k-Structure를 이용한 한국어 상품평 단어 자동 추출 방법

(Automatic Extraction of Opinion Words from Korean Product Reviews Using the k-Structure)

강 한 훈 † 유 성 준 †† 한 동 일 †††
(Hanhoon Kang) (Seong Joon Yoo) (Dongil Han)

요 약 감정어 추출과 관련하여 기존 영어권 연구에서 제시된 방법의 대부분은 한국어에 직접 적용이 쉽지 않다. 한국어권 연구에서 제시된 방법 중 수작업에 의한 방법은 감정어 추출에 많은 시간이 걸린다는 문제점이 있다. 영어 시소러스 기반 한국어 감정어 추출 기술은 한국어와 영어 단어간 일대일 부정합에서부터 기인하는 정확도의 저하를 제고해야 하는 과제를 갖고 있다. 한국어 구문 분석기를 기반으로 한 연구는 출현 빈도가 낮은 감정어를 선정하지 못할 수 있는 문제점을 내포하고 있다. 본 논문에서는 한국어 상품평 중 단순한 문장에서 감정어를 자동으로 추출하는 데 있어 기존에 제안된 한국어권 연구에 상호 보완적으로 정확도를 향상시킬 수 있는 k-Structure(k=5 또는 8) 기법을 제안한다. 단순한 문장이라 함은 패턴 길이를 최대 3으로 한다. 이는 평가 대상 상품(예를 들어 '카메라')의 속성 명 f(예를 들어 카메라의 '배터리')를 기준으로 ±2의 거리에 감정어가 포함되어 있는 문장을 의미한다. 성능 실험은 국내 주요 쇼핑몰로부터 수집한 1,868개의 상품평을 대상으로 미리 주어진 8개의 속성 명에 대한 감정어를 k-Structure를 이용하여 자동으로 추출하고 그 정확도를 평가하였다. 그 결과, k=5일 경우 평균 79.0%의 재현률, 87.0%의 정확률을 보였고, k=8일 경우 평균 92.35%의 재현률, 89.3%의 정확률을 얻을 수 있었다. 또한, 영어권 연구에서 제안된 방법 중 PMI-IR(Pointwise Mutual Information-Information Retrieval) 기법을 이용하여 실험을 수행하였다. 이 결과, 평균 55%의 재현률과 57%의 정확률을 보였다.

키워드 : 감정어, 리뷰 패턴, 상품 속성, 오피니언 마이닝

Abstract In relation to the extraction of opinion words, it may be difficult to directly apply most of the methods suggested in existing English studies to the Korean language. Additionally, the manual method suggested by studies in Korea poses a problem with the extraction of opinion words in that it takes a long time. In addition, English thesaurus-based extraction of Korean opinion words leaves a challenge to reconsider the deterioration of precision attributed to the one to one mismatching between Korean and English words. Studies based on Korean phrase analyzers may potentially fail due to the fact that they select opinion words with a low level of frequency. Therefore, this study will suggest the k-Structure (k=5 or 8) method, which may possibly improve the precision while mutually complementing existing studies in Korea, in automatically extracting opinion words from a simple sentence in a given Korean product review. A simple sentence is defined to be composed of at least 3 words, i.e., a sentence including an opinion word in ±2 distance from the attribute name (e.g., the

· 서울시 산학연 협력사업(10581)'의 지원을 받아 수행된 연구임
† 학생회원 : 세종대학교 컴퓨터공학과
kangcom@paran.com
†† 종신회원 : 세종대학교 컴퓨터공학과 교수
sjyoo@sejong.ac.kr
(Corresponding author임)
††† 비 회 원 : 세종대학교 컴퓨터공학과 교수
dihan@sejong.ac.kr
논문접수 : 2009년 12월 8일
심사완료 : 2010년 4월 2일

정보과학회논문지: 소프트웨어 및 응용 제37권 제6호(2010.6)

'battery' of a camera) of a evaluated product (e.g., a 'camera'). In the performance experiment, the precision of those opinion words for 8 previously given attribute names were automatically extracted and estimated for 1,868 product reviews collected from major domestic shopping malls, by using k-Structure. The results showed that k=5 led to a recall of 79.0% and a precision of 87.0%; while k=8 led to a recall of 92.35% and a precision of 89.3%. Also, a test was conducted using PMI-IR (Pointwise Mutual Information - Information Retrieval) out of those methods suggested in English studies, which resulted in a recall of 55% and a precision of 57%.

Key words : Opinion Word, Review Pattern, Product Attribute, Opinion Mining

1. 서 론

본 논문은 오피니언 마이닝 관련 연구 중 속성단어를 포함하는 한국어 상품평으로부터 속성단어와 관련된 감정어를 추출하는 방법을 제안한다. 오피니언 마이닝 연구는 1)문서 단위 오피니언 마이닝 연구[1-6], 2)속성 단위 오피니언 마이닝 연구[7-14]로 나눌 수 있다. 1)에서는 전통적인 문서 분류기법에서 다룬 기계학습 기법 또는 통계적 기법 등에 의해 문서 전체의 의미 극성(Polarity)을 긍정 또는 부정으로 분류하는 방식이다. 2)에서는 문장 단위에서 속성명을 추출하고, 속성명과 인접한 곳에서 감정어를 추출한다. 그런 후에 감정어에 따른 의미 극성을 결정한다. 이때, 속성명 또는 감정어를 추출하기 위해 단어의 품사는 중요한 역할을 한다. 속성명은 명사일 확률이 높고, 감정어는 동사나 형용사일 확률이 높기 때문이다. 따라서 2)와 관련된 연구에서는 상품의 속성별 의미 극성을 분석하기 전에 전처리 과정으로 POS(Part-Of-Speech) Tagger나 형태소 분석기를 사용하여 품사의 태깅을 수행한다.

예를 들어 (1)과 같은 예문에서 속성명은 "화질(명사)"이고, 감정어는 "깨끗합니다(형용사)"이다. 그리고 그에 따른 의미 극성은 "긍정"이다. 궁극적으로 2)의 연구 목표는 모든 상품의 속성에 대해 긍정과 부정을 결정하여 요약된 형태로 사용자에게 제공해주고자 하는 것이다.

"화질이 정말 깨끗합니다." (1)

따라서 2)와 관련된 대부분의 연구에서는 상품평으로부터 속성단어 추출 방법, 감정어 추출 방법, 감정어에 따른 의미 극성의 결정 방법에 대한 해결책을 다양한 형태로 제시한다.

영어권 연구에서 제안한 감정어 추출 기법들은 영어를 대상으로 구문분석기, POS-Tagger등을 사용했기 때문에 한국어에 직접적 적용은 불가능한 것들이 대부분이다. 한국어 문서에 대한 연구 중에는 수작업 분석을 통해 감정어를 찾는 방법이 있고[12], 영어 시소러스를 기반으로 감정어를 찾는 연구[13]가 있다. 이외에도 한국어 구문분석기를 기반으로 감정어를 찾는 연구도 있다[14]. [12]의 경우 감정어를 찾는 데 있어 수작업 분석 과정을 거치므로 정확성은 향상될 수 있지만 감정어를 찾는 시간이 오래 걸린다. [13,14]는 자동적인 방법으로 감정어를 추출하지만 [12]보다 정확도가 낮을 수 있다. [13]은 영어 시소러스를 기반으로 감정어를 추출하기 때문에 한국어와 영어 단어가 일대일로 일치하지 않는데서 오는 정확도의 저하를 제고해야 한다. [14]는 한국어 구문 분석기를 이용한 방법으로 출현 빈도가 낮은 감정어를 선정하지 못하는 경우가 있을 수 있다.

본 논문에서는 한국어로 작성된 상품평 문서로부터 자동으로 감정어를 추출하는 데 있어서 기존 연구와 상호보완적으로 사용할 수 있는 방법을 제안한다. [12]는 감정어를 추출하는 데 있어서 높은 정확도를 요구할 때 사용할 수 있다. [13]에서는 일부 단어(14개)를 씨드 워드로 주고, 영어 시소러스 사전을 이용하여 유의어를 확장하였다. 그러나 본 논문에서 찾은 47개의 감정어를 [13]의 씨드 워드로 주어 유의어로 확장한다면 추출 정확도가 향상될 수 있을 것으로 보인다. [14]는 구문 분석기를 이용한다. 형태소 분석기를 이용할 때는 본 논문의 방법을 적용하여 감정어를 추출할 수 있다.

본 논문에서는 k-Structure(k=5 또는 8) 기법을 제안한다. k-Structure는 한국어 상품평을 분석했을 때 감정어가 포함되어 있을 확률이 높은 5개 또는 8개의 문장 구조를 발견하여 정리한 것이다. 각 문장 구조에는 형태소 분석기를 통해 찾아낸 품사 정보가 포함되어 있다.

본 논문은 패턴 길이를 최대 3으로 하는 단순한 문장에서의 감정어 자동 추출에 대한 연구로만 한정한다. 단순한 문장은 속성명 f를 기준으로 ±2의 거리에 있는 단어(w_i)를 가지고 해당 속성에 대한 감정어가 포함되어 있는 문장을 의미한다.

[단순한 문장의 예]

1. "배송(f)이 빠르네요(w_1)."
2. "화질(f)이 정말(w_1) 깨끗합니다(w_2)."
3. "너무(w_1) 저렴한(w_2) 가격(f)"

위와 같은 단순한 문장은 태깅된 품사정보를 통해 다음과 같은 패턴임을 알 수 있다.

[품사가 태깅된 단순한 문장의 예]

1. 명사('배송'), 형용사('빠르네요')

2. 명사('화질'), 부사('정말'), 형용사('깨끗합니다')
3. 부사('너무'), 형용사('저렴한'), 명사('가격')

즉, 분류 대상의 문장이 이 패턴 중 하나에 해당되는 구조를 갖는다면 이 문장은 감정어를 포함하고 있을 확률이 높다고 볼 수 있다.

한편, 복잡한 문장은 속성명 f를 기준으로 $\pm d(d>2)$의 거리에 있는 단어(w_i)를 가지고 해당 속성에 대한 감정어를 찾을 수 있는 것으로 부정어도 포함될 수 있다. 이러한 문장 구조는 향후 연구에서 다루도록 한다.

[복잡한 문장의 예]

1. "가격(f)이 그렇게(w_1) 싼(w_2) 것(w_3) 같지(w_4) 않습니다(w_5)."
2. "배송(f)이 생각한(w_1) 것(w_2) 보다(w_3) 너무(w_4) 빠릅니다(w_5)."

성능 실험은 국내 주요 쇼핑몰로부터 수집한 실제 상품평 데이터에 k-Structure 방법을 적용하여 감정어를 추출하고 평가하였다. 영어권 연구에서 제안한 것 중 PMI-IR 기법[1]을 본 연구에 적용하여 k-Structure 기법과 정확도를 비교하였다.

본 논문의 2절에서는 관련 연구의 문제점에 대해 예를 들어 설명한다. 3절에서는 k-Structure를 이용하여 기존 연구의 문제점을 해결할 수 있는 방법을 설명한다. 또한 k-Structure에 대한 자세한 설명과 감정어를 자동으로 추출할 수 있는 방법에 대해 설명한다. 4절에서는 제안하는 시스템 구조를 보여준다. 5절에서는 실험 내용 및 결과에 대해 설명한다. 마지막으로 6절에서는 결론을 맺는다.

2. 관련 연구

오피니언 마이닝 관련 연구에는 크게 문서단위로 처리하는 방법과 속성 단위로 처리하는 방법이 있다. 감정어 추출 관련 연구는 주로 속성 단위 오피니언 마이닝 연구에서 이루어지고 있다. 이 절에서는 각각의 관련 연구 내용과 문제점에 대해 기술한다.

2.1 문서 단위 오피니언 마이닝 연구

[1]은 PMI-IR 기법을 이용하여 문서 전체를 대상으로 긍정 또는 부정으로 분류하였다. PMI-IR은 특정한 문장의 패턴을 만족하는 여러 개 구문들의 Semantic Orientation(SO)[1]을 계산한다. 그런 후에 각 구문에 대한 SO의 총합이 양수일 경우 긍정으로, 음수일 경우 부정으로 분류하는 방법이다.

[2-5]는 전통적인 주제기반 문서 분류의 개념을 오피니언 문서에 적용하여 긍정 또는 부정으로 분류하였다. 이 중 [2]는 Score Function 계산식을 이용하여 확률적으로 긍정 또는 부정을 결정한다. [3-6]에서는 기계학습 알고리즘을 적용하여 문서의 긍정 또는 부정을 결정한다.

[6]에서는 자체적으로 정의한 계산식을 이용하여 Document Sentiment Value(DSV)를 산출한다. 그런 후에 DSV의 값에 따라 긍정 또는 부정으로 분류하였다.

2.2 속성 단위 오피니언 마이닝과 감정어 추출에 관한 연구

[7]에서는 명사를 속성명이라고 가정하고, 속성명에 인접한 형용사를 감정어라고 판단한다. 예를 들어 다음 문장에서 'picture'는 명사로서 속성명이고, 'clear'는 형용사로서 감정어이다.

"The pictures are very clear."

실제로는 속성명과 인접한 동사/명사가 감정어일 수 있음에도 불구하고 이를 감정어로 판단하지 못하는 문제를 갖는다. 이에 따라 감정어 추출의 재현율이 높지 못할 수 있다.

[8]은 영문 WordNet을 이용하여 감정어를 추출한다. 초기에 다수의 감정어를 Seed Word로 주고, 각 Seed Word에 대한 유의어와 반의어를 확장하여 감정어 사전을 구축한다. 그러나 Seed Word에 대한 유의어/반의어 확장시 감정어에 해당되지 않는 단어가 추출되는 경우가 있다. 예를 들어 'regular(보통의)'를 Seed Word로 주었을 때, 이에 대한 유의어로 'amend(수정하다)', 'lawful(합법적인)'이 파생될 수 있다. 이렇게 해서 수집한 감정어 집합 역시 정확률(precision)이 떨어지는 문제가 있을 수 있다.

[9]는 MINIPAR라는 영어 구문 분석기의 결과를 분석하여 감정어 추출 규칙을 정의하였다. 예를들어 "lamp has problem"이라는 문장 내에 주어('lamp')와 서술어('has'), 목적어('problem')가 있을 때 주어가 리뷰 대상일 경우 목적어를 감정어로 취급한다. [9]에서는 이러한 규칙 10가지를 정의하였다. 그러나 위의 예문을 한국어로 번역한 예문("램프는 문제를 가지고 있다.")을 한국어 구문 분석기로 분석한 결과 주어('램프')와 목적어('문제')는 명확히 찾지만 서술어는 찾지 못하였다. 또한 번역한 예문을 한국어 문맥상에 자연스럽게 표현하면 " 램프에 문제가 있다."로 번역하여야 올바른 문장인데, 이 문장을 한국어 구문 분석기를 적용하였을 때, 의도한 것과는 다른 주어('문제')를 찾았다. 서술어와 목적어는 추출하지 못하였다. 즉, [9]에서 정리한 규칙 10개를 한국어, 영어 문장 구조의 차이상 한국어 문장에 그대로 적용하기가 어렵다. 이러한 문제를 해결하기 위해 우리는 k-개(k=5 또는 8)의 한국어 문장 구조 관련 규칙을 찾아내었다.

[10,11]에서는 감정어 후보와 속성명을 인자로 하여 PMI-IR식을 계산한다. PMI-IR값이 가장 높은 것을 해당 속성에 대한 감정어로 취급하는 방법이다. 실험 결과

우리가 제안한 k-Structure 방법을 적용하여 감정어를 추출한 것보다 PMI-IR 방법을 적용한 정확도가 더 낮았다.

[12]에서는 수작업 분석을 통해 감정어를 찾아낸다. 현재까지 10개 카테고리에 대해서 약 9,000개 정도의 감정어를 구축하였다. 그러나 감정어를 찾는 데 시간이 많이 걸리므로 모든 분야의 감정어를 찾아내는 데에는 많은 시간이 소요될 것으로 예상된다. 우리 연구는 자동으로 감정어를 추출하는 연구이므로 이 연구는 비교 대상에서 제외한다.

[13]에서는 네이버 포털 사이트에서 제공하는 영어 유의어 사전을 활용하여 대표 어휘를 선정하고 어휘의 확장을 통해 감정어 워드넷을 구축하였다. 그러나 영어 감정어를 한국어로 번역하는 과정에서 리뷰 대상 문서에서 잘 사용되지 않는 단어로 번역될 경우가 종종 발생할 수 있다는 사실을 발견했다. 예를 들어 'inexpensive', 'cheap'은 네이버 포털 사이트에서 번역한 결과 각각 '덜 비싼','돈이 적게 드는'의 의미로 번역할 수 있었다. 그러나 실제의 상품평에서는 이러한 단어 대신 '저렴하다', '싸다'와 같은 단어가 많이 활용되고 있다.

[14]에서는 한국어 구문 분석기를 통해 상품평을 분석하였다. 구문 분석 결과 서술어(Predicate)가 감정어가 될 수 있다. 이들 중 문서 전체에서의 출현 빈도수가 임계치 이상이 되는 것을 감정어로 채택하는 방법을 제안했다.그러나 실제로는 임의의 단어가 임계치 이하의 값을 가지더라도 감정어인 경우가 종종 있을 수 있어서 이런 단어는 감정어로 선정이 되지 않을 수 있다. 예를 들어 다음 문장에서 '굿굿굿'은 긍정의 의미를 갖지만 많이 출현하지 않는 단어이기 때문에 감정어로 채택되지 않을 수 있다.

"배송이 굿굿굿입니다."

3. k-Structure

본 논문은 기존 연구에서 고려하지 못한 사항들의 개선을 위해 k-Structure를 제안하였고, 본 절은 k-Structure에 대한 설명과 이를 적용하여 자동으로 감정어를 추출하는 방법에 대해 기술한다.

3.1 감정어 후보 문장 추출

본 절에서는 상품평으로부터 감정어 후보 문장의 추출 방법을 설명한다. 본 논문에서 대상으로 하는 감정어 후보 문장은 패턴 길이를 최대 3으로 하는 단순한 문장(ss)이다. 단어의 최대 패턴 길이가 4이상인 것은 복잡한 문장(sc)이라 하여 본 연구의 범위에 포함하지 않는다.

ss는 다음과 같이 속성명 f와 인접해서 좌우측으로 2개의 단어(w_i)만 존재해도 해당 속성명에 대한 감정어를 추출할 수 있는 문장을 의미한다.

$$ss_1 = \{f, w_1, w_2\}$$
$$ss_2 = \{w_1, w_2, f\}$$

반면에 sc는 다음과 같이 속성명 f와 인접해서 좌우측으로 $d(d>2)$개의 단어(w_i)가 존재하여야 해당 속성명에 대한 감정어를 추출할 수 있는 문장을 의미한다.

$$sc_1 = \{f, w_1, w_2, \ldots, w_k\}$$
$$sc_2 = \{w_1, w_2, \ldots, w_k, f\}$$

k-Structure는 단순한 문장을 대상으로 형태소 분석을 하여 품사 태깅한 결과, 감정어가 포함될 확률이 높은 문장의 패턴을 일반화한 것이다. 본 논문에서는 5-Structure(표 1, 표 2)와 8-Structure(표 3, 표 4)를 제안하고 이를 이용하여 감정어 후보 문장을 추출한다.

표 1 5-Structure

구조번호	품사패턴	구조번호	품사패턴
1	*NV*	4	*VN*
2	*NZV*	5	*ZVN*
3	*NNN*		

N: 명사, *V*: 동사/형용사, *Z*: 부사

표 2 5-Structure에 해당하는 문장의 예문

구조번호	예문
1	배송(*N*)이 빠르다(*V*)
2	화면(*N*)이 정말(*Z*) 크다(*V*)
3	품질(*N*)에 완전(*N*) 만족(*N*)함
4	싼(*V*) 가격(*N*)
5	너무(*Z*) 빠른(*V*) 배송(*N*)

k-Structure의 품사 태깅 결과는 형태소 분석기[15]에 의존적이다. 8-Structure는 5-Structure에서 올바르게 태깅하지 못하여 놓칠 수 있는 구조까지 포함한 것이다. 예를 들어 표 2의 구조 번호 4의 "싼 가격"에서 '싼'은 '싸다'를 기본형으로 하였을 때 형용사(*V*)이기 때문에 올바르게 태깅하였다. 그러나 "선명한 화질"에서 '선명하다(기본형)'는 형용사이지만 본 논문에서 사용한 형태소 분석기는 '선명하다'에서 '선명'을 명사로 추출하여 태깅한다. 즉, 형용사 및 동사인 일부 단어를 명사로 태깅하는 경우가 있었다. 8-Structure에서는 올바르게 태깅하지 못한 것들도 고려한다. 표 3의 구조 번호 중 [n]-2의 구조번호는 형태소 분석기가 동사/형용사를 명사로 태깅하는 경우이고, [n]-1은 동사/형용사의 품사를 갖는 단어가 본래의 의미대로 태깅하는 경우이다.

3.2 감정어 추출

3.1에서는 k-Structure의 구조에 대해 기술하였다.

본 절에서는 감정어 후보 문장으로부터 감정어를 추출하는 과정을 설명한다. 본 논문에서 정의한 규칙은 한

표 3 8-Structure

구조번호	품사패턴	구조번호	품사패턴
1	NV	4-1	VN
2-1	NZV	4-2	N_1N_2
2-2	NZN	5-1	ZVN
3	$N_1N_2N_3$	5-2	ZN_1N_2

표 4 8-Structure에 해당하는 문장의 예문

구조번호	예문
1	배송(N)이 빠르다(V)
2-1	화면(N)이 정말(Z) 크다(V)
2-2	가격(N_1)에 너무(Z) 만족(N_2)함
3	품질(N_1)에 완전(N_2) 만족(N_3)함
4-1	싼(V) 가격(N)
4-2	선명(N_1)한 화질(N_2)
5-1	너무(Z) 빠른(V) 배송(N)
5-2	정말(Z) 저렴(N_1)한 가격(N_2)

국어 형태소 분석기의 분석 결과를 기반으로 한 것이다.

정의 1 (리뷰 문서 R)

리뷰 문서 R은 n개의 연속된 문장으로 이루어진다. 각 문장은 $S_{i(i=1,\ldots,n)}$로 표시한다.

$$R = S_1 S_2 \ldots S_n$$

(예) R ='배송이 빠릅니다(S_1).
디자인도 예쁘네요(S_2).'

정의 2 (문장 S_i)

문장 S_i는 m개의 연속된 단어로 이루어진다. 각 단어는 $w_{i(i=1,\ldots,m)}$로 표시한다.

$$S_i = w_1 w_2 \ldots w_m$$

(예) S_i ='배송이(w_1) 빠릅니다(w_2)'

정의 3 (품사 P_i)

임의의 단어 w_i 에 대한 품사는 P_i 로 표시한다.

정의 4 (품사 패턴 $Pattern_i$)

표 3(8-Structure 기준)에 따라 구조 번호 i에 대한 품사 패턴은 $Pattern_i$로 표시한다. $Pattern_i$는 z 개 품사의 연속된 나열로 이루어진다.

$$Pattern_i = P_1 P_2 \ldots P_z$$

8-Structure에서는 8개의 패턴을 정의했으며, 각 패턴은 $Pattern_1 = NV, Pattern_{2-1} = NZV$, ... ,

$Pattern_{5-2} = ZN_1N_2$로 표현할 수 있다.

정의 5 (품사 태깅 $POSTag(s_i)$)

$POSTag(s_i)$는 임의의 문장을 입력으로 받아 각 단어의 품사를 태깅하는 함수이다. $POSTag(s_i)$의 결과로 다음과 같은 집합을 구성할 수 있다.

$$POSTag(s_i) = \{w_1 : P_1, w_2 : P_2, \ldots, w_k : P_k\}$$

(예) {'배송':N, '빠릅니다' : V}

정의 6 (구조 $ST(s_i)$)

구조 $ST(s_i)$는 $POSTag(s_i)$의 원소 중 P_i만 추출하여 품사 시퀀스를 구성하는 함수이다. $ST(s_i)$의 결과로 다음과 같은 집합을 구성할 수 있다.

$$ST(s_i) = P_1 P_2 \ldots P_l$$

(예) ST('배송이 빠릅니다')→NV

정의 7 ($Value(x)$)

$Value(x)$는 $POSTag(s_i)$로부터 품사 P_i와 연결되는 단어 w_i를 구하는 함수이다.

$$Value(P_1) \rightarrow w_1$$

(예) $Value(N)$→'배송'

정의 8 (속성 집합 F)

속성 집합 F는 제품의 속성명으로 사용되는 단어로 이루어진 집합이다.

$$F = \{f_1, f_2, \ldots, f_n\}$$

(예) {'배송','디자인', ... }

정의 9 (감정어 집합 $OPWord$)

감정어 집합이란 특정 속성에 대해 감정을 표현한 단어로 이루어진 집합이다.

$$OPWord = \{ow_1, ow_2, \ldots, ow_l\}$$

(예) {'추천하다','빠르다', ...}

정의 10 (강조어 집합 $Strength$)

강조어 집합이란 감정어 앞에서 감정어에 대한 강도를 표현하는 단어로 이루어진 집합이다.

$$Strength = \{sw_1, sw_2, \ldots, sw_n\}$$

(예) {'너무','정말', ...}

앞서 서술한 10개의 정의를 바탕으로 5-Strucrue Rules과 8-Structure Rules을 정의한다.

그림 1은 표 1에 나타난 품사 패턴을 통해 감정어를 추출하기 위한 규칙을 정의한다.

if $ST(s_i) = Pattern_1$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
if $ST(s_i) = Pattern_2$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
and put $Value(Z)$ into $Strength$
if $ST(s_i) = Pattern_3$ and $Value(N_1) = f_i$
then put $Value(N_3)$ into $OPWord$
and put $Value(N_2)$ into $Strength$
if $ST(s_i) = Pattern_4$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
if $ST(s_i) = Pattern_5$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
and $Value(Z)$ into $Strength$

그림 1 5-Structure Rules

if $ST(s_i) = Pattern_1$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
if $ST(s_i) = Pattern_2$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
and put $Value(Z)$ into $Strength$
if $ST(s_i) = Pattern_3$ and $Value(N_1) = f_i$
then put $Value(N_2)$ into $OPWord$
and put $Value(Z)$ into $Strength$
if $ST(s_i) = Pattern_4$ and $Value(N_1) = f_i$
then put $Value(N_2)$ into $OPWord$
and put $Value(N_2)$ into $Strength$
if $ST(s_i) = Pattern_5$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
if $ST(s_i) = Pattern_6$ and $Value(N_2) = f_i$
then put $Value(N_1)$ into $OPWord$
if $ST(s_i) = Pattern_7$ and $Value(N) = f_i$
then put $Value(V)$ into $OPWord$
and $Value(Z)$ into $Strength$
if $ST(s_i) = Pattern_8$ and $Value(N_2) = f_i$
then put $Value(N_1)$ into $OPWord$
and $Value(Z)$ into $Strength$

그림 2 8-Structure Rules

표 5 규칙에 따른 감정어 및 강조어 추출 예

구조번호	감정어 ($OPWord$)	강조어 ($Strength$)
1	V = 빠르다	없음
2-1	V = 크다	Z = 정말
2-2	N_2 = 만족	Z = 너무
3	N_3 = 만족	N_2 = 완전
4-1	V = 싼	없음
4-2	N_1 = 선명	없음
5-1	V = 빠른	Z = 너무
5-2	N_1 = 저렴	Z = 정말

그림 2는 표 3에 나타난 품사 패턴을 통해 감정어를 추출하기 위한 규칙을 정의한다. 표 5는 그림 2의 규칙을 통해 추출한 감정어와 강조어의 예를 보여준다.

4. 시스템 구조

본 절에서는 k-Structure를 적용한 감정어 자동 추출 시스템의 구조 대해 기술한다. 시스템은 Crawling, Morphemic Analyzer, POS Pattern Extraction, Opinion Word Extraction 단계를 거쳐 감정어를 추출한다. 여기서, POS Pattern Extraction의 결과인 pp가 k-Structure(k-S.)의 비교 대상이 된다. 각 단계별 주요 기능은 이후에 설명한다.

Crawling[16]단계에서는 국내 주요 쇼핑몰 3곳으로부터 '모니터', '노트북', '디지털 카메라', 'MP3 플레이어' 카테고리에 대한 상품 평을 수집하여 Product Review DB에 저장한다.

Morphemic Analyzer는 Product Review DB로부터 리뷰 문서 r_i와 속성 f_j를 가져온다. 그런 후에 r_i의 문장을 분석하여 품사 태깅한다. 본 논문에서 사용한 속성 f는 [12]에서 정의한 것 중 8개를 선택하였다.

f = {*가격, 디자인, 배송, 색상, 품질, 소음, 음질, 화질*}

POS Pattern Extraction 단계에서는 품사 태깅된 tp로부터 속성명 f_j를 기준으로 ±2의 단어를 추출한다(pp). pp를 k-Structure와 비교하여 k-Structure에 해당한다면 Opinion Word Extraction단계에서 감정어 추출 규칙에 따라 감정어와 강조어를 추출한다. 추출한 감정어와 강조어는 Opinion Word DB로 추가된다. pp가 k-Structure에 해당하지 않는다면 동일한 리뷰 r_i에 대해서 다음 속성 명 f_{j+1}을 기준으로 동일한 분석과정이 진행된다. 하나의 r_i에는 k개의 속성명(예:'배송', '디자인')과 그에 따른 감정어(예:'빠릅니다', '예쁘네요')가 포함되어 있다.

r_i = {"배송이 빠릅니다. 디자인도 예쁘네요"}

따라서 r_i에 포함되어 있는 모든 속성과 그에 따른 감정어를 추출하기 위해서는 f_1부터 f_n까지 반복하여야 한다.

그림 3에 나타난 시스템 구조도는 하나의 리뷰 r_i를 분석하는 과정으로 r_n까지 분석하고자 한다면 그림 3의 과정을 n번 반복하여야 한다.

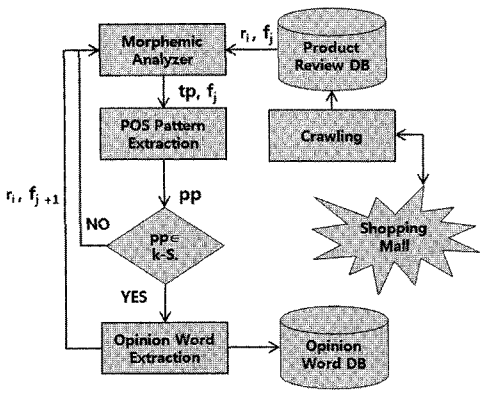

그림 3 감정어 자동 추출 시스템

5. 실험

본 절에서는 k-Structure를 이용하여 실험한 방법과 결과를 분석한다. 아울러 영어권 연구에서 제안한 PMI-IR 기법을 본 논문의 데이터에 적용하여 k-Structure의 정확도와 비교하였다.

5.1 실험 데이터

실험 데이터는 래퍼 기반 크롤러[16]를 이용하여 국내 주요 쇼핑몰 3곳으로부터 수집하였다. 실험에 사용한 상품의 카테고리는 모니터, 노트북, 디지털카메라, MP3플

표 6 상품평 수집 건수

	M1	N	D	M2	합계
S1	8,724	3,919	250	769	13,662
S2	3,010	2,413	3,106	381	8,910
S3	19,731	11,329	5,804	2,275	39,139
합계	31,465	17,661	9,160	3,425	61,711

S1:사이트1, S2:사이트2, S3:사이트3

M1:모니터, N:노트북, D:디지털카메라, M2:MP3플레이어z

표 7 사이트별 8-Structure의 구성 비율

구조번호	S1	S2	S3
1	8.8%	10.5%	14.7%
2-1	19.5%	20.4%	21.7%
2-2	9.4%	7.9%	8.7%
3	4%	3.1%	3.6%
4-1	4.3%	3.5%	4.9%
4-2	10.9%	16.7%	16.8%
5-1	2.2%	2.0%	1.7%
5-2	0.7%	0.4%	0.2%
기타	40.2%	35.5%	27.7%
	100%	100%	100%

레이어이다. 총 수집한 상품평의 수는 61,711개이며, 수집된 데이터의 세부 내역은 표 6에서 보여주고 있다.

표 7은 수집한 상품평을 분석하여 8-Structure의 구성 비율을 산출한 것이다. S1 사이트에서 13,662건의 상품평 중 약 8,170건이 8-Structure에 해당하였고, S2사이트에서는 약 5,747건, S3 사이트에서는 약 28,297건이 8-Structure에 해당하였다.

8-Structure에 해당하지 않는 기타 구조에서도 감정어가 발견될 수 있다. 그러나 그 감정어의 수가 많지 않다. 이는 8-Structure를 적용하여 실험한 결과 중 재현율을 통해 알 수 있다. 실험 결과에서 평균 재현율이 92%로 나타나 있는데, 이는 리뷰 문장 중 8-Structure에 해당하는 것 중에 감정어가 포함되어 있는 비율을 나타낸다. 나머지 8%는 기타 구조의 문장에서 감정어가 포함되어 있는 비율을 나타낸다.

5.2 k-Structure를 이용한 감정어 추출 실험 및 결과

Product Review DB로부터 각각의 속성명을 포함하는 300개씩 총 2,400개(8개의 속성명 * 300개의 상품평)를 추출하였다. 그 중 일부 복잡한 문장은 실험 대상에서 제외하였다. 표 8은 추출한 상품평 중 단순한 문장과 복잡한 문장의 현황을 보여준다.

실험 방법은 단순한 문장으로부터 *pp*를 추출한 후에 8개 각각의 속성 명을 기준으로 k-Structure에 해당하는 문장 중에서 올바른 감정어를 포함하는 문장의 수(TP), k-Structure에 해당하는 문장 중에서 감정어를 포함하고 있지 않는 문장의 수(FP), 그리고 k-Structure

표 8 문장 유형별 추출된 상품평 수

	단순한문장	복잡한문장	합계
가격	220	80	300
디자인	236	64	300
배송	275	25	300
색상	193	107	300
품질	225	75	300
소음	229	71	300
음질	242	58	300
화질	248	52	300
합계	1,868	532	2,400

표 9 정확률

	5-Structure	8-Structure
가격	0.86	0.91
디자인	0.93	0.94
배송	0.87	0.88
색상	0.74	0.81
품질	0.85	0.87
소음	0.88	0.89
음질	0.90	0.91
화질	0.93	0.93
평균	0.87	0.89

표 10 재현률

	5-Structure	8-Structure
가격	0.66	0.95
디자인	0.80	0.94
배송	0.81	0.90
색상	0.76	0.89
품질	0.78	0.96
소음	0.87	0.94
음질	0.83	0.89
화질	0.81	0.92
평균	0.79	0.92

에 해당하지 않는 문장 중에서 감정어를 포함하는 문장의 수(FN)를 산출하였다. 산출한 각각의 값(TP, FP, FN)을 통해 재현률과 정확률을 계산하였다.

$$\text{재현률} = \frac{TP}{TP+FN}$$

$$\text{정확률} = \frac{TP}{TP+FP}$$

표 9와 표 10은 5-Structure와 8-Structure에 대한 정확률과 재현률을 나타낸다.

5.3 PMI-IR 식을 이용한 감정어 추출 실험 및 결과

PMI-IR[10]은 속성명 *f*와 *k*개 감정어(*OPWord*)와의 연관성을 계산한다. 계산한 점수가 상위 N 순위에 해당하는 감정어를 *f*에 대한 감정어로 취급하는 방법이다.

일반적인 PMI-IR 기법은 검색 엔진에 속성명과 감정

어를 질의한 후, 결과로 나온 문서 수를 이용하여 PMI-IR 수식을 계산한다. 이 때, 검색 엔진은 문서에 포함된 속성명과 감정어가 서로 인접해 있지 않더라도 문서 내에 두 질의어가 포함되어 있다면 결과 문서로 결정한다. 즉, 전통적인 PMI-IR 기법은 문서 단위로 속성의 감정어를 찾는 방법이다.

그러나 본 논문에서 제안하는 k-Structure 기법은 속성과 인접한 감정어를 찾는 방법이다. 따라서 k-Structure기법과 PMI-IR 기법을 비교하기 위해 동일한 데이터를 사용하였고, PMI-IR 기법을 본 논문에 적용시 속성명과 인접한 거리(±10)에 감정어(*OPWord*)가 있는 경우를 검색 결과로 결정하도록 했다.

다음은 본 논문에서 적용한 PMI-IR 식을 나타낸다. 수식은 일반적인 PMI-IR 기법과 동일하다. 다만, 본 논문에서 사용한 PMI-IR 기법에서는 웹 검색 엔진을 이용하지 않고 자체적으로 수집한 Product Review DB를 활용하였다.

$$PMI-IR(w,f)=\log\frac{hits(w,f)+\varepsilon}{hits(f)}$$

여기서 $hits(w,f)$는 그림 3의 Product Review DB에 저장되어 있는 상품평 중 속성별 300건을 대상으로 f 와 w를 동시에 질의하였을 때 검색된 문서(레코드)의 개수이다. $hits(f)$는 f 만 질의하였을 때 검색된 문서(레코드)의 개수이다. ε은 $hits(w,f)$의 결과가 0이 될 경우 전체 수식에 영향을 미치지 않도록 설정해주는 상수 값이다. 본 논문에서는 1로 설정하였다. 실험에서 사용한 속성명 f 는 4절에서 명시한 8개이다. 감정어는 표 5의 방법을 통해 추출한 것으로 사람이 찾아내어 맞춤법을 검사하고 기본형으로 변환하였다. 그렇게 추출한 감정어는 표 11과 같다.

표 11에서 속성명이 '공통'으로 되어 있는 것은 8개의 속성명에서 감정어로 사용 가능한 것이고, 특정한 속성명으로 명시되어 있는 것은 그 속성명에서만 사용가능한 것이다. 따라서 각 속성별 감정어는 공통으로 사용하는 감정어와 해당 속성에서만 사용하는 감정어를 포함한다.

표 11 속성별 감정어 리스트

감정어	속성명
추천하다	공통
적당하다	공통
나쁘다	공통
....	
저렴하다	가격
싸다	가격
...	
느리다	배송
빠르다	배송
...	...

표 12 속성별 감정어의 수

속성	공통	해당 속성	전체
가격	12개	5개	17개
배송		7개	19개
품질		1개	13개
음질		2개	14개
디자인		5개	17개
색상		8개	20개
화질		5개	17개
소음		2개	14개
전체	12개	35개	47개

표 12는 속성별 감정어의 수를 나타낸다. 여기서 '공통'으로 사용되는 12개의 감정어는 모든 속성에서 사용 가능한 감정어로 '추천하다','적당하다','나쁘다' 등의 단어로 이루어져 있다. 반면에 '해당 속성'에서 사용 가능한 감정어는 '가격'에서 '저렴하다','싸다' 등으로 이루어진 것이고, '배송'에서는 '빠르다','느리다'등으로 이루어진다.

표 11에서 정의한 감정어 *OWList*를 일반화 하면 다음과 같이 나타낼 수 있다.

$$OWList=\{ow_1,\ldots,ow_{12},ow_{13},\ldots,ow_{47}\}$$

w_1에서 w_{12}까지는 모든 속성에서 사용 가능한 공통 감정어이고 w_{13}에서 w_{47}까지는 8개 각각의 속성에서 사용할 수 있는 감정어이다. PMI-IR식에서는 각각의 속성명과 *OWList*에 있는 모든 감정어를 인자로 하여 점수를 계산하다. 그런 후에 점수가 상위 N 순위에 해당하는 것을 감정어라고 본다. 여기서 n은 각 속성별로 *OWList*에서 포함하고 있는 감정어개수이다. 즉, '가격'에서 n은 17이고, '배송'에서 n은 19이다.

표 13은 PMI-IR 기법을 통해 실험한 결과를 보여준다.

'색상'과 '소음'의 재현률과 정확률에 있어서 다른 속성들보다 상대적으로 높은 결과를 보이고 있다. 반면에 '품질'에서는 가장 저조한 성능을 보이고 있다.

표 13 실험 결과

	재현율	정확률
가격	0.50	0.53
디자인	0.53	0.59
배송	0.50	0.47
색상	0.68	0.65
품질	0.40	0.46
소음	0.65	0.61
음질	0.56	0.64
화질	0.56	0.59
평균	0.55	0.57

5.4 결과 분석

[13]에서는 [17]에서 사용한 영어 감정어 14개(긍정:7개, 부정:7개)를 네이버 영어 시소러스 사전을 이용하여 14개 단어에 대한 유의어를 찾았다. 그런 후에 영어로 된 유의어를 한글로 번역하여 감정 자질로 사용하였다. 하지만 [13]에서 사용된 감정어는 일반적인 상황에서 사용될 수 있는 감정어로 실제 쇼핑몰에서 자주 사용되지 않는 단어일 수 있다. 따라서 본 논문에서 추출한 47개의 감정어를 [13]에서 씨드 워드로 사용하여 확장한다면 감정어 추출의 정확도를 향상시킬 수 있다.

[14]는 구문분석기를 사용할 경우의 감정어 추출 방법이다. 아울러 출현 빈도가 낮은 감정어를 선정하지 못할 수 있는 문제점을 내포하고 있다. [14]에서 제안한 방법과 다르게 본 논문에서는 형태소분석기를 사용하여 감정어를 추출하는 방법을 제안하므로 구문분석기를 구하기 어렵고 형태소 분석기만을 사용할 수 있는 경우에 해당하는 감정어 자동 추출 방법이다. 그러면서도 [14]의 문제점일 수 있는 출현빈도가 낮은 감정어를 선정하지 못할 수 있는 문제도 해결한다.

k-Structure는 한국어 감정어 추출 관련 연구에서 제안한 방법에 상호보완적으로 추출의 정확성을 높이고자 제안하는 방법이다. 5-Structure와 8-Structure의 실험에서 두 방법 간의 정확률에서 크게 차이는 없었으나 재현률에서 많은 차이를 보였다. 이는 당연히 8-Structure에서 5-Structure보다 감정어가 포함된 문장을 더 많이 보유하고 있었다는 것을 의미한다.

본 논문에서 실험한 PMI-IR 기법의 경우 검색 엔진을 이용하지 않고, k-Structure에서 사용한 데이터를 이용하였다. 아울러, k-Structure 기법과 비교를 위해 속성 단위로 감정어를 찾도록 하였다. 이는 속성명과 인접한 거리에 감정어가 포함된 문서(레코드)를 결과 문서로 선정함으로써 가능했다. PMI-IR 기법을 적용한 결과와 비교하여, 5-Structure의 재현률은 24%, 정확률은 20% 높게 나왔다. 8-Structure의 재현율은 37%, 정확률은 32% 높았다. 이로써 본 논문에서 제안하는 k-Structure 기법이 효과적임을 보여주고 있다.

6. 결 론

본 논문에서는 k-Structure를 이용하여 한국어 상품평으로부터 자동으로 감정어를 추출하는 방법을 제안하였다. 영어권 연구에서 제시한 방법들을 한국어에 적용하려고 해도 분석하려는 언어의 구조가 다르기 때문에 직접적 적용이 쉽지 않다. 물론, 한국어권 연구 중에도 감정어 추출과 관련된 연구는 존재하지만 수작업 분석을 통해 감정어를 찾는 연구의 경우, 추출 시간이 오래 걸린다. 또한 자동적인 방법으로 추출한다 하더라도 정확도의 저하를 제고해야 하는 문제를 가지고 있다. 따라서 본 논문에서는 기존 연구에서 제안된 방법에 상호보완적으로 정확도를 향상 시킬 수 있는 방법을 제안하였다. 아울러, 성능의 비교를 위해 영어권 연구에서 제안한 PMI-IR 기법을 본 논문에 적용하여 실험하였다. 그 결과, 8-Structure가 가장 높은 성능을 보였다.

향후 연구에서는 상품평의 종류에 따라 k-Structure에서의 구조가 확장될 수 있으며, 이 밖에도 복잡한 문장 속에서 감정어를 추출하는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *In Proceedings of the Meeting of the Association for Computational Linguistics(ACL'02)*, pp.417-424 (2002).

[2] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.79-86, 2002.

[3] K.Dave, S. Lawrence, and D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *In Proceedings of the 12th Intl. World Wide Web Conference (WWW '03)*, pp. 512-528, 2003.

[4] Qiang Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destination by supervised machine learning approaches," Expert Systems with Applications, Elsevier, pp.1-9, 2008.

[5] Hanhoon Kang, Seong Joon Yoo, Dongil Han, "Accessing Positive and Negative Online Opinions," In *Proceedings of the 13th International Conference on Human-Computer Interaction, HCII 2009, LNCS 5616*, pp.359-368.

[6] Youngho Kim, Yuchul Jung, and Sung-Hyon Myaeng," An Opinion Analysis System Using Domain-Specific Lexical Knowledge," *In Proceedings of the 4th Asia Information Retrieval Symposium, AIRS 2008, LNCS 4993*, pp.466-471.

[7] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *In Proceedings of ACM SIGKDD Intl. Conf on Knowledge Discovery and Data Mining(KDD '04)*, pp.168-177, 2004.

[8] Soo-Min Kim, Eduard Hovy, "Determining the Sentiment of Opinons," *Proceedings of the COLING conference*, pp.1-8, 2004.

[9] Ana-Maria Popescu, Oren Etzioni, "Extracting Product Features and Opinions from Reviews," *Proceedings of the conference on Human Language Technology and Empirical Methods in*

Natural Language Processing, pp.339-346, 2005.
[10] Qi Su, Kun Xiang, Houfeng Wang, Bin Sun, Shiwen Yu, "Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews," *International Conference on the Computer Processing of Oriental Languages*, pp.22-30, 2006.
[11] Qingliang Miao, Qiudan Li, Ruwei Dai, "An integration strategy for mining product features and opinions," *Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 1369-1370, 2008.
[12] http://www.moransoft.com/sentidict.pdf,Technical Report
[13] Jaewon Hwang and Youngjoong Ko, "A Korean Sentence and Document Sentiment Classification System Using Sentiment Features," *Journal of Korean Institute of Information Scientists and Engineers (KIISE): Computing Practices and Letters*, vol.14, no.3, pp.336-340, May, 2008. (ISSN 1229-6848)
[14] J. Myung, D. Lee, 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.347-405, Jun. 2008. (in Korean)
[15] Morphemic Analyzer Tool (Korean Language Technology Ver. 2.10b), http://nlp.kookmin.ac.kr
[16] Hanhoon Kang, Seong Joon Yoo, Dongil Han, "Modeling Web Crawler Wrappers to Collect User Reviews on Shopping Mall with Various Hierarchical Tree Structure," In *Proceedings of The 2009 International Conference on Web Information Systems and Mining, IEEE Computer Society*, pp.69-73, 2009.
[17] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," *ACM*, pp.617-624, 2005.

강 한 훈

2006년 2월 세종대학교 컴퓨터소프트웨어학과 졸업(학사). 2008년 2월 세종대학교 컴퓨터공학과 졸업(석사). 2008년 3월~현재 세종대학교 컴퓨터공학과 박사과정. 관심분야는 데이터마이닝, 오피니언 마이닝, 기계학습, 웹 크롤러

유 성 준

1982년 2월 고려대학교 전자공학과 졸업(학사). 1990년 2월 고려대학교 전자공학과 졸업(석사). 1996년 2월 시라큐스대학교 전산학과 졸업(박사). 2002년 3월~현재 세종대학교 컴퓨터공학과 부교수. 관심분야는 러닝시스템, 패턴 인식, 데이터마이닝, 이미지 프로세싱

한 동 일

1988년 2월 고려대학교 전자전산공학과 졸업(학사). 1990년 2월 한국과학기술원 전기 및 전자공학과 졸업(석사). 1995년 2월 한국과학기술원 전기 및 전자공학과 졸업(박사). 1995년 2월~2003년 2월 LG 전자 디지털TV연구소 책임연구원. 2003년 3월~현재 세종대학교 컴퓨터공학과 교수. 관심분야는 영상 처리, 신호 처리, 컴퓨터 비전, 데이터마이닝