인지과학, 제21권 제4호
Korean Journal of Cognitive Science
2010, Vol. 21, No. 4, 669~696.

감정어휘 평가사전과 의미마디 연산을 이용한 영화평 등급화 시스템*

고 민 수 신 효 필†
서울대학교 언어학과

본 논문은 한 문서의 전체 의미는 각 부분의미의 합성이라는 관점에서 미리 반자동으로 구축된 감정어휘 평가사전을 기반으로 한 시스템을 제안한다. 인간의 의사 결정 과정과 유사한 방식으로 의사 결정 과정을 모델링하려는 노력으로써 본 ARSSA 시스템은 개별 리뷰의 의미값 연산과 자료 분류를 통해 감정 표현이 나타난 영화평 리뷰의 자동 등급화에 대한 연구를 수행한다. 이는 {'평점' : '리뷰'} 이항구조로 이루어진 현재의 평점 부여 형식에서 발생하는 두 변항의 불연속성 문제를 해결해보려는 목적을 가진다. 이는 어휘 의미 합성 과정에서 반영된 추상적 의미들의 합성 함수를 통해 실현될 수 있다. 시스템의 성능 실험에서 네이버 무비에서 확보한 1000개의 리뷰에 대한 10-fold 교차 검증 실험이 수행되었다. 이 실험은 기존에 부여된 평점과 비교하여 감정어휘 평가사전을 이용하였을 때 85%의 F1 Score를 보였다.

주제어 : 감정어휘 평가사전, 유의어, 유의어집합, 등급화, 의미마디, SVM, ARSSA

---

* 이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2008-327-A00325)

† 교신저자: 신효필, 서울대학교 인문대학 언어학과, 연구 분야: 자연언어처리
E-mail: hpshin@snu.ac.kr

서 론

이 연구는 자연언어처리의 관점에서 관련 대상에 대한 사용자의 평가를 일관된 기준으로 일반화시킴으로써 대상의 가치에 대한 정확한 판단과 관리의 편의성을 향상시키는 자동 등급화 시스템을 제안한다. 본 연구의 목표는 다음 연구 동기에서 발생한 과제를 해결할 수 있는 자동화 시스템을 개발하는 것이다. ① 평가/감정 표현 관련 어휘 의미에 대해 개인과 한 언어 사회에서 공유하는 기준의 교집합을 찾아 일반화시킬 수 있는가. 만약 가능하다면 어떻게 검증할 수 있는가. ② 유의어 간 등급 설정이 가능한가. 가능하다면 어떤 기준이 있는가. ③ 일반화된 기준을 이용하여 영화평을 자동 분류했을 때, 기존에 부여된 평점과의 비교 결과가 얼마나 일치하는가. ④ 결과의 차이를 오류라고 했을 때, 보정 결과의 타당성을 확보할 수 있는가.

현재 리뷰 작성 양식은 일반적으로 사용자에게 '내용'과 '평점'을 각각 요구한다. 만약 사용자가 따르는 일관된 기준 체계가 없다면, 해당 리뷰에 대한 신뢰성이 의심받을 수 있다. 만약 대다수의 사용자가 공유하는 기준을 확보할 수 있다면 입력 리뷰를 자동 분석하여 등급에 따른 평점을 부여하는 시스템의 구축이 가능할 것이다. 리뷰 텍스트를 구성하는 어휘 중 극성/평점의 결정에 직접 관계된 요소는 감정/평가를 표현해주는 어휘, 즉 '감정어휘'이다. 이 어휘들의 의미값은 내용어로써 리뷰 의미를 합성적으로 결정하는데 절대적 영향력이 있다. 감정어휘로 이루어진 '감정어휘 평가사전'은 리뷰 자동 등급화를 위한 시스템의 의미 사전으로 이용되었다.

인터넷 리뷰 문서는 형식적 자유로움이 허용되는 정도가 매우 높다. 인터넷 화자의 의사표현은 은어, 속어, 유행어, 독립어, 철자변이형, 문법파괴형, 이모티콘 등의 요소를 통해 풍부한 감정을 드러내고 있다. 전처리 과정에서 이러한 특징을 반영하기 위한 작업이 수행되었다. 감정어휘 평가사전은 어휘 간의 유의어 관계에 대한 연구를 바탕으로 구축되었다. 이는 어휘에 대한 의미값을 부여의 기준 설정을 위한 것으로, 어휘 간 체계적 관계를 형성하고 등급화가능성을 확보함으로써 방법론적으로 일관성있는 어휘의미값의 할당을 지향한다.

최종 의미값은 통사구조, 접속어미, 부정어, 중의성 등을 고려하여 의미마디 단

위 및 문장단위 연산과정을 거치면서 해당 리뷰에 대한 의미합성값으로 도출된다. 이 의미합성값은 등급화된 각 어휘를 기초로 연산된 값이기 때문에 연산과정 이후에도 결합 패턴에 따라 문장별로 일정한 등급 패턴이 발생한다. 연구의 타당성은 Naver Movie 영화평 말뭉치를 이용한 시스템 성능 실험에서 검증된다.

관련 연구

본 연구는 고민수(2010)의 연구를 기초로 크게 두 가지의 과제가 추가적으로 수행되었다. 첫째, 모든 개별 어휘의미값을 일관된 기준에 따라 자동 할당받는 시스템을 갖추었다. 따라서 개념적으로 존재하던 연속적 실수형 자료로 구성된 인간의 어휘의미에 대한 정도를 나타내는 수치값을 이용할 수 있게 되었다. 둘째, 어휘의미의 도메인 독립성과 감정어휘의 일반성을 확보하기 위한 시도로써 말뭉치를 추가 · 확장시키고 이에 대한 실험이 수행되었다.

'의견 정보'와 '사실 정보'의 구분

문서의 의견 추출 방법에 관한 연구들을 간략하게 살펴보자면, 주관성 여부의 판단을 위해서 Wiebe(2000)는 주관적 언어(subjective language)를 '저빈도어, 연어 관계, 형용사/동사'의 세 가지 주관성 단서로 표현 수준과 문서 수준에서 수작업으로 주석 처리된 말뭉치를 학습시켜서 93%의 높은 정확도를 얻을 수 있다는 것을 보였다. Hatzivassiloglou(2003)는 복합 질문에 대한 질의 응답의 문맥에서 정보 구조화를 위한 필요성으로 의견 검출 및 분류 시스템을 구축했다. 그는 사실로부터 의견과 극성을 분류하고 그에 따라 문서 수준에서 극성을 분류하기 위해서 Naive Bayes 기계학습 알고리즘을 통해 97%의 정확도를 보였다. 강인호(2007)는 문장 전체를 이용한 분류 결과와 감정 표현구를 이용한 분류 결과를 결합하여 주/객관적 문장 분류기의 성능 향상 방법을 제안한다. 이와 같은 접근 방식은 같은 범주에서 공통의 특징을 기준으로 의견이 표현되는 문서의 처리에 적합한 특징 기반 요약의 일종이다. 특징 기반 요약은 문장 수준에서 수행되지만 특징을 식별하는 단계를 통

해 특징을 포함하는 문장만을 대상으로 한다.

의견 극성 분류

의견 극성 분류는 가장 주된 감정 분석의 연구 과제이다. 기존 연구들에서 문장이나 용어의 극성을 판별하는 방법으로 여러 가지가 제안되었다. 제안된 방법에 따라 이를 몇 가지 유형으로 분류하면 다음과 같다.

첫째, 의미 사전에 기반한 방법을 이용하는 연구 유형이다. 본 연구의 경우 유의어 관계가 반영된 감정어휘 평가사전을 구축해서 극성을 판별하고, 명재석(2008) 역시 의미사전을 구축한다는 점에서 본 연구와 비교해볼 수 있다. 본 연구의 경우 Appraisal Theory(White 2005)를 통해 말뭉치로부터 검출된 각 어휘의 유의어 관계에 따라 사전을 구축한 것에 비해, 명재석(2008)에서는 사전을 White(2005)에 근거하여 작성된 Semantic Clause의 형태로 반자동으로 구성한다. 한편 Whitelaw(2005)는 White (2005)에 근거하여 Appraisal Taxonomies를 구축하여 극성을 분석했다.

둘째, 기존의 유의어 사전을 이용해서 의미를 바탕으로 어휘를 확장하여 자질로 이용하는 방법의 연구 유형이다. 의미값 기반 감정어휘 평가사전의 구축 방법은 이 유형에 연관성이 있다. 고영중(2008)은 감정 분류는 문서에 나타나는 단어 형태가 아닌 의미에 기반해야 한다는 점에 입각해서 유의어 관계에 주목했다. 유의어 추출을 위해 영단어 시소러스의 유의어 정보를 이용해 어휘를 확장하고, 대역사전을 이용해 번역하는 방법을 취했다. 이 방법은 유의어 사전을 거쳐서 생성한 유의어 목록을 확보한다는 점에서 해당 사전의 영향을 받게 된다. 영단어 시소러스를 이용하는 방법은 외국어로 기술된 사전의 의미와 단어에 원천적으로 영향을 받고, 대역사전의 오류 및 중의성 문제에 노출될 가능성이 있다.

셋째, 기계 학습을 이용해서 극성을 분류하는 방법의 연구 유형이다. 이는 자동 문서 분류에 가장 일반적으로 이용된다. Pang(2002), Dave(2003)에서와 같이 학습 자료의 평점을 기계 학습 시킨 후 극성을 예측할 수 있다는 것이 특징이다. 이는 정답으로 가정한 자료에서 극성 판별 기준이 되는 평점을 신뢰할 수 없다는 문제점이 있다.

넷째, WordNet과 같은 언어 자원을 활용하여 극성을 분류하는 방법의 연구 유형이다. 이 방법은 WordNet의 유의어, 반어의 관계를 이용하여 형용사의 극성을 예

측하는 것이다. Liu(2004)는 유의어 관계의 형용사는 중심이 되는 형용사와 같은 극성을 가질 것이라고 예상했다. Esuli(2006, 2010)는 WordNet Synset과 WordNet Gloss Corpus의 어휘자원을 기반으로 극성 어휘의 정규화된 의미값을 부여하고 중의성 해소를 위해 노력했다. 유의어, 반의어 관계를 이용한다는 점은 감정어휘 평가사전의 주요 구성을 이루고 있는 유의어집합과 연관성이 있다. 이 방법의 문제점은 감정어휘가 문맥적으로 다른 용법을 갖더라도 변경하기 어렵고 기존 언어 자원에 의존하고 있고, WordNet과 같이 검증된 한국어 자원이 없기 때문에 즉시 적용할 수는 없는 방법이라는 점이다. 한국어에서는 본격적인 감정어휘 사전이 없고 현재 이를 구축하려는 노력이 시작되고 있다. 따라서 영화평 어휘를 중심으로 얻어진 어휘를 일반어휘로 확대하여 감정 어휘 사전을 구축한다는 점에서 본 연구의 의의가 있다.[1)]

시스템 구조

이번 단락에서 연구의 전체 논의 과정을 바탕으로 제안하는 리뷰 자동 등급화 시스템 ARSSA의 시스템 구조에 대해 설명한다.

[과정 1] 형태소 분석을 포함한 전처리 과정이다. 우선 입력 리뷰는 띄어쓰기

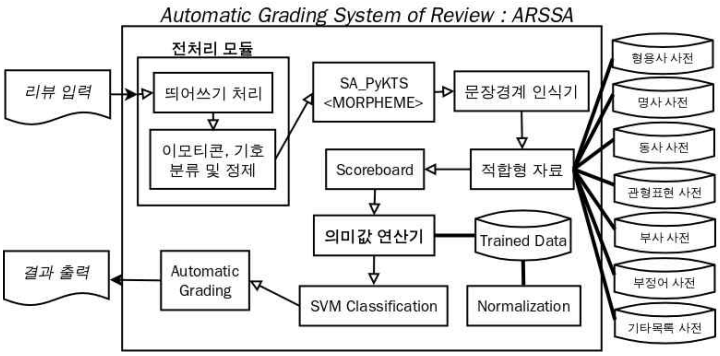


그림 1. 리뷰 자동 등급화 시스템 ARSSA

1) 현재 감정 온톨로지에 관한 연구가 윤애선(2010) 등에서 시도되고 있다.

처리와 이모티콘/기호 등의 정제 과정을 거친다. 이후 형태소분석기를 이용하여 텍스트에 형태소 표지를 부착한다. 이후 각 어절에 부착된 형태소 표지를 근거로 문장경계 인식기가 리뷰의 문장경계를 자동 구분한다.

{과정 2} 앞 단계에서 적합형 자료를 확보하면 감정어휘 평가사전과 어휘유형이 매칭되는 모든 감정어휘에 대한 의미값이 순서대로 연산을 위해서 일괄적으로 Scoreboard에 기록된다. 이 값을 이용해서 의미값 연산기는 W_n으로 구성된 문장 내에서 표 3의 구조화된 몇 가지 의미 결합 패턴을 인식하는 '$RULE_x$'에 따라 구성된 의미마디 mc_n는해당 연산 Cal_mc에 투입된다. 각 문장단위 의미값 ms_n은 문장단위 연산 Cal_ms에서 입력 리뷰의 최종 결과 의미값 $Value_R_n$을 얻는다. 이후 단계에서 자동 등급화된다.

{과정 3} 이는 자동 등급화를 위한 분류 단계이다. 고성능의 다차원 분류를 위해 SVM(Support Vector Machine)을 시스템의 최종 분류기로 이용한다. 100개의 기존 훈련 자료를 바탕으로 새로 입력된 리뷰의 의미값은 SVM 분류기를 거치고 9개의 결정 초평면이 나타내는 10개의 분할 공간 중 어느 곳에 속하는가에 따라 자동 등급화가 이루어진다. 분산되어있는 자료의 분포 상태를 보정함으로써 각 분할 공간의 영역과 결정 초평면을 최적화시키기 위해서 기계 학습 이전에 학습 집합의 전체 의미합성값은 훈련 자료에 대해 Normalization(Bolstad 2003)을 거친다. 이를 통해 자료는 각 등급에 대해 보다 견고한 특징 공간을 갖도록 유도되었다.

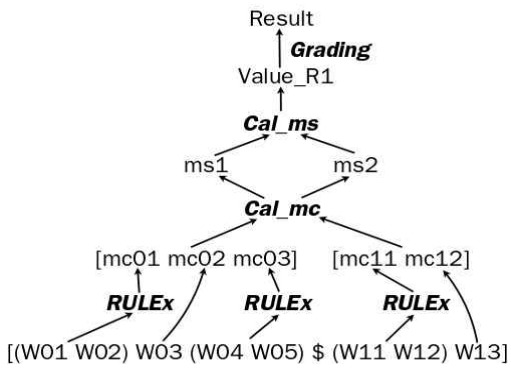


그림 2. 의미값 연산 과정

전처리와 감정어휘 평가사전

전처리 과정

인터넷 문서의 특징상 발생하는 문제는 입력 데이터의 상당량을 분석 이전에 소실시킨다. 이는 연구에 큰 장애가 되기 때문에 작업의 필수 선결과제이다. 형태소 분석을 위해서 오픈소스로 공개된 KTS 형태소분석기를 보완한 감정분석용 형태소 분석기 SA_PyKTS[2]를 이용했다.

띄어쓰기가 잘 지켜지지 않는 현상은 구어체 말뭉치에서 관찰되는 특징이다. 띄어쓰기 부분 오류는 형태소분석기에서 형태소 별로 분할 처리되지만 처리 가능 비율을 현저히 감소시키는 문장 전체의 띄어쓰기가 무시된 경우는 우선 띄어쓰기 교정기를 통해 자료를 처리한다. 이후 입력 자료는 이모티콘 분류 및 기호 정제를 거친다. 이모티콘 분류는 형태소 분석 오류를 줄이고 다양한 이모티콘의 사용이 감정표현 텍스트에서 갖는 의미를 보존하기 위해 자체적으로 구현된 이모티콘 분류기를 이용했다. 일관된 패턴에 따라 분류되기 때문에 이후 연구에서 관련 정보가 필요할 때, 처리하기에 적합한 형태로 남는다.

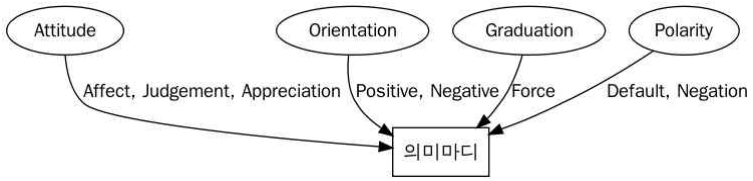
한편 어휘유형별 분류를 통해 빈번한 철자오류 유형을 파악해서 사용가능 형태로 반영하고, 형태소분석기 개발 당시 사전에 미반영된 어휘들이나 해당 도메인에서 많이 쓰이는 어휘를 찾아내어 처리한다. 전체 과정은 형태소분석기 의존적이기 때문에, 빈번한 철자오류나 미반영 어휘 목록을 확보할 것이 요구된다. ‘Cine21 영화평말뭉치’를 띄어쓰기 단위로 청킹 했을 때, 총 953,935개의 청크와 227,769개의 어휘유형으로 분할된다. 고빈도 순 정렬시, 어휘유형별 빈도 5회 이하의 저빈도 어휘유형들에서 의도적인 맞춤법 파괴현상이 관찰되는데, 이러한 형태 어휘의 사용자는 핵심 어휘뿐만 아니라, 조사, 어미등을 활용하여 다양하게 변형시켜 사용한다. 예를 들어, ‘재미있-’의 이형태 ‘잼있-’의 경우, 가장 먼저 발견되는 ‘잼있-’ 포함 어휘유형은 ‘잼있게’(순위: 1280, 빈도: 82회)이지만, 총 222개의 유형중 212개의 유형이 빈도 5 이하의 어미/조사 변화형을 보여준다. 이상의 전처리 과정을 거침으

2) http://clab.snu.ac.kr/sakts/

로써 형태소 오분석을 감소시키고 시스템이 요구하는 수준의 정제된 텍스트를 얻을 수 있다.

의미마디와 유의어 관계

감정어휘 평가사전은 특정 기준에 따라 감정표현으로 분류되는 모든 어휘의미값이 저장된 의미사전이다. 특정 기준이란 Appraisal Group(Whitelaw 2005)의 개념을 응용한 의미마디이다. Whitelaw(2005)에서 사용된 형용사 감정어휘의 분류 방식은 Appraisal Theory(White 2005)를 따르고 본 연구에도 적용되었다. 의미마디에서 각 어휘가 Attitude 하위 유형에 속하는지 여부에 따라 감정어휘가 추출되고, 선정된 표제어는 연산 과정에서 그림 3의 각 항목에 대한 값이 부여된다.



	Attitude	Orientation	Graduation	Polarity
좋/pa	affect	positive	0	default
안/a 좋/pa	affect	negative	0	negation
아주/a 안/a 좋/pa	affect	negative	1	negation

그림 3. 의미마디의 구조와 분석

감정어휘 평가사전 1.0

감정어휘 평가사전 1.0은 고민수(2010)에서 검증된 사전의 도메인 특화성을 기초로 일반 감정어휘가 추가된 의미사전이다. 기존에 영화평 말뭉치를 기반으로 구축된 감정어휘 평가사전은 사전의 일반성을 확보하기위해 현재 기존 목록에 세종전자사전 체언 사전 속성인간 하위 의미부류 전체 및 관련 부류 36항목과 KOLON[3)]

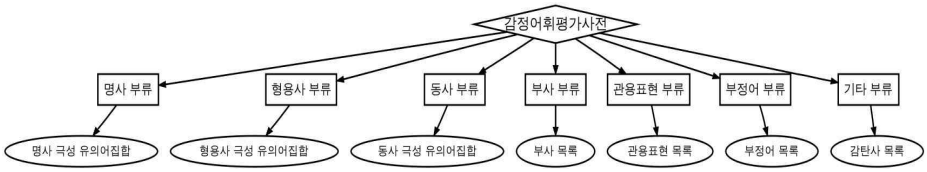


그림 4. 감정어휘 평가사전의 주요 구성

의 POSITIVE/ NEGATIVE-STATE에 속하는 해당 극성 어휘가 추가되었다. '감정어휘 평가사전 1.0'의 목록은 홈페이지[4]에 공개되어 있다. 이는 그림 4와 같이 총 7 부류로 구성된다.

감정어휘란 '문자의 형태로 인간의 감정을 표현하기 위해 사용되고, 등급화가능성을 갖는다고 판단되는 모든 어휘'라고 정의한다. 고빈도 감정어휘일수록 해당 말뭉치에서 기본 감정표현과 관련된 기능을 한다. 감정표현을 통해 전달하고자 하는 모든 정보는 감정이 포괄하는 몇 가지 하위 영역에 속한다. 만약 고빈도 어휘 목록을 확보했다면 이는 전체 말뭉치에서 자주 표현되는 감정과 관련 어휘일 가능성이 크다고 가정하고 나머지 저빈도 어휘들이 감정어휘에 속하는지 여부는 유의어 대사전(최운천 2010)의 검색 결과를 참조해서 작업자의 직관에 따라 고빈도 어휘와 유의어 관계에 있는지 여부를 판별하여 기존 목록에 포함/탈락 여부를 결정하고 유의어 목록을 추가했다. 이는 Esuli(2010)의 *bag of synsets* 모델과 유사한 방법이다.

저빈도 어휘의 발생은 크게 다음 두 가지 경우에 해당된다. 첫째, 일반적으로 자주 사용되지 않는 표현을 사용한 텍스트에서 발생된다.[5] 그러나 여기에 해당하는 어휘들 역시 결국 어떤 감정의 한 부분을 표현하기 위해 쓰였기 때문에, 고빈도 감정표현 어휘가 나타내고자하는 의미에 속할 가능성이 크다.[6] 만약 어떤 감정표

3) http://word.snu.ac.kr/kolon/

4) http://clab.snu.ac.kr/arssa/

5) '일반적으로 자주 사용되지 않는 표현'의 판단 기준은 다음과 같다. 예를 들어, 유의어집합 Sn(힘들다)에 속한 어휘 '굴곡지/pa'의 경우와 같이 사전적 의미와 관계없이 해당 말뭉치에서 사용 빈도가 1에 근접하는 표현들을 의미한다.

6) 장기간 다수의 사용자가 빈번히 사용한 고빈도 감정어휘는 인간이 공통적으로 느끼는

표 1. 감정어휘 평가사전 1.0 주요 품사별 통계 (현재 2010. 10. 12.)

말뭉치 seed			긍정	부정	총합
형용사	195	유의어집합	50	56	106
		어휘	552	625	1177
명사	278	유의어집합	497	223	275
		어휘	1411	1764	3175
동사	54	유의어집합	13	20	33
		어휘	25	80	105

현에 속한다고 판단되는 경우 해당 유의어집합에 '포함' 과정을 거친다. 둘째, 형태적 변이형의 발생이다. 기본형에서 벗어난 변이형에 직접적으로 관계된 기본형이 이미 목록화[7]된 경우, 기존 목록의 변이형으로 본다.

유의어 관계와 유의어집합

감정어휘 평가사전은 말뭉치에 출현하는 감정어휘 모두를 포착할 수 있어야 한다. 따라서 전체 어휘에 대한 일관된 기준에 따른 값 부여가 중요한데, 이를 위해 전체 표제어에 대한 긍정/부정 유의어집합을 구성하였으며 목록의 확보를 위해 유의어 대사전을 참고했다. 만약 둘 또는 그 이상의 단어가 '같은 의미'를 갖는다면 서로 유의어 관계이다. 같은 의미를 가진다는 것은 같은 원형의미에 속하는 진리값을 갖는다고 정의된다. 본 연구의 유의어집합은 아래의 세 가지 성립조건을 따른다.

한 집합 내부의 동일 등급의 원소는 집합별로 각각 값할당 기준이 동일한 방법

기본 감정에 속할 것이라는 직관적 가정은 고민수(2010). p.22에서 제시한 순수 감정 어휘 범위에 해당 어휘가 속하는 유의어집합이 할당되는 것으로 증명된다. 감정어휘 중 저빈도 어휘가 독자적으로 유의어집합을 설정하는 경우는 거의 없기 때문에 이와 같은 설명이 가능하다.

7) 형태적 변이형이 감정어휘로 포착된 경우 같은 의미의 기본형이 존재한다. 그 기본형이 감정어휘 평가사전에 이미 포함되어 있는 경우를 목록화라고 설명한다.

상에 있고 서로 다른 차원에서 등급화되었다. 기본적으로 유의어집합 내부 원소의 할당 값은 등급별로 실수 형태의 연속적 스케일 상의 어떤 한 점을 의미한다. 유의어집합은 기본적으로 각 품사별로 극성 유의어집합이 생성된다. 유의어집합 간 등급화작업은 다음 정의에 따라 유의어집합 내부 어휘들 간에 이루어졌다.

<유의어집합 성립조건 1>

$Synset_i = \{m_j | m_j \in S_i,\ S_i \cap S_{i\pm1} = \varnothing\}$

(단, i, j $\geq$ 1, S_i는 유의어집합, m_j는 어휘의미)

<유의어집합 성립조건 2>

1. 결합법칙이 성립한다 : S_i의 모든 원소 m_i, m_{i+1}, m_{i+2}에 대해 $(m_i * m_{i+1}) * m_{i+2} = m_i * (m_{i+1} * m_{i+2})$ 가 성립한다.
2. 항등원이 존재한다 : S_i의 모든 원소 m_i에 대해, m_i * e = e * m_i = m_i 인 S_i의 원소 e 가 존재한다.

<유의어집합 성립조건 3>

1. × 에 대해 닫혀있다. : 임의의 원소 a, b에 대해 a × b ∈ G 이다.
2. × 에 대해 결합법칙이 성립한다.
3. 항등원이 존재한다.
4. 임의의 원소는 역원이 존재한다.

$$S_i = \{\{m_{1(g1\ min)}\cdots\}, \{m_{2(g2)}\cdots\}, \{m_{3(g3)}\cdots\}, \{m_{4(g4)}\cdots\}, \{m_{5(g5\ max)}\cdots\}\}$$

[정의 1] 서로 다른 유의어집합에 속했다면, 서로 교집합이 공집합인 독립적 집합들 간의 관계라고 가정했을 때, 다른 차원에 속하는 같은 순번의 등급은 서로 무관하다. 서로 다른 의미의 유의어집합 *S1*~*S4* 총 4개가 있다면, 각 집합 내부의 어휘들은 각각 가장 강한 표현부터 가장 약한 표현의 5등급으로 분할된다. 만약 *S1* 집합의 3등급 어휘와 *S4* 집합의 3등급 어휘가 같은 등급에 속했다고 해서, 같은 의미값을 갖는다고 할 수 없다.

[정의 2] 이 집합은 열린 집합이다. 새로운 의미 원소 'm_j'는 현존하는 집합에 추가될 수 있다. 어휘의미는 시간의 흐름에 따라 소멸, 변형, 생성을 반복하기 때문에, 개별 집합의 내부 원소는 '추가/삭제' 가능성을 가진다.

[정의 3] 만약 현재 특정 의미값이 존재하지 않는다면, 이는 해당 어휘의 사멸, 변형, 미생성을 의미한다. 가상의 'm_j'를 설정함으로써 기본 등급 사이에는 수많은 원소의 의미값이 위치할 가능성이 있다. 최대값과 최소값 사이의 무한히 많은 어휘의미가 실수값 형태로 저장될 수 있다.

유의어집합 등급화와 어휘의미값 할당

유의어집합의 내부 원소들은 TFIDF 통계값을 기준으로 기본 등급이 분할된다. 의미란 문맥에 따라 제한되어 있고 의미 관계는 어휘 자체의 의미보다 우선한다는 점에서 기본 의미와 실제 적용의 구분은 타당하다. 어휘 등급화를 위해 리뷰에 사용된 어휘들의 의미는 개념적 의미와는 차이가 존재하며 사용자의 어휘 의미 판단과정이 말뭉치에 반영되었다고 가정하고, 각 어휘의 TFIDF 값이 곧 어휘간 의미값의 유의미한 차이를 보여줄 것이라는 가정을 뒷받침 하는 통계값이라는 점에 주목한다. 각 해당 어휘에 대한 언어 직관이 반영된 말뭉치의 통계값은 불특정 다수의 사람들이 사용하는 의미에 대한 내적 판단에 따른 결과의 외연에 대한 관찰값이다.

Esuli(2006)의 SENTIWORDNET은 WordNet Synset 개별 어휘에 TFIDF 값을 기준으로 정규화된 확률 분포에 따라 값을 배정했다. 이와 같은 방법은 본 연구에서 유의어집합의 의미 조정값 할당 과정에 포함된다.

이 과정을 설명하자면, 먼저 TFIDF 검출기는 극성에 따라 각 유의어집합을 한 라인으로 인식하고 각 평점에 대한 값 중 가장 유의미한 값을 집합의 대표값으로 선택하고 기존의 점수를 위계로 삼아 원소를 정렬한다. 같은 위계에 속한 개별 어휘 각각은 해당 점수에 대한 비중에 따라 내부 위계가 정렬된다. 그 다음 각 평점에 속하는 전체 어휘에 대한 TFIDF 값의 표준정규분포 Z-transformation에 따라 [0.0, 1.0] 범위로 정규화된 확률값을 개별 어휘의 기본 등급값에 자동 부여한다. 이에 따라 어휘의미값이 출력된다. 즉, 어휘의미값은 $Value_{meaning}$은 기본 등급값 $Value_{grade}$에

확률밀도함수값 $Value_{PDF}$를 합한 값이다.

표 2. 유의어집합 어휘의미값 할당 과정 예시

S_1(헛수고하다) = {'맥빠지/pa', '헛되/pa', '부질없/pa', '허무하/pa', '허황/nca', '덧없/pa', '허무하/pa', '무상/nca', '뭣하/pa', '희망없/pa', '터무니없/pa', '새되/pa'}				
단어	**대표평점**	**TFIDF 값**	**Z-transformation**	**어휘의미값**
부질없/pa	9	1.012888	0.50061450203	-1.34575
덧없/pa	8	0.675387	-0.0231734195595	-1.748
무상/nca	8	1.125645	1.97682658044	-1.98805
허무하/pa	7	0.47657	-1.0	-2.08055
허황/nca	6	0.469679	-1.50473393973	-2.53405
맥빠지/pa	6	0.704518	-0.495266060269	-2.65605
터무니없/pa	6	1.878715	2.08426053576	-2.9906
뭣하/pa	5	0.604932	-1.33233006835	-3.0467
희망없/pa	5	1.209865	0.667669931645	-3.3743
새되/pa	4	1.446055	1.0	-3.92065

$$Value_{meaning} = Value_{grade} + Value_{PDF}$$

$$PDF_{Probability\ Density\ Function}:$$
$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

유의어집합을 갖는 명사, 형용사, 동사, 관용표현 부류의 최종 어휘의미값은 [-5.5, +5.5] 구간에 분포한다. 또한 유의어집합 내부 원소의 값이 기본 등급의 이산적 특성을 넘어 연속적 분포를 갖게 된다. 극성 어휘가 아닌 객관적 어휘로 분류된 경우는 이미 목록에서 제외되었기 때문에 0.0 값에는 어휘가 할당될 수 없다.

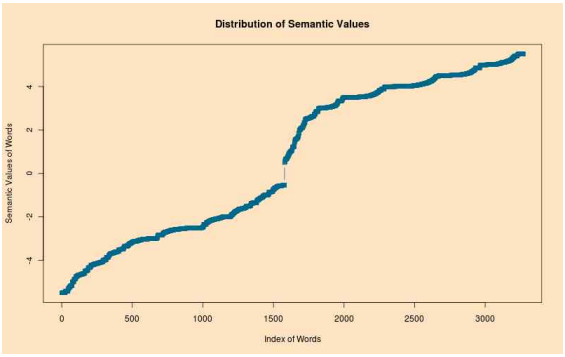


그림 5. 최종 어휘의미값 분포 (명사+형용사 부류)

의미값 연산과 다차원 분류

통사구조와 접속어미 및 접속부사 처리

연산의 기본 단위는 문장을 경계로 감정어휘 평가사전에서 추출된 감정어휘로 이루어진 의미마디이다. '의미마디'이란 하나 또는 몇 개의 어절이 독립적 의미를 형성하는 최소 단위를 뜻한다. 접속어미/접속부사가 포함된 표 3의 문장은 연산과정에 특징이 있다. 이들은 접속어미를 중심으로 문장 내에서 의미값 연산에 분절이 생기는 병렬문장이다. 본 연구는 기본적으로 관련 내용어 중심의 연산과정이 핵심이기 때문에 접속어미/ 접속부사에 관련된 유의미한 통사 패턴에 따라 연산이 이루어진다.

접속어미/접속부사의 기능은 통사적으로 결합된 두 문장 사이의 의미관계를 제시하는 것이다. 이들은 문장 내부에서 의미값 흐름 변경에 고유한 기능을 갖는다. 특히 대립관계 접속어미/접속부사 *"지만/cs, 나/cs, 아도/cs, 어도/cs, 나마/cs, ㄴ데/cs 는데/cs 그러나/ajs, 그래도/ajs, 근데/ajs, 그런데/ajs, 하지만/ajs"*의 목록이 연산에 반영되었다. 이는 그림 7에서 평점 6을 중심으로 분포한다는 점에서 알 수 있듯이 부정/긍정적인 평가가 혼재하는 병렬문장의 정확한 통사적 분석을 위해 필수적이다. 이를 중심으로 문장내 $\pm n$ 스팬 형태소 분석 단위 연쇄는 그림 6과 같이 각 주제

표 3. 병렬문장의 예

	스토리는 너무 짜증났었는데 연기는 일품이었다.
(1)	[S0 [S1 [NP 스토리/nc 는/jx] [VP [A **너무/a**] [V **짜증나/pv** 았었/efp] [C **는데/ecs**]] [S2 [NP 연기/nc 는/jx] [VP **일품/nc** 이/jcp 었/efp 다/ef]]]
	스토리는 짜증났었는데 연기는 너무 일품이었다.
(2)	[S0 [S1 [NP 스토리/nc 는/jx] [VP [V **짜증나/pv** 았었/efp] [C **는데/ecs**]] [S2 [NP 연기/nc 는/jx] [VP [A **너무/a**] [VP **일품/nc** 이/jcp 었/efp 다/ef]]]
	스토리가 너무 짜증났었고 연기는 일품이었다.
(3)	[S0 [S1 [NP 스토리/nc 가/jc] [VP **너무/a** [V **짜증나/pv** 았었/efp] [C 고/ecq]] [S2 [NP 연기/nc 는/jx] [VP **일품/nc** 이/jcp 었/efp 다/ef]]]
	스토리가 짜증났었고 연기는 너무 일품이었다.
(4)	[S0 [S1 [NP 스토리/nc 가/jc] [VP **짜증나/pv** 았었/efp] [C 고/ecq]] [S2 [NP 연기/nc 는/jx] [VP [A **너무/a**] [**VP 일품/nc** 이/jcp 었/efp 다/ef]]]

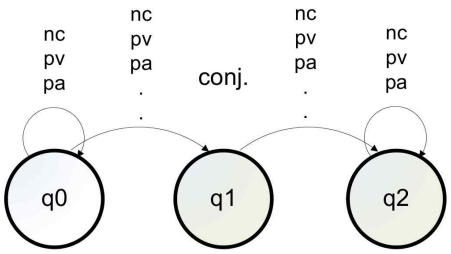


그림 6. 접속어미/접속부사 중심 통사 패턴[8)]

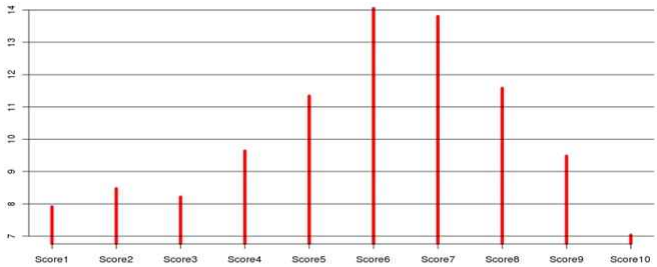


그림 7. 대립관계 접속어미 출현 전체 백분율

표현에 해당하는 품사 연쇄가 관찰된다.

부정어 처리

[8]

‘부정’은 소속된 명제의 진리값을 반대로 변화시킨다. 본 연구에서 리뷰의 진리값은 감정표현에 관련되어 있기 때문에, 직접적으로 극성과 관계되고 따라서 각 부정어는 연산시에 극성을 변화시키는 기능을 갖는다.

한국어 부정어는 일반적으로 *‘못’*은 형용사를 수식할 수 없고, 동사 앞에 오면 능력부정으로 기준미달을 나타내고, *‘안’*은 동사와 함께 나타나면 행위부정, 형용사와 함께 오면 성질 및 상태부정의 의미로 정의된다. 그 형태소는 *‘안/a’, ‘않/px’, ‘아니/a’, ‘못/a’, ‘못하/px’, ‘말/px’*의 형태로 출현하는데, 관계되는 의미군과 결합되어 방향성 결정과 의미값 연산에서 중요한 역할을 하기 때문에 부정어의 적용 범위를 파악해야 한다.

부정어의 영향력은 최대한 문장경계를 넘지 않기 때문에 이를 최대 한계로 설정한다. 내용어 중심 처리 과정에서 형용사가 부정어에 연쇄되는 경우 부정어는 바로 뒤에 오는 형용사를 수식하고, 부정어 이후에 동사가 출현하는 경우 전체 의미값을 부정하도록 규칙화되었다.[9]

---

8) 기본적인 품사 연쇄는 표 3과 같이 전체 품사 연쇄로 이루어져 있지만 실제 연산과정은 감정어휘로써 추출된 내용어를 중심으로 이루어지기 때문에 이와 같은 통사 패턴을 보이게 된다.

9) 부분부정의 문제에 있어서, 텍스트 처리는 실제 입력된 문자열에 의존하기 때문에 ‘다 안갔다.’와 같은 문장은 처리할 수 없다. 현재 ‘다 간 것은 아니다.’와 같은 명시적 형태만 처리 가능하다.

어휘의미값 연산

의미마디 추출과 단위 연산

해당 어휘에 대한 사용자의 언어 직관에 근거하여 수치화된 어휘의미값을 확보함으로써 감정표현과 관계된 다양한 언어 표현의 연산을 위한 기초를 마련했다. 의미값 연산 단계는 분류를 위한 특징 생성 과정이다. 한 문장의 특정 의미는 다수의 관련 의미마디 합성을 통해 형성된다. 독립 의미가 될 수 있는 범위는 하나의 의미마디를 이루고, 해당 연산 과정을 통해 의미마디값이 합성된 수치는 의미 합성값이 된다. 연산은 각 의미마디의 평균을 계산하는 방식으로 이루어진다. 의미마디는 의미값 연산기에 입력된 표 4의 규칙 패턴과 미리 결정된 관용표현 목록으로부터 인식된다.

표 4. 의미마디 패턴 규칙과 예

1. $Basic\ Rule_{concat}$ BC_{R01} : $md_{[0,1]} + (md\|a*)_{[0,1]} \leftrightarrow nc$ BC_{R02} : $md_{[0,1]} + (md\|a*)_{[0,1]} \leftrightarrow pa$ BC_{R03} : $md_{[0,1]} + (md\|a*)_{[0,1]} \leftrightarrow pv$ BC_{R04} : $md_{[0,1]} + (md\|a*)_{[0,1]} + pa + nc$ 2. $Sub\ Rule_{pa}$ pa_{R01} : $nc + xn$ pa_{R02} : $nca + xpv$ pa_{R03} : $ncs + xpa$ pa_{R04} : $nc + pa + exm$ pa_{R05} : $nc + jcm$ 3. $Sub\ Rule_{a}$ a_{R01} : $pv + ecs$ a_{R02} : $pa + xa$	1. $Basic\ Rule_{concat}$ 추출 예시 BC_{R01} : 이런/md 개/a 쓰레기/nc BC_{R02} : 이/md 얼마나/a 뛰어나/pa BC_{R03} : 이런/md 잘/a 알리/pv BC_{R04} : 역시/a 괜찮/pa 영화/nc 2. $Sub\ Rule_{pa}$ 추출 예시 pa_{R01} : 시사회/nc 적/xn pa_{R02} : 간단/nca 하/xpv pa_{R03} : 거대/ncs 하/xpa pa_{R04} : 연기력/nc 뛰어나/pa ㄴ/exm pa_{R05} : 최고/nc 의/jcm 3. $Sub\ Rule_{a}$ 추출 예시 a_{R01} : 미치/pv 도록/ecs a_{R02} : 더럽/pa 게/xa

패턴 규칙 $RULE_x$는 4가지 기본 규칙과 부사/형용사 역할 부분 규칙으로 나뉘고, 해당 품사의 유무 여부에 따라 기본 규칙이 부분 규칙으로 치환되어 총 104가지의 규칙으로 확장된다. 추출된 내용어 연쇄가 규칙에 해당하지 않으면 각 어절이 독

립적으로 의미마디를 이루고 이항연산을 거친다.

연산을 위한 함수 선택의 문제는 의미마디값이 연산된 이후 얻어지는 값의 범위에 대한 실험에 근거해서 결정했다. 산술평균(A), 기하평균(G), 조화평균(H)의 분포 범위에 대한 결과 비교에 따라 의미마디 단위 연산을 위해 기하평균이 선택되었다.

$$\sqrt[n]{\prod_{i=1}^{n}(mc_i)^n}$$

표 5는 세 가지 평균 방법을 비교한 자료이다. 이들은 항상 |H|≤|G|≤|A|의 관계를 갖는다. 이 평균값은 같은 평점의 다른 리뷰의 평균값과 비교해서 차이가 심하지 않아야 한다. 중심 내용어 역할을 하는 의미값은 관련 리뷰의 의미합성값에서 리뷰 의미의 유의미한 비중을 차지하기 때문에 결과값에 큰 차이가 발생한다면 적절한 방법이 아니다. 한편 출현 의미마디 개수가 증가해도 기하평균은 n 차원 평균값을 취하기 때문에 분절의 개수가 증가하는 경우를 문제없이 처리해준다. 서로 반대 극성의 의미마디 연쇄가 교차로 출현하는 경우, '+' 연산을 통해 서로의 의미값을 상쇄시키고 차이값이 남는다.

표 5. 문장 내 연산 산술/기하/조화평균 값 비교 예시

이런/md 개/a 쓰레기/nc 구역질/nc 영화/nc 에/jca 10/nnn 점/nbu 주/pv 는/exm 변태/nc 들/xn 은/jx 무엇/npd 이/jcp 지/ef ?/s. (평점 1)		
mc_1 {이런/md:1.100000 개/a:1.500000 쓰레기/nc:-5.36285}		$f : mc_i \times mc_{i+1}$
mc_2[구역질/nc:-5.4992]		f(mc_1, mc_2) = Val_1
mc_3[변태/nc:-2.5011 무엇/npd 이/jcp 지/ef:-2.5087]		f(Val_1, mc_3) = Val_2
$Mean_{Arithmetic}$	$Mean_{Geometric}$	$Mean_{Harmonic}$
-6.72423	-6.615835	-6.518846

문장단위 의미값 연산

문장단위 연산을 위해 형태소의 어미 정보 표지를 사용하여 문장경계 인식기가

구축되었다. 의미마디 단위의 연산을 거친 의미값은 문장을 경계로 의미마디값의 추가적인 연산이 제한된다. 만약 입력 리뷰가 여러 문장으로 이루어졌다면 문장 단위 연산 과정을 거치게 된다. 문장 간 연산은 이항적으로 이루어지는데, 만약 반대 극성이 출현하는 경우는 방향성 전환을 위해 두 값을 더하고, 같은 극성이 연쇄되는 경우는 조화평균을 이용한다.

$$Mean_{Harmonic} = \frac{n}{\sum_{i=1}^{n} \frac{1}{ms_i}}$$

의미마디 단위 연산과 달리 조화평균이 사용된 이유는 이미 부분 의미가 합성된 한 문장이 다음 문장과의 연산을 거칠 때, 그 문장의 의미는 발화행위의 최소 단위로써 다음 문장의 의미에 경도되어서는 곤란하기 때문이다. 표 6에서 두 리뷰의 차이는 부사어 '정말'에 의해 발생하는데, 두 리뷰 모두 전달하고자 하는 의미 정보는 '이 영화는 어이없다'와 '윤감독의 결과에 실망했다'이다. 의미값 연산기는 앞 문장에 대한 연산값으로 -5.501815, (1)번 리뷰 뒷 문장에 대한 연산값으로 -7.00231, (2)번 리뷰 뒷 문장에 대한 연산값으로 -9.803234를 배출한다. 그런데 산술평균/기하평균을 사용하면, 부사어와 같이 감정의 강도를 더 강하게 하려는 어휘에 의해 최종 의미값이 크게 영향을 받는다. 특히 (2)번 리뷰의 경우, 산술평균을 사용했을 때, 최종 분류기가 도출해낼 평점 1에 해당하는 리뷰의 최종 연산값 분

표 6. 문장 간 연산 산술/기하/조화평균 값 비교

(1)	윤감독님이 이런 어이없는 영화를 만드시다니. 정말 어이없네요. (평점 1)		
	이런/md:1.100000 어이없/pa:-5.00165	정말/a:1.400000 어이없/pa:-5.00165	
	$Sent_1$: -5.501815	$Sent_2$: -7.00231	
	$Mean_{Arithmetic}$: -6.252062	$Mean_{Geometric}$: -6.206884	$Mean_{Harmonic}$: -6.162033
(2)	윤감독님이 이런 어이없는 영화를 만드시다니. 정말 정말 어이없네요. (평점 1)		
	$Sent_1$: -5.501815	$Sent_2$: **-9.803234**	
	$Mean_{Arithmetic}$: -7.652525	$Mean_{Geometric}$: -7.344085	$Mean_{Harmonic}$: -7.048077

포 범위를 부사어만으로 뛰어넘는다. 표현이 더 강조될 경우, 기하평균값 역시 일반적 분포 범위를 뛰어넘을 가능성이 크다. 이는 다른 등급의 영역을 침범하거나 역전시키므로 분류기의 성능을 저하시킨다. 따라서 특정 문장의 의미값에 경도되지 않고, 전체 텍스트의 의미를 잘 유지해주는 조화평균을 사용함으로써 문장 간 의미값 합성과 등급 분류를 위한 만족스러운 결과를 얻을 수 있다.

SVM을 이용한 등급화

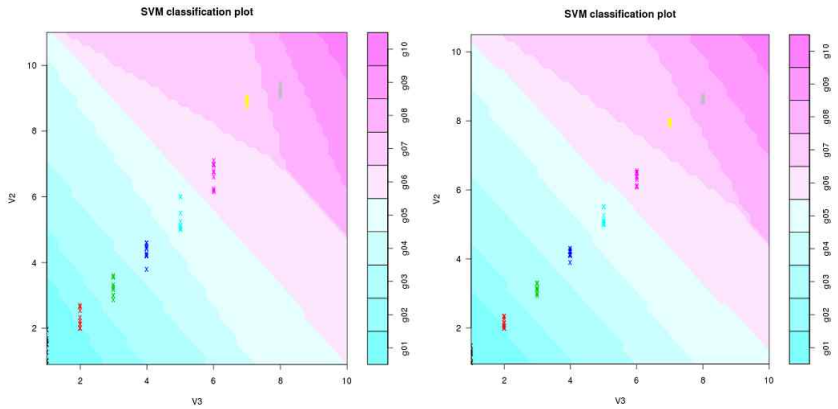


그림 8. 기계 학습 SVM 분류 결과
(좌: Unnormalized, 우: Normalized)

본 연구를 위한 실험 및 기계 학습을 위해 Cine21 영화평 말뭉치에서 100개의 훈련 집합을 일정한 기준에 따라 선정된 1000개의 리뷰 자료에서 평점별로 10개씩 무작위로 선정하였다. 감정어휘 평가사전에서 최대한 다양한 의미값을 추출해서 상호 교차적 연산을 실험을 위해 자료의 선정 기준을 다음과 같이 정의했다.

※ 자료 선정 기준 *리뷰 문서상에*

① 감정/평가 관련 표현이 풍부한가.

② 관련 표현이 포함되었다면, 성실하게 작성된 문서인가.

기준 ①을 기계가 추출해주면 기준 ②는 작업자의 직관에 따르는 반자동 작업이 진행된다. 위 기준을 위배하지 않았다면 기존에 할당된 해당 평점과 리뷰의 내용 일치도가 높을 것으로 판단되는 리뷰를 선정했다.

SVM은 두 부류 간에 존재하는 '여백을 최대화'하여 일반화 능력을 극대화한다. 즉, 두 부류간의 경계에 있는 자료에 초점을 맞춘다. 결정 초평면 $d(x)$는 전체 특징 공간을 두 영역($d(x)>0$, $d(x)<0$)으로 분할한다.

$$d(x)=w^T x + b = 0$$

이는 일정한 분포 자료를 분류하는 결정 초평면의 전체 부류에 대한 여백이 최대화되어 변형에 강한 특징을 갖기 때문에 SVM을 이용해서 등급화에 이용 가능한 다차원 분류 결과를 얻을 수 있다. 그림 8은 훈련 집합에 따라 학습된 결정 초평면과 그에 따른 특징 공간을 시각화한 자료이다. V3 축은 훈련 집합의 각 등급에 속한 의미합성값의 개별 인덱스, V2 축은 실제 합성된 의미값을 의미한다. 훈련 집합에 속한 개별 원소들은 지도 학습을 통해 표식된 상태이다. 각 분할 공간은 여백이 최대화되었기 때문에 이후 SVM 분류기가 등급의 '예측(prediction)' 작업을 수행할 때, 자료에 Outlier가 발생하더라도 오류율이 최소화된 결과를 얻는다. 분류기는 값을 새로 입력받으면 기존의 학습 모델과 비교하여 예측하는 과정을 통해 리뷰의 최종 등급을 결정해서 출력한다. 예를 들어, "*[ms00:-5.000000:손발/nc 이/jc 오그라들/pv] 다/ecx [ms00:1.500000:못하/px 어/ecs 없/pa 어/ecx 지/px] ㄹ/exm 것/nb 같/pa ㄴ/exm [ms01:-4.07115:유치하/pa] ㅁ/exn*"의 문장이 입력되면 다음과 같이 출력된다.

```
@ 손발/nc 이/jc 오그라들/pv 다/ecx 못하/px 어/ecs 없/pa 어/ecx 지/px
ㄹ/exm 것/nb 같/pa ㄴ/exm 유치하/pa ㅁ/exn # 2MC((-*+)-) : -5.477226

# 입력하신 리뷰는 1 등급 입니다.
# 등급 스케일 : 부정(최저) 1 - 긍정(최고) 10
```

그림 9. ARSSA 출력 결과

실 험

실험의 기초 및 방법

ARSSA는 자동 등급화를 위한 시스템이므로 SVM 분류기의 등급화에 대한 정확도를 통해 일반화 능력을 평가하고, 감정어휘 평가사전 1.0이 영화평 도메인에 특화되었다는 가정의 유효성을 확인하는 것이 본 실험의 두 가지 목적이다. 이를 위해서 자료 선정 기준에 따라 Naver Movie 영화평 말뭉치에서 각 평점 당 100개씩 총 1000개의 리뷰를 선정하여 무작위로 100개씩 총 10개의 부분집합으로 분할했고, 전체를 9:1의 비율로 훈련집합과 실험집합으로 나누어 10-fold Cross-Validation을 실시했다. 전체 자료 집합은 지금까지의 논의를 통해 구축된 시스템 ARSSA 모듈을 통해 분류기 투입 이전 단계까지의 모든 과정을 동일하게 거친 결과물이다. 이 실험을 통해 인간의 직관적 평점 할당이 일반화된 감정어휘 평가사전을 이용한 자동 등급화 시스템의 산출 결과와 얼마나 일치하는가를 살펴보았다.

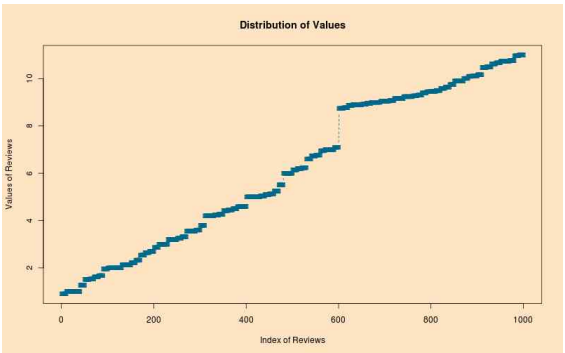


그림 10. 자료 집합 의미합성값 분포 결과

그림 10은 Naver Movie에서 선정된 1000개의 자료 집합 의미합성값 분포를 보인다. 전체 자료는 극성에 따라 일정 구간에 따라 분포되어 있을 것으로 짐작케 한다. 실험은 기존의 말뭉치에서 리뷰와 함께 주어진 각 평점을 기준으로 이루어진

다. 확보된 자료 집합에서 각 평점에 대해 무작위로 선정해서 검증을 위한 실험 집합으로 삼는다. 실험 집합의 평점 n 영화평이 입력되었을 때, 분류기가 n으로 분류시 TP, 입력 n에 대해 $\sim n$(not n)으로 분류시 FP, 입력 $\sim n$에 대해 n으로 분류시 FN, 입력 $\sim n$에 대해 $\sim n$으로 분류시 TN이다. 실험을 통해 이들 각각에 해당하는 수치는 분류 행렬에 입력된다.

표 7. Confusion matrix

분류결과 / 참부류	ω_1	ω_2
ω_1	True Positive (tp)	False Positive (fp)
ω_2	False Negative(fn)	True Negative (tn)

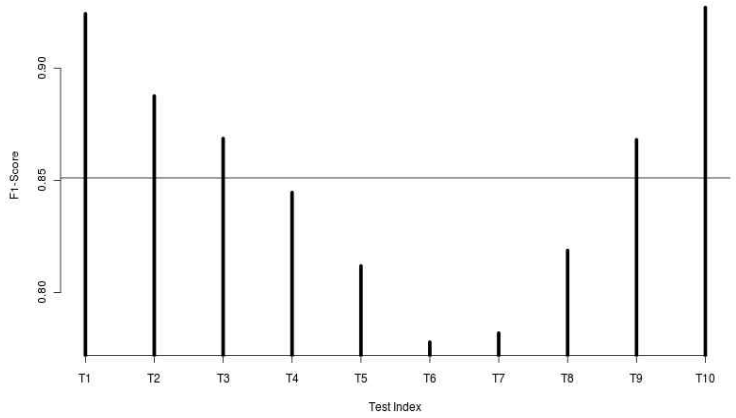
실험 결과 및 분석

리뷰 작성자가 부여한 평점을 정답 기준으로 했을 때, 본 시스템의 성능은 각 평점의 분류 결과별로 차이가 있고 평균적으로 약 85% 정도의 분류 성공률을 보인다. 이는 고민수(2010)에서 Cine21 영화평 말뭉치를 대상으로 실험한 결과와 거의 일치한다. 평균 분류 성공률을 기준으로 했을 때, F1 Score가 .85 보다 낮은 경우는 평점 5~8이다. 이러한 평점을 부여하는 리뷰 작성자는 영화에 대한 평가가 복합적이기 때문으로 판단된다. 이 경우는 사용된 어휘의 극성이 명확하지 않거나, 접속어미를 통한 극성의 방향에 변곡점이 생성되는 경우이다. 반대로 평점 1~4, 9~10에 해당하는 리뷰는 각각 부정과 긍정에 관련된 표현이 명확히 드러난다. 사용된 어휘의 극성이 일관적이며 감정표현이 비교적 강하게 드러나 있고 복합적 평가가 거의 없는 판단에 따라 해당 평점을 부여한 것이다.

같은 등급의 어휘와 접속어미를 사용해서 작성된 평점 6의 두 영화평이 있을 때, 한 영화평은 일반화된 기준에 따른 등급 6으로 정확히 분류된 반면, 다른 영화평은 등급 5로 분류되었다고 가정해보자. 이 차이의 이유는 첫째, 각 어휘와 표현에 대한 기준이 사람마다 서로 다르기 때문이고, 또는 둘째, 접속어미를 기준으로 어느 쪽 표현에 더 가중치를 부여해서 표현했는지 여부이다. 두 번째 이유에 의한 차이라면 똑같이 어휘 수준의 단서가 있지 않는 한, 기계적으로 판단이 불가능하

표 8. 분류기 전체 성능 평가 실험 결과 (단위: %)

평점	Accuracy	Precision	Recall	F1 Score
1	.9240792	.928	.9215694	.924229
2	.882	.884	.8914798	.8875584
3	.876	.859	.8788506	.8686324
4	.854	.842	.8473927	.8445679
5	.813	.806	.8181973	.811839
6	.781	.774	.7819678	.777881
7	.783	.778	.7861446	.7819342
8	.819	.818	.8197327	.818743
9	.87	.87	.8703761	.8680643
10	.926	.938	.9165059	.9269936
AVE:	.8539958	.8497	.8532217	.8510443



지만, 첫 번째 이유에 의한 차이라면 일반화된 의미 기준에 따라 기계에 의해 일괄적으로 처리된 경우라고 할 수 있다.

$$Performance_{ARSSA} = Performance_{Classifier} + Error_{reasonable}$$

따라서 분류 결과와 평점 간 오차의 유형을 '설명 가능 오류'와 '설명 불능 오류'로 나누어 볼 수 있다. 설명 가능 오류는 엄밀히 말해서 시스템을 통해 의도된 보정 결과이다. 따라서 본 실험 결과에 수치로 나타나있지 않지만, ARSSA 시스템의 실제 성능은 분류 결과에 설명 가능 오류를 합한 결과가 된다.

결 론

본 논문에서는 등급화가능성 갖는 유의어집합으로 이루어진 감정어휘에 대한 의미사전을 구축하고, 추출되는 어휘와 표현의 패턴에 관한 의미마디 연산을 통해 의미값을 얻고 리뷰 문서의 의미 등급을 분류하는 시스템에 대해 살펴보았다. 또한 검증 실험을 통해 시스템의 분류 성능을 확인함으로써, 분류 결과와 사람이 직관적으로 부여한 평점과의 일치도를 평가했다. 일반화된 기준을 기초로 한 시스템의 연산 체계에 의해, 영화평 리뷰의 등급 자동 분류가 가능했고, 실험에서 기존 평점과 비교해서 평균적으로 약 85%의 일치도를 보였다.

현재 감정어휘 평가사전의 처리 가능 범위를 확대시키기 위해서 수시로 관리자에 의해 신규 어휘 목록이 추가되고 각 도메인에 대한 어휘의미값이 갱신되고 있다. 이는 특정 도메인에 제한된 처리를 뛰어넘어 중의성 및 일반화 처리를 가능케 하는 사전 구축의 기본이 되는 것으로써 Esuli(2010)에서 어휘의 다양한 사전적 정의에 따른 처리에 의해 중의성을 해소하는 접근법과 차이를 보인다. 이후 연구에서 구축될 예정인 '감정어휘 평가사전 2.0'은 실용적으로 중의적 의미의 처리를 지향한다. 즉, 기존에 확보된 유의어집합의 범주에서 소속 어휘의 의미값을 어휘의 도메인 별 용례와 관찰값에 따라 의미값을 세분화함으로써 중의성을 해소한다.

향후 통사적으로 발생하는 문제의 적절한 해결과 문맥에 따른 중의성에 관련된 정보를 처리하는 방법에 대한 연구와 함께 입력 문서의 주제 분석에 관한 연구가 병행되어야 한다.

참고문헌

[1] 고민수. 2010. 감정어휘 평가사전에 기반한 영화평 리뷰 자동 분류 및 등급화 시스템. 서울대학교 석사학위논문.

[2] Hatzivassiloglou V., Wiebe J.. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. COLING-2000.

[3] Hong Yu and Hatzivassiloglou V.. 2003. Towards Answering Opinion Questions:

Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. Proceedings of EMNLP-03.

[4] 강인호. 2007. 감정 표현구 단위 분류기와 문장 단위 분류기의 결합을 통한 주관적 문장 분류의 성능 향상, 정보처리학회논문지: B 14-B권 7호.

[5] 명재석 · 이동주 · 이상구. 2008. 반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템. 정보과학회논문지.

[6] Martin J. R., White P. R. R.. 2005. The language of evaluation: appraisal in English. Palgrave Macmillan.

[7] Whitelaw C., Garg N., Argamon Shlomo. 2005. Using Appraisal Taxonomies for Sentiment Analysis. In Paper session IR-8 (information retrieval): sentiment and genre classification.

[8] 황재원, 고영중. 2008. 감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템. 정보과학회논문지.

[9] PANG, B., LEE, L., and VAITHYANATHAN, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp.79-86, Philadelphia, PA.

[10] Kushal Dave, Steve Lawrence, David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. Proc. of the WWW.

[11] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. KDD 04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining.

[12] Esuli A., Sebastiani F., 2006, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06

[13] Baccianella S., Esuli A., Sebastiani F.. 2010. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC 2010 - Seventh conference on International Language Resources and Evaluation (Valletta, Malta, 18-22 maggio 2010). Proceedings, pp. 2200 - 2204. ELRA.

[14] Bolstad B. M., Irizarry R. A, Åstrand M., Speed T. P.. 2003. A Comparison of

Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. Bioinformatics Vol. 19 no. 2.

[15] 윤애선, 권혁철. 2010. 감정 온톨로지의 구축을 위한 구성요소 분석. 인지과학. Vol. 21 No. 1.

[16] 최운천 · 김기형. 2008. 한국어 낱말망(Korean Wordnet)과 우리말유의어분류대사전 구축. 한국사전학 제 11호.

1 차원고접수 : 2010. 10. 19
2 차원고접수 : 2010. 12. 1
최종게재승인 : 2010. 12. 2

(*Abstract*)

Grading System of Movie Review through the Use of An Appraisal Dictionary and Computation of Semantic Segments

Minsu Ko Hyopil Shin

Dept. of Linguistics, Seoul National University

Assuming that the whole meaning of a document is a composition of the meanings of each part, this paper proposes to study the automatic grading of movie reviews which contain sentimental expressions. This will be accomplished by calculating the values of semantic segments and performing data classification for each review. The ARSSA(The Automatic Rating System for Sentiment analysis using an Appraisal dictionary) system is an effort to model decision making processes in a manner similar to that of the human mind. This aims to resolve the discontinuity between the numerical ranking and textual rationalization present in the binary structure of the current review rating system: {rate: review}. This model can be realized by performing analysis on the abstract menas extracted from each review. The performance of this system was experimentally calculated by performing a 10-fold Cross-Validation test of 1000 reviews obtained from the Naver Movie site. The system achieved an 85% F1 Score when compared to predefined values using a predefined appraisal dictionary.

Keywords : Appraisal Dictionary, Synonymy, Synset, Grading, Semantic segment, SVM, ARSSA