \sum_{j : u = t[j]} \lambda^{l(i)+l(j)}$$

위 식에서 커널함수 Φ는 부분문자열의 길이 n이 작을 경우에도 연산이 많아 이를 이용하여 벡터를 구하는 데에는 무리가 있다. 그러나 String Kernel에서는 아래와 같은 식으로 재정리함으로써 계산을 단순화시켜 효율적인 계산이 가능하다.

$$K'_i(s,t) = \sum_{u \in \Sigma^n} \sum_{i : u = s[i]} \sum_{j : u = s[j]} \lambda^{|s|+|t|-i_1-j_2+2}$$

$$i = 1,...,n-1$$

개선된 함수 $K'_i(s,t)$는 $l(i)$와 $l(j)$ 대신 특정 문자열의 시작 위치부터 각 문자열 마지막까지의 길이를 이용한다. 이 함수는 동적 프로그래밍 기반의 재귀적 방법에 의해 계산 되며, 이를 바탕으로 K_n을 계산한다.

String Kernel은 문자열의 길이가 길어질수록 값이 커지는 문제가 있다. 이를 해결하기위해 아래 식을 이용하여 정규화를 한다.

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

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

그림 1 String Kernel의 알고리즘
Fig. 1 The algorithm of String Kernel

3.3 음절 커널

음절 커널(Syllable Kernel)은 김상도[6] 등에 의해 String Kernel을 한국어에 맞게 확장한 모델로 String Kernel에서 사용된 개별 문자에 비해 보다 큰 구조적 단위인 음절을 기본 비교 단위로 사용한 커널 함수이다. 비교되는 부분문자열이 알파벳과 같은 개별 문자 대신에 음절에 기반 한다는 점을 제외하고는 String Kernel에서 정의된 동일한 가중치와 유사성 척도를 사용한다. 즉, 문자의 유한 집합인 Σ는 $\{a,b,c,...,X,Y,Z\}$ 대신에 한글 음절 집합인 $\Sigma \in \{가,각,갂,...,힢,힣\}$으로 구성된다.

음절 커널은 주어진 두 문자열이 상호 공유하는 음절 단위의 연속·비연속부분문자열이 많을수록 두 문자열의 유사도를 높게 계산한다.

표 1은 음절 단위 부분문자열에서 두 문자열이 상호 공유하는 부분문자열을 비교한 예시이다. 예에서 사용된 두 문자열은 띄어쓰기나 철자의 오류가 포함되어 있어 비슷한 의미의 두 문장임에도 불구하고 서로 공유하는 단어가 없으므로 별도의 복잡한 어근 추론 모델을 사용하지 않는다면 BOW에서의 성능이 떨어질 수밖에 없다. 그러나 음절 커널은 유사한 감성을 가지는 두 문자열로부터 3음절에서 "진짜재", "진짜밌", "진짜영"이나, 4음절의 "진짜재밌", "진짜영화", "재밌영화" 등 공통 부분문자열을 가지므로 두 문장의 의미가 유사함을 추론 가능하다.

표 1 음절 부분문자열을 통한 공유 문자열 탐색
Table 1 String search using syllable sub-strings

문장A		내 생에 진짜루재밌는영화였어
문장B		그 영화 진짜로 재밌던 옛날영화였지
공유문자열	3음절	"진짜재", "진짜밌", "진짜영", ... , "재영화", "재영였", "영화였"
	4음절	"진짜재밌", "진짜영화", "재밌영화", ...

모바일 텍스트는 빠르게 자신의 상황을 전달하는 목적을 지니므로 오탈자가 많고 약어나 통신용어, 이모티

콘 등이 자주 사용된다. 이러한 환경에서 텍스트 감성 분류기를 적용하기 위해서는 기존의 정적인 분석기법으로는 한계가 존재하며 보다 유연한 분석기가 필요하다.

본 연구에서는 모바일 텍스트의 특징을 초성 중심의 변형, 중성, 종성의 탈락과 같은 오탈자, 음운으로 구성된 이모티콘 사용, 두음 표기에 따른 어휘 변화로 보았다.

표 2 모바일 텍스트의 특징
Table 2 The features of mobile text

음운 단위의 오탈자	“에뻐요” → “이뻐용” “좋아요” → “조아요” “많이 주네” → “마이 주네” “그게 뭔데” → “그게 먼데”
두음 표기	“싫어요” → “시러요” “밥 먹어” → “밥 머거”
음운 중심의 이모티콘 또는 의성어	“ㅠㅠ”, “ㅜㅜ”, “ㅋㅋㅋ”, “:D”, “—-;;”, “ㅇㅋ”, “ㅇㅅㅇ”

표 3 음운패턴과 음절패턴 비교
Table 3 The comparison between the phoneme pattern and the syllable pattern

원문	음운 분석($k=3$)	음절 분석($k=1,2$)
싫어	{ㅅ,ㅣ,ㄹ}, {ㅅ,ㅣ,ㅎ}, {ㅅ,ㄹ,ㅎ}, ...	{싫}, {어}, {싫어}
싫은데	{ㅅ,ㅣ,ㄹ}, {ㅅ,ㅣ,ㅎ}, {ㅅ,ㄹ,ㅎ}, ...	{싫}, {은},{데}, {싫은},{싫데},{은데}
시로	{ㅅ,ㅣ,ㄹ}, {ㅅ,ㅣ,ㄴ}, {ㅅ,ㄹ,ㄴ}	{시}, {로}, {시로}
시러	{ㅅ,ㅣ,ㄹ}, {ㅅ,ㅣ,ㅓ}, {ㅅ,ㄹ,ㅓ}	{시}, {러}, {시러}
시른데	{ㅅ,ㅣ,ㄹ}, {ㅅ,ㅣ,ㅡ}, {ㅅ,ㄹ,ㅡ}, ...	{시},{른},{데}, {시른},{시데},{른데}

이와 같은 어휘 변화는 음절 단위에서는 서로 다른 단어로 의미 추론이 되지만 음절로 나눌 경우 그 패턴을 공유하므로 유사성을 보다 쉽게 찾아낼 수 있다. 표 3은 ‘싫어’라는 단어의 원형 및 변형을 각각 음운과 음절로 나누어 유사성을 비교한 것이다.

4. 음운 커널 기반 텍스트 감성분류

본 논문에서 제안하는 음운커널(Phoneme Kernel) 기법은 String Kernel을 기초로 하여 확장한 한국어 텍스트 감성 분류를 위한 커널함수이다. 기존 한국어 텍스트 감성분류는 음절 단위로 의미를 분석한 것에 반해 음운 커널은 모바일 텍스트의 특징에 알맞은 분석을 위해 음절보다 작은 단위인 음운(초성/중성/종성을 이루는 단위로 자음과 모음을 이르는 단위)으로 분리하여 이들의 부분순서(Sub-sequence)를 비교함으로써 이들의 내적을 구하고 커널 값을 추출하는 방법을 제안한다.

본 논문에서는 한글로 된 모바일 텍스트를 음운 단위의 순서열(Phoneme Sequence)로 변환하여 이를 String Kernel을 통해 벡터 값을 추출하고 이를 통해 SVM이 감성을 분류하는 텍스트 기반 감성추출 모듈을 구축하였다. Training Data는 하나의 텍스트에 대한 감성이 Labeling이 되어있는 음운 순서열이며, 일반적인 모바일 텍스트로부터 변환된 데이터이다. 아래 그림 2는 본 논문에서 제안하는 텍스트 기반 감성분류 기술의 개념도이다.

입력데이터는 일반적인 SNS 데이터로 본 연구에서는 140자 이내의 한글, 영문, 각종 이모티콘을 포함한 트위터 글을 사용하였다. 다만 감성에 영향을 미치지 않는 URL 링크 등은 제거하였으며, 이모티콘은 특수문자, 일어, 한자 등을 포함하는 경우가 많으므로 이러한 문자들은 포함시켰다.

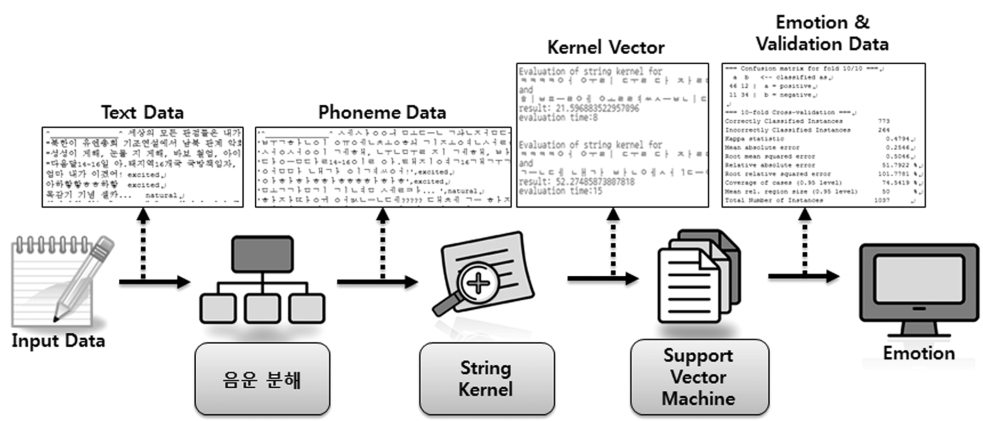

그림 2 제안하는 음운커널 기반 감성 분류기
Fig. 2 A proposed emotion classifier based on phoneme kernel

음운 분해에서는 한글만을 선별하여 초성/중성/종성으로 나누어 분해된 형태로 문자열을 재구성하는 단계로 영문자나 특수문자, 타국어 등은 By Pass 된다.

String Kernel은 분해된 Phoneme Data를 이용하여 Vector값을 추출하는 Kernel 함수이다. Vector 값은 학습 데이터의 유사 부분문자열 패턴을 찾아 구하게 된다. String Kernel은 입력 텍스트 외에도 두 가지의 파라미터를 설정해 주어야 하는데, 최대부분문자열 길이 k와 연속부분문자열과 비연속부분문자열을 반영하는 비율을 지정하는 λ가 그것이다. 이에 대한 설정 값은 4.2절에서 자세히 다루겠다.

4.1 음운 커널

음운커널은 크게 음운 분리와 String Kernel의 조합으로 이루어져 있다. 한글 코드는 UTF-8을 기초로 하였으며, 각 음운의 Base code에 원본 코드를 대입하여 각 음운 코드를 분할한다. 트위터나 채팅 용어에는 특수문자, 영문자, 일본어 등이 이모티콘으로 활용되는 경우가 많으며, 한글은 아니지만 감성을 나타내는데 의미가 있으므로 순서에 맞춰 음운과 통합한다.

4.2 학습모델 생성

String Kernel은 SVM을 통해 학습모델을 생성시키기 위해 두 가지 중요한 파라미터를 설정해주어야 한다. 첫 번째 파라미터는 소멸계수(Decay Factor)로 앞 장의 String Kernel의 소개에서 논한 바와 같이 연속·비연속 부분문자열의 가중치를 주는 값이다. 이 값은 글의 종류나 방식에 따라 다를 수 있으므로 다양한 실험을 통해 도출해야한다. 두 번째 파라미터는 부분문자열의 최대 길이로 Lodhi 등에 의하면 부분문자열의 최대 길이 k가 4보다 커질 경우 처리 속도가 급격히 느려짐을 밝혔다 [5]. 본 연구에서는 음운커널과 음절커널 모두 $k=3$으로 설정하여 모델을 생성하였다.

5. 실험 및 평가

이 장에서는 본 논문에서 제안하는 음운 커널 기반의 SVM 감성 분류기의 성능을 평가하기위한 실험에 대하여 논의한다. 실험의 목적은 본 연구의 차별성으로 언급한 모바일 텍스트에 대한 정확도가 개선됨을 보일 것이며, 음절 커널과 동일한 데이터셋을 이용하여 실험, 비교함으로써 그 타당성을 입증한다. 더불어 모바일 환경에 String Kernel을 적용하기 위해 상수로써 미리 지정되어야 하는 소멸계수(Decay Factor) λ의 최적 값을 찾는 실험을 함께 진행 할 것이다.

5.1 실험 데이터

데이터 셋은 Naver의 실시간검색을 통한 트위터 글을 수집하였으며, 2012년 10월 12일부터 31일까지 무작위 시간 및 임의의 사용자로부터 1037개의 글을 긍정 및 부정으로 분류하였다. 단순 기사 글을 옮긴 트윗은 제외했으며, 비교적 일상적이고 감성이 나타나는 것으로 판단되는 글을 위주로 수집하였다.

5.2 실험 방법

본 논문에서 제안하는 음운 커널 기반의 SVM 감성 분류기를 직접 수집한 데이터 셋을 이용하여 다음과 같은 항목을 기준으로 실험하였다. 정확도는 10-folds Cross-Validation을 통해 산출하였다.

실험 1. 음절과 음운 커널의 소멸계수 λ을 0과 1사이에서 0.1간격으로 실험, 각각 최고의 성능을 보이는 지점 도출

실험 2. 각 구간에 대한 정확도 및 최고 성능간의 비교

5.3 실험 결과

5.3.1 소멸계수별 정확도

트위터를 통해 수집된 1037개의 데이터셋을 음절 커널을 적용한 SVM과 음운 커널을 적용한 SVM에 각각 학습시켜 10-folds Cross-Validation으로 정답률을 도출하였으며, 각 소멸계수별 정답률은 아래와 같다.

소멸계수 0.1~0.2 구간에서 음운 커널과 음절 커널의 정답률은 약 56.5%로 동일하게 유지됨을 알 수 있다. 그러나 0.2~0.6 구간에서는 음운 커널이 음절 커널에 비해 높은 정확도를 보임을 할 수 있으며, 특히 음운 커널의 경우, 0.5에서 약 78.1%로 가장 높은 정답률을 보여준다. 0.7 이후 구간에서는 음절 커널이 더 높은 결과가 나왔으나 그 차이가 2% 이내로 최대 7%의 차이를 보인 0.2~0.6 구간에 비해 미미하다. 또한 음절 커널의 경우 0.7에서 약 74.7%로 가장 높은 정답율을 보여주나, 동일한 값의 음운 커널 정답률인 약 74.5%와 큰 차이가 없으며, 음운 커널의 최고 정답률인 약 78.1%와 약 3.4%의 정답률 차이를 보였다.

음절 커널과 음운 커널이 나타내는 최고 성능구간은 음절에서 0.7, 음운에서 0.5로 다소 차이가 발생하였다. 하나의 글자를 음절에서는 연속·비연속부분문자열의 구분이 없지만 음운에서는 초성, 중성, 종성으로 이루어진 하나의 패턴으로 인식하므로 음절에 비해 비연속부분문자열의 가중치를 줄어든 보다 낮은 소멸계수에서 더 높은 정답률을 나타낸 것으로 보인다.

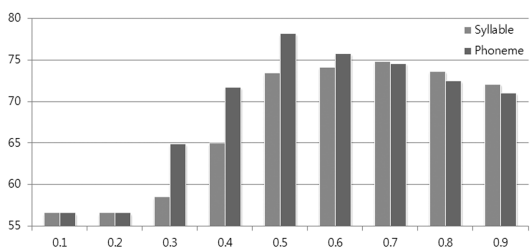

그림 3 음절과 음운커널의 소멸계수별 정답률 비교

Fig. 3 Comparison of a percentage of correct answers by syllable and phoneme kernels

5.3.2 최고 성능 비교

음절 커널과 음운 커널에서 정답률이 가장 높은 모델을 선택하여 각각의 F-1 척도(F-measure), 정확률(Precision), 재현율(Recall) 등을 평가한다. 데이터 셋에서 긍정적 문장을 추출하는 것으로 성능을 분석하는 F-1 척도의 경우에 정답률의 차이보다는 작지만 비교적 음절 커널보다 음운 커널이 더 높은 성능을 나타내고 있다. 뿐만 아니라 다른 성능측면에서도 모두 음운 커널이 뛰어남을 보여주고 있다. 하지만 재현율의 경우 긍정 분류 능력이 음절에 비해 비슷하거나 미미하게 낮은 성능을 보여줘 다소 낮게 나온 경향이 있다. 하지만 현실의 모바일 텍스트의 경우 긍정문과 부정문은 비슷한 비율로 혼재해 있으며, 음운 커널의 긍정 및 부정의 성능차가 더욱 적다는 면에서 안정성이 더 높다고 할 수 있다.

표 4 음절 및 음운 커널의 정답률

Table 4 A percentage of syllable and phonem kernels

		Classified Result			
		Syllable		Phoneme	
		Pos.	Neg.	Pos.	Neg.
Origin	Pos.	500	86	497	89
	Neg.	176	275	138	313

표 5 음절과 음운 커널 성능 비교

Table 5 Comparison of performance of syllable and phoneme kernels

	F1	Presicion	Recall	Overall
Syllable	79.24	73.96	85.32	55.29
Phoneme	81.41	78.27	84.81	61.26

6. 결 론

스마트폰 시장이 급격히 발전하면서, 사용자 중심의 모바일 환경에 대한 요구사항이 갈수록 다양해지고 있다. 또한 기계학습이나 인지 기술 분야가 발전 할수록 사용자의 생활 패턴이나 인간의 행위 및 감성분류에 대한 중요성은 커져가고 있다. 이러한 배경에서 스마트폰을 통해 사용자의 감성을 인지하여 적절한 서비스를 해줄 수 있는 감성 기술은 매우 중요한 기술로 부각되고 있으며, 특히 텍스트 기반 커널 기법은 감성을 인지하기 위한 분류 기술에서 중요한 부분을 차지한다.

본 논문에서는 SVM 알고리즘에 텍스트를 매핑시켜 감성 분류가 가능하도록 하기위한 음운 커널 기법을 논하였다. 여기서 얻는 결과는 각 사용자의 감성에 맞는 맞춤형 서비스를 제공할 수 있는 기반 기술이 될 것이다. 하지만 현재 실험은 단순한 이진 분류 문제만을 다루었으며, 사용자에게 보다 몰입도 높은 서비스를 제공하기 위해서는 보다 다양한 감성으로 분류하기 위한 방법이 고려되어야 한다. 이 후 연구로써 본 연구의 결과물인 음운커널을 다양한 감성에 적용하여 분류 실험을 하고 이에 대한 정확도를 향상 시키는 연구 또한 가치 있는 주제가 될 것이다.

참 고 문 헌

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

[3] J. Shin, H. Kim, "A Robust Pattern-based Feature Extraction Method for Sentiment Categorization of Korean Customer Reviews," *Journal of KIISE : Software and Application*, vol.37, no.12, pp.946-950, Dec. 2010. (in Korean)

[4] 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.

[5] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, "Text Classification using String Kernels," *Journal of Machine Learning Research*, vol.2, pp.419-444, 2002.

[6] S. Kim, S. Park, S. Park, S. Lee, K. Kim, "A Syllable Kernel based Sentiment Classification for Movie Reviews," *Journal of Korean Institute of Intelligent Systems*, vol.20, no.2, pp.202-207, Apr. 2010. (in Korean)

김 현 우

2011년 협성대학교 컴퓨터공학과(학사) 2013년 경희대학교 컴퓨터공학과(석사) 2013년~현재 LG전자 MC연구소. 관심분야는 유비쿼터스컴퓨팅, 상황인지, 모바일 기반 감성인지 등

이 승 룡

1978년 고려대학교 재료공학과(학사). 1987년 일리노이공과대학교 전산과학과(석사) 1991년 일리노이공과대학교 전산과학과(박사). 2001년~현재 경희대학교 컴퓨터공학과(교수). 관심분야는 u-Healthcare, 상황인지, 운영체제 등