http://www.calsec.or.kr/jsebs
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.19, No.4, Nov 2014, pp.1-19
http://dx.doi.org/10.7838/jsebs.2014.19.4.001

빅데이터 분석을 위한 한국어 SentiWordNet 개발 방안 연구 : 분노 감정을 중심으로

The Study of Developing Korean SentiWordNet for Big Data Analytics : Focusing on Anger Emotion

최석재(Sukjae Choi)*, 권오병(Ohbyung Kwon)**

초 록

빅데이터 내에 존재하는 감정 정보를 추출하여 사용자들이 특정 대상에 대하여 갖고 있는 인식이 어떠한지를 파악하고자 하는 노력이 활발히 이루어지고 있다. 상품, 영화, 그리고 사회적 이슈 등에 대한 문장을 분석하여 사람들이 해당 주제에 어떠한 견해를 가지고 있는지를 분석하고 측정하여 구체적인 선호도를 알아내는 것이다. 문장에서 드러나는 감정 정도를 얻기 위해서는 감정어휘의 목록과 정도값을 제시할 수 있는 감정어휘사전이 필요하므로 본 연구에서는 감정어휘를 발견하는 방법과 이들의 정도값을 결정하는 문제를 다룬다. 기본적인 방법은 기초 감정어휘의 목록 수집과 이들의 정도값은 선행연구 결과와 직접 설문 방식을 이용하고, 확장된 목록의 수집과 정도값은 사전의 표제어 설명부(glosses)를 이용해 추론하는 것이다. 그 결과 발견된 감정어휘는 전형성을 띠고 있는 기본형 감정어휘, 기본형 감정어휘의 gloss에 사용된 확장형 1단계 1층위 감정어휘, 비 감정어휘 중 gloss에 기본형 또는 확장형 감정어휘를 가지고 있는 확장형 2단계 1층위 감정어휘, gloss의 gloss에 기본형 또는 확장형 감정어휘가 사용된 확장형 2단계 2층위 감정어휘의 네 종류로 나뉜다. 그리고 확장형 감정 어휘의 정도값은 기본형 감정어휘의 정도값을 기초로 문형의 가중치와 강조승수를 적용하여 얻었다. 실험 결과 AND, OR 문형은 내포된 어휘의 감정 정도값을 평균 내는 가중치를, Multiply 문형은 정도 부사어의 종류에 따라 1.2~1.5의 가중치를 갖는 것으로 파악되었다. 또한 NOT 문형은 사용된 어휘의 감정 정도를 일정 정도로 낮추어 역전시키는 것으로 추정된다. 또한 확장형 어휘에 적용되는 강조승수는 1층위에서 2, 2층위에서 3을 갖는 것으로 예상된다.

ABSTRACT

Efforts to identify user's recognition which exists in the big data are being conducted actively. They try to measure scores of people's view about products, movies and social

이 논문은 2014년도 BK21 플러스 사업의 지원을 받아 수행된 연구임(경희대학교 경영대학 데이터 과학에 기반한 경영 전문 연구인력 양성팀).
This work was supported by the IT R&D program of MSIP/KEIT. [10047255, Development of Smart Wellness Companion Service System]'.

* First Author, School of Management, Kyung Hee University(sjchoi@khu.ac.kr)
** Corresponding Author, School of Management, Kyung Hee University(obkwon@khu.ac.kr)

2014년 07월 04일 접수, 2014년 09월 05일 심사완료 후 2014년 09월 11일 게재확정.

issues by analyzing statements raised on Internet bulletin boards or SNS. So this study deals with the problem of determining how to find the emotional vocabulary and the degree of these values. The survey methods are using the results of previous studies for the basic emotional vocabulary and degree, and inferring from the dictionary's glosses for the extended emotional vocabulary. The results were found to have the 4 emotional words lists (vocabularies) as basic emotional list, extended 1 stratum 1 level list from basic vocabulary's glosses, extended 2 stratum 1 level list from glosses of non-emotional words, and extended 2 stratum 2 level list from glosses' glosses. And we obtained the emotional degrees by applying the weight of the sentences and the emphasis multiplier values on the basis of basic emotional list. Experimental results have been identified as AND and OR sentence having a weight of average degree of included words. And MULTIPLY sentence having 1.2 to 1.5 weight depending on the type of adverb. It is also assumed that NOT sentence having a certain degree by reducing and reversing the original word's emotional degree. It is also considered that emphasis multiplier values have 2 for 1 stratum and 3 for 2 stratum.

키워드 : 빅데이터, 감성워드넷, 감정어, 감정 정도, 문장비중
Big Data, Sentiwordnet, Emotional Words, Emotional Degree, Sentence Weight

1. 서 론

빅데이터 내에 존재하는 감정 정보를 추출 및 활용함으로써 특정 대상에 대한 사용자들의 인식을 파악하고, 서비스 만족도를 높이려는 노력이 다양한 분야에서 이루어지고 있다. 연구들은 주로 인터넷 게시판, SNS, 블로그 등에 올려진 사용자 텍스트를 분석하여 특정 주제에 대한 사용자의 감정을 인식하고, 향후 사용자의 긍정적인 반응이 극대화될 수 있게 하는 자료를 제시하는 데 초점이 있다. 이를 위해 빅데이터에서는 감정분석(sentiment analysis)에 대한 연구가 활발하다[3, 25].

그런데 비정형 대규모 텍스트를 통하여 사용자들이 대상에 대해 가지고 있는 감정을 이해하려면 텍스트를 구성하는 기본 단위에 대한 감정 정보를 가지고 있어야 한다. 문장은 형태소 또는 단어라는 기본 단위가 매번 다른 방법으로 서로 연결되는 것이므로, 이에 대한 정보가 있어야 절(節, clause)과 문장, 그리고 텍스트 전체라는 보다 큰 단위에 대해서도 감정 정보도 얻을 수 있기 때문이다.

그러나 아직까지 감정 정보를 얻을 수 있는 적절하고도 공개된 자원이 마련되어 있지 않아 감정 인식과 관련된 연구와 서비스가 활성화되지 못하고 있다. 영어를 대상으로는 각 단어에 대한 긍정/부정 값이 부여된 SentiWordNet이 공개되어 있기도 하지만, 자체의 한계가 있을 뿐만 아니라, 영어를 기계적으로 번역하면 감정 정도가 그대로 유지되지도 않는다. 언어는 그 언어 사용자의 문화를 다면적으로 함축하고 있기 때문이다. 따라서 한국어에 특화된, 그러면서도 개선된 결과를 내놓을 수 있는 별도의 자원을 구축하는 것이 필요하다.

이에 본 연구는 언어의 기본 단위인 단어에 대하여 감정 정보를 부여하는 한국어 감정어휘사전을 구축하는 방법을 제안한다. 단어

의 의미를 가장 잘 기술하고 있는 사전 표제어의 해설 부분(glosses)을 이용하여 감정어휘 목록을 확장하고, 감정 정도를 계산하는 방법을 제시하여 다양한 환경에서 사용자의 감정 정보를 얻을 수 있는 한국어 SentiWordNet을 구축할 수 있게 하고자 한다.

2. 관련 연구

2.1 감정어휘의 개념과 기준

감정이라는 개념은 인간에게 가장 추상적인 영역 중 하나로서 그 실체를 파악하기가 쉽지 않다. 기존 연구에 의하면, 감정은 내적 또는 외적 사상(事象)에 대한 직접적인 반응으로서, 비의도적으로 나타나는 정신적인 상태이며[21], 대상을 평가하고, 생리적 변화를 수반한다[17]. 따라서 감정어휘는 무엇을 평가하는 기능을 가질 뿐만 아니라, 비의도적인 신체적 변화를 연상시킬 수 있어야 한다.

이러한 기준에 따라 '기쁘다, 슬프다'는 감정어휘가 되지만 '웃다, 울다'는 평가성이 없어 감정어휘가 되지 않는다. 또한 '흡족하다, 못마땅하다'는 감정어휘가 되지만, '부드럽다, 거칠다'는 생리적 변화를 수반하지 않으므로 감정어휘가 되지 못한다. Shaver 등[22]은 이를 토대로 설문조사를 거쳐 135개의 영어 감정어휘를 도출하였고, 같은 취지로 Gim[12]은 한국어에 대해서 494개의 감정어휘를 제시한 바가 있다.

그러나 위의 기준은 너무 엄밀하여 사용자의 감정을 인식하는 응용 시스템에서 사용하기에는 수가 제한되는 문제가 있다. 예를 들어, '웃다, 울다'가 감정어휘로 여겨지지 않으면 영화 리뷰 분석 등이 제대로 이루어지지 않을 것이며, '부드럽다, 거칠다'가 배제되면 의류 구매 후기를 분석할 때 문제가 될 것이다.

그런데 서비스 만족도를 제고하려는 시스템, 예컨대 상품평을 분석[1, 13, 23]하거나, 영화, 호텔 리뷰 문서를 분석[18, 7]하는 시스템을 위해서 위에서 언급된 감정어휘의 기준이 모두 필요하지는 않다. 이러한 경우에는 사용자가 특정 대상을 어떻게 평가하는지가 중요하지, 그로 인해 생리적인 변화가 생겼는지 여부는 크게 중요하지 않은 것이다. 그렇다면 감정어휘의 기준인 '평가성', '비의도성', '생리적 변화' 중 사용자의 평가를 반영하는 '평가성' 기준만 충족되면 사용자 만족도를 파악할 수 있는 것이다.

다시 말해, 그것이 감정어휘인지 여부보다는 그 어휘를 통하여 사용자가 긍정적인 마음을 갖고 있는지 또는 부정적인 마음을 갖고 있는지를 직접적 또는 암시적인 이유에서라도 알 수 있다면 충분히 수집 대상 어휘가 될 수 있고, 그 결과 우리의 요구를 충족하는 감정어휘의 수는 훨씬 늘어날 수 있다. 다만, 이러한 어휘는 긍정/부정적 평가와 관련이 없는 경우에도 많이 쓰이므로 전형적인 감정어휘와는 차이를 두어야 한다.

2.2 어휘의 긍정/부정 속성 판단

전형적인 감정어휘의 조건을 모두 충족시키지는 않지만, 어휘가 긍정(positive) 또는 부정(negative) 속성을 띠는지를 밝히려는 시도로는 다음과 같은 연구가 있다. Hatzivassiloglou and McKewon[14]은 코퍼스에서 1,336개의 형

용사를 임의 선택하고 이들의 긍정/부정 속성을 결정하였다. 그리고 이들이 접속사로 어떠한 단어가 연결되어 있는지를 보아 and로 연결된 형용사에는 동일한 긍정/부정 속성을, but로 연결된 형용사에는 반대의 속성을 부여하여 어휘를 확장하였다.

Turney and Littman[24]은 보다 적은 양의 기초 어휘(seed terms)를 결정하고 이를 확장하는 방안을 사용하였다. 긍정 어휘로 good, nice, excellent, fortunate, correct, superior의 7개를, 부정 어휘로 bad, nasty, poor, unfortunate, wrong, inferior의 7개를 기초 어휘로 삼고 이 어휘들과 함께 사용되는 어휘들의 빈도를 계산하여 긍정/부정 어휘를 확장하였다.

WordNet이 구축된 이후에는 이를 활용하여 어휘의 감정 속성을 결정하는 방안이 제시되었다[20]. Kamps 등[16]에서는 긍정 기초 어휘로 good을, 부정 기초 어휘로 bad를 두고 동의어 관계로 목표 어휘와 good 및 bad와의 최단 길이를 구하였다. 그리고 전자가 후자보다 더 짧으면 긍정적인 단어로, 그 반대이면 부정적인 단어로 보았다.

WordNet을 통하여 어휘의 긍정/부정 속성을 얻으려는 시도에서 가장 대표적인 것은 SentiWordNet의 구축과 관련된 연구[2, 8, 9, 10, 11]이다. SentiWordNet은 [24]에서 제시된 14개 기초 긍정/부정 어휘와 WordNet을 이용하여 대상 어휘가 기초 어휘와 어떠한 관계에 있는지를 파악한다. 동의어 관계와 상/하의어 관계에 있다면 같은 긍정/부정 속성을 가지게 하고, 반의어 관계에 있다면 반대 속성을 가지게 하여 감정 속성이 부여된 어휘의 숫자를 늘려나간다.

그러나 WordNet이 보여주는 관계망은 매우 성글어서 이를 통해 속성을 확장할 수 있는 어휘는 제한적이다[19]. 때문에 관계망에 의하여 파악되지 못한 어휘들에 대해서는 이제까지 얻어진 감정어휘들이 WordNet 각 어휘의 해설 부분(glosses)에 얼마나 나타났는지를 양적으로 분석하여 긍정/부정 속성을 결정하고, 감정 정도값을 결정한다.

이처럼 SentiWordNet은 각 어휘에 대하여 긍정적인 어휘에 가까운지, 부정적인 어휘에 가까운지를 결정해줄 뿐만 아니라, 각 속성의 정도값까지 알려주기 때문에 잠재적 활용도가 높다. 현재 300여 이상의 연구팀에서 다양한 프로젝트를 위하여 사용되고 있다[2].

하지만 SentiWordNet은 정확성이 떨어지는 점이 가장 큰 문제이다[26]. 예를 들어 동사 blame은 부정적인 견해를 드러내는 단어임이 틀림없어 보이는데도 SentiWordNet에서는 이들의 긍정/부정 정도값이 모두 0으로 기록되어 있다. 또 다른 예로, 해부학 용어로서 중립적으로 여겨지는 'pia mater(연뇌막)'라는 어휘는 부정값은 없이 긍정값만 1.0만점에 0.5에 이른다.

SentiWordNet의 정확도가 떨어지는 일차적인 이유는 정도값 계산의 근거 자료로 사용된 glosses가 WordNet의 것이기 때문이다. WordNet은 어휘 사이의 관계를 형성하는 데 주목적이 있는 것이어서 일반 사전처럼 자세한 해설을 하지 않는다. 예를 들어, 일반 사전인 Collins Cobuild Advanced Learner's English Dictionary[6]는 'blame'의 뜻으로 'If you blame a person or thing for something bad, you believe or say that they are responsible for it or that they caused it'라고 기술하고 있는데 특히 'bad'라는 단어를 통해 표제어 'blame'

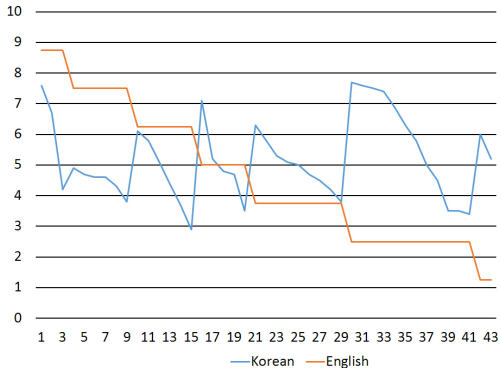
에는 부정적 속성이 있음을 알 수 있다. 그러나 WordNet의 gloss는 'attribute responsibility to'로서 핵심만을 소략하게 기술하여 감정 속성을 얻기에 충분한 자료를 제공하지 못하는 경우가 발생한다.

정확도가 떨어지는 또 다른 이유로 문형에 관한 고려가 충분하지 못한 점도 들 수 있다. 'honest'라는 단어는 WordNet에서 7가지 다의어를 가진 것으로 분석되는데(#1~#7), SentiWordNet에서는 이 중 5개에는 긍정 정도값을 주고 있지만, honest#4와 honest#7의 2개에 대해서는 부정 정도값을 훨씬 많이 주고 있다(긍정값 0.25, 부정값 0.625). 그런데 honest#4의 해설을 보면, '위선적이지 않음(without pretensions)'이고, honest#7도 '속이거나 훔치지 않고 얻음(gained or earned without cheating or stealing)'으로서 모두 'without'이 존재하는 공통점을 가진다. 결국 부정 극어 등 정도값 계산에 크게 영향을 미치는 요소를 충분히 고려하지 않았음을 알 수 있다.

감정에 대한 세부 분류 없이 긍정/부정의 양 극성 정보만을 제공한다는 점도 문제이다. 긍정/부정 견해는 감정의 상위 개념으로서 감정보다 더욱 모호한 개념이기 때문에 서비스 만족도를 위해서는 보다 세분된 감정 정보가 필요한 경우가 많다[26, 15]. 예를 들어, '영화가 너무 슬퍼서 펑펑 울었어요'라는 영화 리뷰가 있다고 할 때 '슬프다'와 '울다'는 부정 어휘에 속하므로 이 리뷰를 영화에 대한 부정적인 리뷰로 파악한다면 이는 오분석한 것이다. 긍정/부정의 의미는 맥락에 따라 달라지는 상대적인 개념이다.

SentiWordNet은 한국어로 직접 번역하여서는 적절한 결과를 얻을 수 없다는 점도 큰 걸림돌이다. 직접 번역할 수 없는 이유는 몇 가지가 있다. 첫째, 한국어에서는 단어인 것이 영어에서는 '구(句, phrase)'로 존재하여 대역 단어를 찾을 수 없는 경우가 있다. 이러한 예로는 '신물나다(sick of), 지긋지긋하다(fed up), 발끈하다(fly into a rage), 부아나다(lose one's temper)'와 같은 것들이 있다.

둘째, 두 언어의 감정 정도값이 일치하지 않는 경우가 많다. 한국어 분노 감정어휘에 대한 연구 결과[5]와 영어 SentiWordNet의 결과를 비교해 보면 양 언어에서 단어로 잘 판별되는 44개 단어에 대하여 다음과 같은 정도값의 차이가 있음을 알 수 있다.



〈Figure 1〉 Comparing the Degree of Korean and English Emotional Words

<Figure 1>은 같은 단어에 대해서 점수의 차이가 큰 것을 보여준다. 가장 큰 차이를 보여준 단어는 '노발대발하다(7.7점)'-'infuriated(2.5점)'으로서 둘의 점수 차이는 5.2점에 이른다. 또한 전체적으로 점수의 폭도 차이가 크다. 영어의 경우에는 정도값의 범위가 1.25~8.75까지로 폭이 넓은 편이나, 한국어의 경우에는 2.9~7.7로 보다 좁다. 이는 한국어와 영

어가 나타내는 표현 영역이 서로 다름을 보여주는 것이다.

셋째, 한국어에서는 서로 다른 어휘로 존재하는 어휘들이 영어로는 같은 어휘로 대역되는 경우 적절한 정도값을 얻지 못한다. 예를 들어 '역정나다, 성질나다, 화나다, 노하다, 분하다, 성나다, 약오르다, 골나다' 등은 서로 다른 정도값을 갖는데 영어로는 모두 'angry'로 대역된다. 그러나 '역정나다'에 비하여 '골나다'는 훨씬 낮은 정도의 분노 감정을 표현한다.

따라서 한국어 감정 분석을 위하여 한국어 SentiWordNet을 별도로 구축할 필요가 있다. 새롭게 구축되는 SentiWordNet은 한국어의 특성을 잘 반영할 뿐만 아니라, 앞서 제기한 영어 SentiWordNet의 정확도와 감정 분류 문제를 해결할 수 있어야 할 것이다. 이제 감정어휘의 목록 수집과 정도값 추론 방법을 중심으로 한국어 SentiWordNet을 구축하는 방안에 대해 논의한다.

3. 감정어휘의 선정

3.1 기본형 감정어휘와 확장형 감정어휘

감정어휘를 선정함에 있어 우선 고려되어야 하는 것은 언어학적 기준을 모두 충족시키는 감정어휘의 수는 제한적이라는 점이다. 따라서 실제에서 사용하기 위해서는 조건을 잘 충족시키는 전형적인 감정어휘, 즉 기본형 감정어휘를 기초로 어휘의 목록을 확장할 필요가 있다. 이 확장형 감정어휘는 감정어휘의 특징을 모두 갖지는 못하나 감정어휘를 필요로 하는 영역에서 현실적 필요는 충족시킨다.

확장형 감정어휘는 상황에 따라 감정을 잘 드러내지 못할 수 있다. 예를 들어, '깨끗하다'라는 단어는 일반적으로는 '집이 깨끗하다'와 같이 쓰여 '기쁨' 감정을 드러낸다고 여겨지나, '밥그릇을 깨끗하게 비웠다'와 같은 경우에는 감정 표현보다는 객관적 기술에 가깝다.

반대로 감정 속성이 없을 것 같은 어휘가 감정을 반영하는 경우도 있다. '하얗다'는 색채어로서 '종이가 하얗다'와 같이 많은 경우 감정을 드러내지 못하는 중립적 어휘이지만 '얼굴이 하얗다', '낯빛이 하얗다' 등 신체 관련 어휘와 함께 쓰이면 '놀람' 혹은 '두려움'의 감정을 보일 수 있다.

따라서 어느 상황에서나 감정을 드러내는 기본형 감정어휘와 일부 상황에서 감정을 드러내는 확장형 감정어휘를 구분하는 것이 필요하다. 그리고 이러한 구분이 어휘의 감정 정도값에도 반영되도록 할 필요가 있다.

3.2 감정어휘 수집의 기본 방법

기본형 감정어휘의 수집은 기존의 연구를 기초로 확보할 수 있다. Gim의 연구[12]는 한국어에 대한 기존의 연구 결과를 토대로 어휘 목록을 다시 검토한 것이고, Choi의 분류[4]는 이 어휘들을 Shaver 등[22] 일반 언어학적으로 많이 쓰이는 분류 체계에 따라 정리한 것이므로 이들을 1차적인 자료로 삼을 수 있다. 감정의 종류는 '사랑(Love), 기쁨(Joy), 놀람(Surprise), 분노(Anger), 슬픔(Sadness), 두려움(Fear)'의 여섯 가지가 사용된다. 확장형 감정어휘에 대해서는 그 목록이 정리된 바 없으므로 새롭게 구축해야 한다.

확장형 감정어휘를 얻는 기본적인 방법은

사전의 해설 부분인 glosses를 이용하여 기본형 감정어휘의 유의어를 찾는 것이다. 사전은 오랜 시간에 걸쳐 단어의 의미를 쉽고 자세하게 기술해 놓은 것이므로 내용의 정확도에 높은 신뢰도를 줄 수 있다. 영어의 SentiWordNet이 단어의 긍정/부정값을 구할 때 WordNet의 glosses를 이용한 것도 같은 맥락으로 해석된다.

그러나 영어 SentiWordNet은 각 단어의 glosses에 쓰인 단어에서 기 파악된 긍정/부정 어휘를 검출하여 양적 조사로 정도값을 결정하는 것이지만, 본 연구는 기본형 감정어휘가 기술될 때 사용된 단어에서 유의어를 찾아 감정어휘를 확장하는 방법을 사용하는 것으로 접근 방법은 다르다. 사전의 어휘 설명은 유사한 어휘를 통해 이루어지는 것이므로 유의어 관계에 속하는 어휘들을 많이 확보할 수 있다. 자세한 수집 방법에 대해서는 다음 절에서 논의한다.

본 연구는 사전의 신뢰도를 더욱 높이기 위하여 국가적 사업으로 구축된 표준국어대사전을 이용한다. 표준국어대사전은 일반적인 목적의 사전으로서 해설 부분이 충실히 기술되어 있으며, 국가 대표적인 사전으로 인식되고 있어 대표성을 가진다. 또한 온라인상에 구축되어 있어 접근이 용이하다는 장점도 있다.

3.3 확장형 감정어휘의 수집

확장형 감정어휘의 수집은 기본형 감정어휘로부터 시작하여 찾는 Top-down 방식의 1단계와 기본형이 아닌 어휘로부터 시작하여 감정어휘를 찾는 Bottom-up 방식의 2단계가 있다.

1단계는 기본형 감정어휘로부터 시작하는 것으로서 표제어 glosses에 사용된 어휘를 파악하는 것이다. 사전에서 어떤 표제어를 기술하는 데 사용된 어휘들은 표제어와 같은 속성을 상당히 갖추고 있는 유의어라고 할 수 있다. 따라서 표제어가 기본형 감정어휘라고 한다면, glosses에 사용된 어휘들도 같은 속성을 상당히 갖추고 있는 유사 감정어휘, 즉 확장형 감정어휘라고 할 수 있다.

기본형 감정어휘인 '만족하다'라는 단어로 예를 들어보면 다음과 같다.

1. '만족하다'의 다의어 중 해설 부분에 감정어휘가 사용되지 않은 것을 찾는다.
2. '만족하다[Ⅱ-2]'는 '모자람이 없이 충분하고 넉넉하다'라는 gloss를 갖는다. 이 중 수식어부를 제외하여 '충분하다, 넉넉하다'라는 두 개의 단어를 얻는다.

'만족하다'의 몇 가지 다의어 중 [Ⅱ-2]는 '모자람이 없이 충분하고 넉넉하다'로서 기본형 감정어휘가 사용되지 않아 조사 대상이 된다. 이 중 '모자람이 없이' 부분은 수식어이므로 소거하고, 술어부에 사용된 단어 '충분하다'와 '넉넉하다' 두 가지를 얻는 것이다.

본 연구에서는 기본형 분노 감정어휘 중 자주 사용되는 61개를 대상으로 위의 방법을 사용하니 확장형 감정어휘 21개를 얻을 수 있었다. 그 목록은 <Table 1>과 같다.

이러한 방법으로 감정어휘의 수를 확장할 수 있으나, 그 수가 충분하지 못한 것이 문제이다. 위의 경우에도 감정어휘의 수를 30% 정도 늘린 정도에 그쳤다. 따라서 보다 적극적인 방법으로 감정어휘의 수를 늘릴 방안을

〈Table 1〉 Examples of Extended Emotional Vocabulary

(GamJeongI)IlEoNaDa((감정이)일어나다), GeoBukHaDa(거북하다), KkeoRiDa(꺼리다), KkeoRimChikHaDa(꺼림칙하다), DalGapJi-AnDa(달갑지 않다), DapDapHaDa(답답하다), TtaBunHaDa(따분하다), (MaEumE)DulJi-AnDa((마음에)들지 않다), MipSalSuRupDa(밉살스럽다), SokJulEopDa(속절없다), SiDulHaDa(시들하다), SsuRiDa(쓰리다), APuDa(아프다), YakPpaRuDa(약빠르다), EokUlHaDa(억울하다), JoChi-AnDa(좋지 않다), JinJeolMeoRiNaDa(진절머리나다), JilRiDa(질리다), ThatHaDa(탓하다), HanThanHaDa(한탄하다), ((HimI)PpaJiDa((힘이)빠지다).

찾아야 하므로 Bottom-up으로 찾아가는 2단계 방안을 제시한다.

2단계는 감정어휘가 아닌 것으로부터 시작하는 것으로서 표제어의 glosses에 감정어휘가 사용되었는지를 보는 것이다. 표제어의 glosses에 감정어휘가 사용되었다면 표제어도 속성을 공유하니 감정어휘라고 할 수 있다. 이때 감정어휘는 기본형 감정어휘와 1단계에서 발견된 확장형 감정어휘이다.

여기에 glosses에서는 감정어휘가 나타나지 않았지만, glosses의 glosses 즉, Nested Glosses에서 감정어휘가 나타나는 경우가 있는데 이들도 전형성은 다소 멀어지지만 감정어휘라고 할 수 있을 것이다. '깨끗하다'라는 단어를 예로 찾는 과정을 살펴보면 다음과 같다.

1. '깨끗하다'의 다의어 '깨끗하다[4]'는 '맛이 개운하다'라는 gloss를 갖는다. '개운하다'는 기본형 감정어휘이므로 '깨끗하다[4]'는 확장형 감정어휘가 된다. 층위 '1'을 기록한다.
2. 그러나 '깨끗하다[1]'의 gloss는 '사물이 더럽지 않다'로서 감정어휘가 발견되지 않으나, gloss에 사용된 단어 '더럽다'의 gloss는 '못마땅하거나 불쾌하다'로서 기본형 감정어휘가 발견된다. 따라서 상위의 '깨끗하다[1]'는 확장형 감정어휘가 된다. 층위 '2'를 기록한다.

즉, '깨끗하다[4]'는 1층위에서 발견된 확장형 감정어휘이고, '깨끗하다[1]'은 2층위에서 발견된 확장형 감정어휘이다. 이 두 단어는 모두 확장형 감정어휘이나, 2층위에서 발견된 것은 보다 낮은 정도값을 가질 것이다. 2층위 이하에서도 찾아볼 수 있으나 현실적으로 더 내려가 찾기는 어렵다는 점과, 정도값이 부여되더라도 매우 낮은 정도만이 부여될 것을 고려하여 더 깊이는 내려가지 않는다.

입력된 값이 여러 다의어 중 어떤 어휘에 속하는지를 가려내기 위하여 각 다의어에 사용된 예문들의 실질 어휘도 기록한다. 예를 들어, '깨끗하다[1]'에 사용된 예문은 '그릇을 깨끗하게 씻다'이므로, '그릇, 씻다'를 기록하고, '깨끗하다[4]'에 사용된 예문은 '입맛이 깨끗하다'이므로 '입맛'을 기록한다. 이 정보는 다의어에 따라 의미 차이가 클 때 유용하게 사용될 수 있다.

이제까지의 내용을 토대로 감정어휘의 종류를 구분하면 다음과 같다.

1. 기본형 감정어휘
2. 확장형 1단계 1층위 감정어휘
3. 확장형 2단계 1층위 감정어휘
4. 확장형 2단계 2층위 감정어휘

이들을 저장하는 DB는 다음과 같이 설계된다.

〈Table 2〉 DB Design

Type	Headword		
	Polysem 1	Polysem 2	Polysem 3
Emotion	type	Type	type
Basic/Extended	text	Text	text
Stratum	1 or 2	1 or 2	1 or 2
Level	1 or 2	1 or 2	1 or 2
Emotional Degree	value	Value	value
Examples' words	word	word	word

4. 감정 정도값 계산

4.1 감정 정도값 계산의 대상과 방법

입력문이 가지고 있는 감정 정도를 파악하기 위하여 정도값 계산을 해야 하는 대상은 두 가지이다. 하나는 이제까지 논의한 감정어휘의 정도값이고, 다른 하나는 입력문의 정도값이다.

감정어휘의 정도값 계산은 단일 어휘에 대한 정도값을 계산하는 것이므로 가장 단순하게는 모든 어휘에 대하여 설문조사로 정도값을 얻는 방법을 생각해 볼 수 있다. 설문조사는 각 어휘의 감정 정도를 얻는 단순하면서도 매우 직접적인 방법이다.

그러나 어휘의 수가 매우 많기 때문에 모든 어휘에 대하여 설문조사를 한다는 것은 현실적으로 불가능하다. 따라서 seed가 되어주는 기초 어휘에 대해서만 설문조사를 하고, 다른 어휘는 그 결과를 바탕으로 유추하는 것이 바람직하다.

감정어휘 중 기본형 감정어휘는 감정어휘의 성격을 모두 잘 갖추고 있어 기초 어휘로서의 자격이 충분하다. 또한 그 수가 약 500개 정도이므로 설문조사가 가능한 수준이다. 설문조사는 단어를 제시하고, 그 단어의 감정 정도값을 기술하게 하는 간단한 방법으로 이루어진다.

그러므로 감정어휘 정도값 계산의 초점은 확장형 감정어휘의 정도값을 어떻게 유추할 것인가에 있다. 본 연구에서는 설문조사된 기본형 감정어휘의 정도값을 기초로 gloss내 감정어휘 사이의 연산을 통해 확장형 감정어휘의 정도값을 계산하는 방법을 사용한다. 예를 들어, 확장형 감정어휘 '더럽다'의 감정 정도값은 그 어휘의 gloss '못마땅하거나 불쾌하다'에 사용된 기본형 감정어휘 '못마땅하다'와 '불쾌하다'의 정도값을 기초로 얻는 것이다.

4.2 Glosses 문형 계산 방법

Glosses에 사용된 감정어휘를 통하여 표제어의 감정 정도값을 얻기 위해서는 각 glosses가 가지고 있는 주요 문형을 파악하고 이들에 대한 계산 방법을 마련해야 한다. 어떤 문형이 사용되었느냐에 따라서 같은 단어가 사용이 되어도 그 의미가 달라지기 때문이다.

사전은 표제어를 가급적 쉽게 기술하는 데 목적이 있으므로 glosses의 문장 구조는 다양하거나 복잡하지 않다. 사전의 glosses에 사용되고 있는 주요 문형은 대체로 다음의 네 가지로 정리할 수 있다.

〈Table 3〉 Patterns of Operators

Patterns	Operator	Examples
~Go ~Hada (~**고** ~하다)	AND	SilEoHaGo MiWeoHaDa (싫어하**고** 미워하다)
~GeoNa ~HaDa (~**거나** ~하다)	OR	SeonEul NaeGeoNa HeungBunHaDa (성을 내**거나** 흥분하다)
~Ge ~HaDa (~**게** ~하다)	Multiply	MopSi BunHaDa (**몹시** 분하다)
~Ji ANiHaDa (~**지** 아니하다)	NOT	KiPpuJi ANiHaDa (기쁘**지** 아니하다)

문형의 계산 방법을 검증하기 위해서는 문형에 사용된 감정어휘의 정도값을 알고 있어야 하는데, 분노 감정어휘의 정도값에 대하여는 설문 조사된 바가 있으므로[5] 여기서는 분노 감정을 중심으로 문형 계산 방법을 알아본다.

먼저, 'AND' 연산자를 가지고 있는 문형에 대한 계산 방법으로 gloss 내 각 감정어휘의 정도값을 더하는 것을 생각해 볼 수 있다. 표제어 '혐오하다'는 '싫어하고 미워하다'라는 gloss를 가지니 '혐오하다'의 정도값은 '싫어하다'와 '미워하다'의 정도값을 더한 값으로 생각해 보는 것이다.

그러나 이 계산 방법은 적절한 결과를 낳지 않는다. 감정 정도값이 알려져 있는 위의 세 단어는 각각 7.4, 4.6, 4.9를 가져 '싫어하다, 미워하다' 두 단어의 정도값 합 9.5는 '혐오하다'의 정도값 7.4를 훨씬 뛰어넘는 것이다.

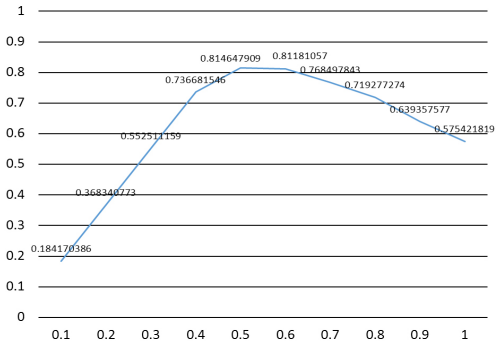
혐오하다 ≠ 싫어하다 + 미워하다

7.4 ≠ 4.6 + 4.9

이것은 문형에서 'AND'가 가지고 있는 의미가 단순히 둘 이상의 어휘 정도값을 더하라는 것이 아님을 의미한다. 만약 단순히 더하게 되면 위의 경우 '싫어하다'와 '미워하다'가 공통적으로 갖는 의미성분이 두 번 계산되므로 전체 합은 한 어휘가 가질 수 있는 정도값을 훨씬 뛰어넘을 수도 있는 것이다. 그러므로 gloss에 사용된 어휘에 대하여 다음과 같이 가중치 W를 주어 적절한 결과값을 가질 수 있게 해야 한다.

혐오하다 = (싫어하다 + 미워하다) × W

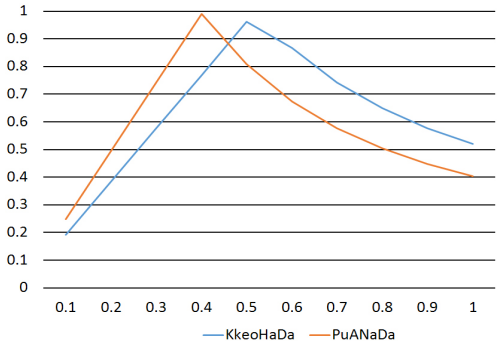
적절한 가중치를 찾기 위하여 정도값이 이미 조사된 분노 감정어휘 표제어에서 gloss에 AND 문형을 가지고 있는 경우를 살펴보니 '분하다, 혐오하다, 신물나다, 괘씸하다'의 4개 어휘가 있었다. 이들에 대하여 0.1에서 1.0까지 가중치를 주어 연산의 정확도를 살펴보니 다음과 같이 가중치 0.5 또는 0.6일 때 가장 높은 정확도 81%를 보였다.



〈Figure 2〉 AND Weight Accuracy

AND의 가중치가 0.5~0.6이라는 것은 이 연산자가 쓰이는 문장은 각 어휘가 가지는 감정 정도값을 약 절반 수준씩만 가지고 와서 결합한다는 의미가 된다.

이번에는 'OR' 연산자를 가지고 있는 문형에 대한 계산 방법을 생각해보자. OR 연산자는 둘 이상의 것 중 어느 하나를 뜻하므로 이 뜻을 그대로 따르면 gloss에 사용된 어휘 중 어느 하나의 정도값을 사용해도 될 것이다. 그러나 어느 하나의 어휘만을 사용하지 않고 복수의 어휘를 사용했다는 것은 어느 하나만으로는 의미를 온전히 살릴 수 없으니 그 어휘들의 평균값을 구하라는 뜻으로 해석된다. 이는 가중치 정확도를 구해보면 확인된다.

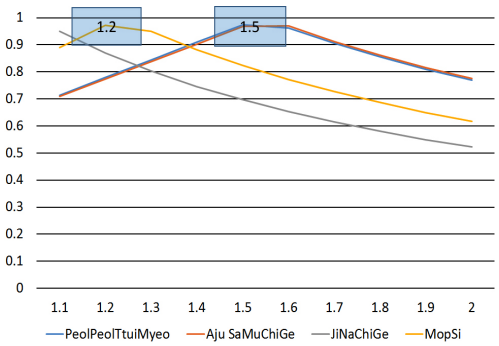


〈Figure 3〉 OR Weight Accuracy

검증이 가능한 OR 연산자가 사용된 표제어는 본 연구에서 '격하다, 부아나다' 두 개여서 0.1~1.0까지의 각 어휘 가중치를 정확도를 어휘마다 별개의 그래프로 나타내었다. 그 결과 '격하다'는 0.4에서 99%, '부아나다'는 0.5에서 96%의 정확도를 보였다.

결국 AND 연산자와 OR 연산자는 뜻하는 의미는 다르지만, 결과는 거의 같다. AND 연산자는 정도값을 1/n 수준씩 가져와 결합하라는 것이고, OR 연산자는 연산자의 의미에 평균값을 구하라는 의미이니 같은 결과를 낳는다.

다음으로 'Multiply' 연산자가 사용된 문형의 계산 방법에 대해서 알아보자. 검증이 가능한 Multiply 연산자가 사용된 표제어는 본 연구에서 '노발대발하다, 증오하다, 질투하다, 분개하다'의 네 가지가 있었는데 이들은 glosses에 각각 '펄펄 뛰며, 아주 사무치게, 지나치게, 몹시'의 부사어를 가지고 있었다. 사용된 부사어에 따라 별개의 그래프로 나타내어 보면 다음과 같다.



〈Figure 4〉 Multiply Weight Accuracy

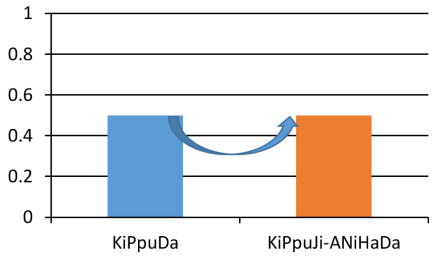
Multiply 가중치 정확도 검사에서 흥미로운 점은 부사어들이 가중치 1.2 수준에서 가장 높은 정확도를 보이는 경우와 1.5 수준에서 가장 높은 정확도를 보이는 경우로 나뉜다는 점이다. '지나치게'와 '몹시'는 1.2에서 95%의 정확도를, '펄펄 뛰며'와 '아주 사무치게'는 1.5 수준에서 97%의 정확도를 보이고 있다.

이것은 정도를 표현하는 정도 부사어마다 갖는 정도값이 서로 다르기 때문이다. 전자인 '지나치게'와 '몹시'도 피수식어의 정도값을 증폭시키기는 하지만 그 수준이 1.2 즉, 120% 정도인 반면, 후자인 '펄펄 뛰며'와 '아주 사무치게'는 부사어가 갖는 정도값이 더 크기 때문에 150%로 증폭시키는 것으로 해석된다.

마지막으로 NOT 연산자는 감정의 종류를

반대로 바꾸는 것인데 현재는 '분노' 감정의 감정 정도값만 조사되어 있기 때문에 적절한 가중치를 찾는 실험을 할 수 없다. 기본 반대 감정인 '기쁨' 감정을 비롯한 모든 감정에 대하여 기본형 감정어휘의 정도값이 우선 파악되어야 한다. 다만, 여기서는 NOT 연산자가 어떠한 방식으로 가중치를 가지게 될 지에 대해 생각해본다.

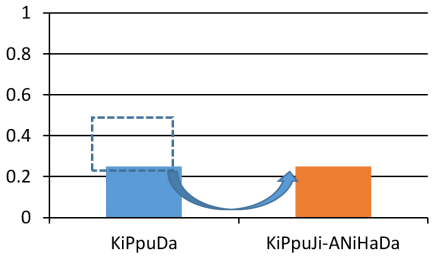
먼저, NOT 연산자의 가중치는 부정되기 전 어휘의 정도값을 유지한 채로 감정의 종류를 역전시킨다고 생각해 볼 수 있다. 예를 들어 '기쁘지 아니하다'라는 표현의 감정 정도값은 기쁨 감정에 속하는 '기쁘다'라는 감정어휘의 정도값을 그대로 가져와 분노 감정의 정도값으로 사용한다는 것이다.



〈Figure 5〉 NOT Weight Concept 1

그러나 NOT 연산자가 쓰였을 때의 감정 정도값이 그대로 유지되지는 않을 것이다. 어떤 사람이 불쾌하거나 언짢을 때 직접 '불쾌하다' 또는 '언짢다'라는 표현을 쓰지 않고 간접적 방법인 기쁨 감정의 어휘를 부정시켜서 표현하는 이유는 분노의 감정 정도가 그만큼 큰 것은 아니기 때문이다. 엄밀히는 '기쁜 것은 아니다'라는 의미를 가질 뿐이다. 다른 예로, '나쁘지 않다'는 것은 '나쁘다'라는 정도가 갖는 만큼 '좋다'는 것이 아니라, '보통이다' 정도의 의미를 가진다.

따라서 NOT 연산자가 사용된 문형은 아래와 같이 해당 어휘의 감정 정도값을 어느 정도 낮춘 뒤 감정 종류를 역전시켜야 한다. 어느 정도 낮출지에 대한 확인은 모든 종류의 기본형 감정어휘에 대하여 설문조사가 이루어져야 가능하다.



〈Figure 6〉 NOT Weight Concept 2

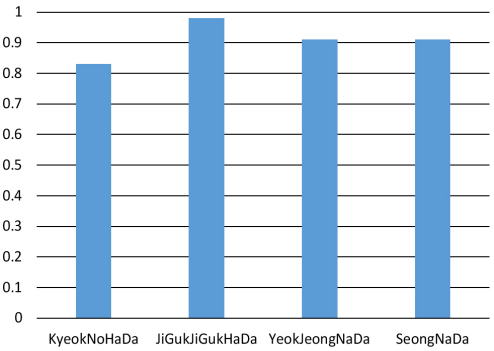
이제까지 논의된 연산자의 가중치를 정리해보면 다음과 같다.

〈Table 4〉 Operators Weight

Operators	Weight
AND	0.5
OR	0.5
Multiply	1.2 or 1.5

이 가중치가 적절한 것인지를 복합문형을 통하여 다시 한 번 검증해 보았다. 여기서 복합문형이란 두 종류 이상의 연산자가 사용된 glosses를 말한다. 부사어가 사용되었으면서 AND 또는 OR 연산자가 사용된 경우가 이에 해당한다. 예를 들어, '격노하다'의 gloss는 '몹시 분하고 노엽다'로서 '몹시'라는 1.2 가중치의 Multiply 연산자와 0.5 가중치의 AND 연산자를 갖는 복합문형이다.

검증이 가능하면서 복합문형을 가지고 있는 표제어는 본 연구에서 '격노하다, 지긋지긋하다, 역정나다, 성나다'의 4개 단어가 있었고 이들의 정확도는 다음과 같았다.



〈Figure 7〉 Verifying through Complex Sentence

이들의 정확도는 최소 83%, 최대 98%였고, 평균적으로는 91%의 정확도를 보여 이제까지의 문형별 가중치 결정 과정이 타당함을 알 수 있다.

4.3 확장형 1단계 감정어휘의 정도값

기본형 감정어휘는 수가 제한적이어서 직접 설문조사를 하여 결정할 수 있지만, 확장형 감정어휘는 수가 많으므로 기본형 감정어휘의 정도값을 토대로 추론하는 방식을 사용한다.

확장형 감정어휘는 표제어가 기본형 감정어휘인 것의 gloss에서 발견되는 확장형 1단계 감정어휘와, gloss에 기본형 감정어휘 또는 확장형 1단계 감정어휘를 보유하는 확장형 2단계 감정어휘의 두 가지가 있다. 이들에 대한 기본적인 추론 방법은 강조승수 m을 두는 것이다. 여기서 m의 의미는 유사 감정어휘가 등장하여 가지는 강조의 정도이다.

예로, 확장형 감정어휘 '속절없다'의 경우를 생각해보자. '속절없다'는 표제어 '환멸하다'의 gloss에 기본형 감정어휘 '괴롭다'와 함께 AND 문형으로 연결되어 나타난다. 기본형 감정어휘는 설문조사를 통하여 정도값을 알고 있으므로 확장형 감정어휘인 '속절없다'의 값을 알아야 하는데 '속절없다'는 확장형으로서 기본형 감정어휘가 아님에도 불구하고 사용되었으니 m만큼 강하게 '환멸하다'의 속성을 설명하고 있다는 것이다.

이를 수식으로 나타내면 다음과 같다. y는 표제어로 사용된 기본형 감정어휘, x_1은 gloss에 사용된 기본형 감정어휘, x_2는 확장형 감정어휘, w는 AND 문형의 가중치이다. 그리고 m은 확장형 감정어휘가 가지는 강조승수이고(단, $m \geq 1$), σ는 강조승수 추정치의 불확실 정도를 의미하는 표준오차이다. 이들은 모두 비음실수(nonnegative real number)이다. 이때 다음과 같은 관계가 성립된다.

$$y = (x_1 + (m \pm \sigma)x_2)w$$

x_2가 기본형 감정어휘일 때는 m의 값은 1이 되어 값에 영향을 주지 않는다. 그러나 확장형 감정어휘일 때는 강조승수 m의 값이 올라가게 된다. 한편 m을 계산하기 위한 표본의 수가 크면 클수록 σ값은 0으로 수렴할 것이다. 일단 본 연구에서는 σ = 0이라고 가정하자.

위의 경우, '환멸하다(6.6)', '괴롭다(5.1)', AND 문형 가중치(0.5)는 알고 있으므로,

$$6.6 = (5.1 + x_2 \cdot m)0.5$$

가 되며, m값을 알면, x_2확장형 감정어휘인 '속절없다'의 정도값을 알 수 있게 된다. 기본형 감정어휘의 m값이 1이므로, 확장형 1단계 감정어휘의 m값을 '2'라고 가정하면, '속절없다'의 감정 정도값은 4.05(4.1)이 된다.

이러한 추론 과정이 적합한지를 검증하기 위하여 복합문형에서 사용된 확장형 감정어휘를 살펴보았다. 확장형 감정어휘 '꺼리다'는 같은 방법으로 5.0의 정도값이 추론되었는데, 이 어휘는 표제어 '질색하다'의 gloss '몹시 싫어하거나 꺼리다'에서 사용되었다. 그리고 다른 어휘들, 즉 '질색하다(5.8), 몹시(1.2), 싫어하다(4.6), OR(0.5)'의 정도값 및 가중치는 기존의 연구와 앞의 과정을 통하여 알고 있으므로 이 값의 정확도를 다음과 같이 알 수 있다.

〈Table 5〉 Inferred Degree of Extended Emotional Vocabulary

Extended emotional vocabulary	Inferred degree of anger emotion
SsuRiDa(쓰리다)	5.1
KkeoRiDa(꺼리다)	5
EokUlHaDa(억울하다)	4.2
DapDapHaDa(답답하다)	4.1
SokJulEopDa(속절없다)	4.1
JinJeolMeoRiNaDa (진절머리나다)	3.7
APhuDa(아프다)	3.6
MipSalSuRupDa(밉살스럽다)	3.5
ThatHaDa(탓하다)	3.4
HanThanHaDa(한탄하다)	2.7
(GamJeongI)IlEoNaDa ((감정이)일어나다)	2.5
(HimI)PpaJiDa ((힘이)빠지다)	2.5
(MaEumE)DulJi-AnDa ((마음에)들지 않다)	2.1
JoChi-AnDa(좋지 않다)	2.1
YakPpaRuDa(약빠르다)	2
JilRiDa(질리다)	1.8
DalGapJi-AnDa (달갑지 않다)	1.75
SiDulHaDa(시들하다)	1.75
KKeoRimChikHaDa (꺼림칙하다)	1.5
GeoBukHaDa(거북하다)	1.15
TtaBunHaDa(따분하다)	1.1

질색하다 = 몹시 싫어하거나 꺼리다

5.8 ≒ 1.2×((4.6+5.0)×0.5) = 5.76(99%)

Gloss 어휘의 계산값은 5.76으로 설문조사로 얻어진 '질색하다'의 정도값 5.8과 99% 일치한다. 따라서 이 추론 방식은 유효하다고 하겠다. <Table 5>에서 찾은 21개의 확장형 1단계 감정어휘의 분노 감정 정도값을 이 방식으로 추론해 보면 다음과 같다.

4.4 확장형 2단계 감정어휘의 정도값

확장형 2단계 감정어휘의 정도값 추론도 기본적으로 1단계의 경우와 같다. 다만 확장형 감정어휘를 gloss에서 찾는 것이 아니라 표제어에서 찾으며, 층위가 1층위는 물론 2층위까지 있다는 점이 다르다.

확장형 2단계 1층위 감정어휘의 예로는 '앓다'가 있다. '앓다[1]'은 '병에 걸려 고통을 겪다', '앓다[2]'는 '마음에 근심이 있어 괴로움을 느끼다'로 모두 gloss에 기본형 감정어휘를 가진다. '고통을 겪다(고통스럽다)'의 감정 정도값은 6.1, '괴로움을 느끼다(괴롭다)'의 감정 정도값은 5.1이므로 강조승수 m = 2를 반영하면 두 다의어의 정도값은 각각 3.1, 2.6이 된다. 사전에서 제시된 예문의 어휘에 따

라 자동적으로 다의어 구분을 할 수 있으나, 구분이 어려울 경우 둘의 평균값인 2.9를 사용한다. DB에는 다음과 같이 정리된다.

〈Table 6〉 Examples of Writing 2 Stratum 1 Level Polysemy

Type	AlTha(앓다)	
	Polysemy 1	Polysemy 2
Emotion	Anger	Anger
Basic/Extended	Extended	Extended
Stratum	2	2
Level	1	1
Emotional Degree	3.1	2.6
Examples' words	bae, I, gamgi, momsal(배, 이, 감기, 몸살)	sok, gotong (속, 고통)

확장형 2단계 2층위 감정어휘의 경우에는 확장형 1·2단계의 1층위 감정어휘보다 감정어휘로서의 비전형성이 더욱 크므로 강조승수 m은 더욱 클 것이다. 기본형 감정어휘와 확장형 1층위 감정어휘의 강조승수 값 연속성을 고려하여 이 경우의 m값을 '3'으로 가정한다.

'궁핍하다'의 경우를 들어보면, '궁핍하다'는 '몹시 가난하다'라는 gloss를 갖고, '가난하다'의 gloss는 '살림살이가 넉넉하지 못하여 몸과 마음이 괴로운 상태에 있다'로서 '괴롭다'라는 기본형 감정어휘를 포함한다. 따라서 '궁핍하다' x의 분노 감정 정도는 다음과 같이 계산된다.

$$1.2 \times 5.1 = x \cdot m$$

m = 3이므로 '궁핍하다'는 2.04(2.0)의 분노 감정 정도값을 얻게 된다.

4.5 문장 감정 정도값 연산

마지막으로 입력문의 계산 방법에 대해서 생각해본다. 입력문의 감정 정도값은 개별 감정어휘의 정도값과 문형의 계산 방법을 결합하여 얻는다. 예를 들어, '방이 청소가 되어 있지 않아 몹시 짜증나고 불쾌하였다'라는 문장이 입력되면 감정어휘 '짜증나다'와 '불쾌하다'의 정도값이 각각 6.2, 5.8이고, '몹시'가 갖는 Multiply 가중치 1.2, AND 문형의 가중치 0.5이므로 전체 문장의 감정 정도값은 다음과 같이 구해질 수 있다.

$$(1.2\times(6.2+5.8))/2 = 7.2$$

이러한 방법으로 감정어휘가 들어간 각 입력문 문장에 대하여 정도값을 계산할 수 있다. 하나의 글에는 여러 개의 문장이 있으므로 감정이 나타난 문장들의 감정 종류, 해당 감정의 최대값, 최소값 및 평균값을 제시하면 글쓴이의 감정 상태를 확인할 수 있다.

5. 결 론

본고는 한국어 입력문에서 감정 정보를 얻을 때 사용될 한국어 감정어휘사전인 SentiWordNet의 구축 방법에 대하여 논의하였다. SentiWordNet의 두 가지 필수 요소는 감정어휘 목록과 개별 감정어휘마다의 정도값인데 이 두 가지를 확보하는 것은 쉬운 일은 아니다. 먼저 전형적인 감정어휘인 기본형 감정어휘는 기존의 연구를 통하여 얻을 수 있지만, 그 수가 적어 이것만으로는 현실적 요구를 감당하는 응용 프로그램 및 서비스를 구

현할 수 없다. 따라서 감정어휘의 속성을 부분적으로라도 가지고 있는 어휘들을 찾아 전체 감정어휘 목록에 포함시킬 방안을 강구해야 한다. 두 번째 어려운 점은 개별 감정어휘의 정도값을 얻는 일이다. 기본형 감정어휘는 그 수가 적으므로 직접 설문조사를 하여 정도값을 얻을 수 있지만, 추가로 확보된 감정어휘까지 모두 설문조사를 수행할 수는 없기 때문이다.

이에 본고는 감정어휘의 속성을 잘 가지고 있어 기존 연구에서 파악된 것을 기본형 감정어휘로, 감정어휘의 속성을 일부 가지고 있어 추가로 확보되는 것을 확장형 감정어휘로 분류하고, 기본형 감정어휘는 직접적인 방법으로, 확장형 감정어휘는 사전의 gloss를 층위별로 이용하는 간접적인 방법을 사용하였다.

따라서 감정어휘는 총 네 가지 종류로 나뉜다. 1) 기본형 감정어휘는 기존의 연구에서 약 500개의 목록이 확보되었으므로 이를 토대로 목록을 정하고, 그 정도값은 직접 설문조사를 통하여 얻는다. 확장형 감정어휘는 세 종류로 나뉘는데 2) 확장형 1단계 1층위 감정어휘는 기본형 감정어휘를 표제어로 하는 gloss에 사용된 어휘에서 얻는다. 3) 확장형 2단계 1층위 감정어휘는 비 감정어휘 표제어 중 gloss에 기본형 감정어휘나 확장형 1단계 감정어휘가 있는 경우이다. 4) 확장형 2단계 2층위 감정어휘는 Nested Gloss에서 기본형 또는 이제까지의 확장형 감정어휘가 발견된 경우이다.

확장형 감정어휘의 정도값은 기본형 감정어휘의 정도값을 기초로 문형의 가중치와 강조승수를 적용하여 얻는다. 실험 결과 AND, OR 문형은 내포된 어휘의 감정 정도값을 평균 내는 가중치를, Multiply 문형은 정도 부사어에 따라 1.2~1.5의 가중치를 갖는 것으로 파악되었다. NOT 문형은 추가 연구가 필요하나 사용 어휘의 감정 정도를 낮추어 역전시키는 것으로 추정된다. 또한 확장형 감정어휘에 적용되는 강조승수는 1층위에서 2, 2층위에서 3을 갖는 것으로 예상된다.

입력문의 감정 정도값 계산은 개별 감정어휘의 정도값과 사용된 입력문의 문형 가중치를 적용하여 얻는다. 하나의 주제에 대하여 여러 문장이 입력될 수 있는데 감정어휘가 사용된 문장을 선별하여 전체 문장에서 최대값, 최소값, 평균값을 구하면 사용자의 감정정보로 제공할 수 있다.

본 논문에서는 분노 감정어휘의 경우를 예시로 한국어 SentiWordNet의 구현 방안을 모색하였다. 향후 전체 감정어휘로 범위를 확장하여 한국어 SentiWordNet을 구현할 계획이다.

References

[1] Abbasi, A., Chen, H., Thome, S., and Fu, T., “Affect Analysis of Web forums and Blogs Using Correlation Ensembles,” IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, pp. 1168-1180, 2008.

[2] Baccianella, S., Esuli, A., and Sebastiani, F., “SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” In Proceedings of the 7th Conference on International Language Resources and Evaluation(LREC

'10), pp. 2200-2204, 2010.

[3] Biswas, S., Yoo, J. H., and Jung, C. Y., "A Study on Priorities of the Components of Big Data Information Security Service by AHP," Journal of Society for e-Business Studies, Vol. 18, No. 4, pp. 301-314, 2013.

[4] Choi, S. J., "The Type and Character of Feeling Verb," EoMunNonJip, Vol. 58, pp. 127-159, 2008.

[5] Choi, S. J., "The level of Feeling Verb : in the case of Anger words," Lingua Humanitatis, Vol. 11, No. 2, pp. 273-295, 2009.

[6] Collins Cobuild Advanced Learner's English Dictionary, 6th Edition, Harper Collins Publishers, 2009.

[7] Dehkharghani, R., Yanikoglu, B. D., and Tapucu, Y., "Adaptation and Use of Subjectivity Lexicons for Domain Dependent Sentiment Classification," IEEE 12th International Conference on Data Mining Workshops(ICDMW), pp. 669-673, 2012.

[8] Esuli, A. and Sebastiani, F., "Determining the Semantic Orientation of Terms through Gloss Classification," In Proceedings of 14th ACM International conference on Information and knowledge management, pp. 617-624, 2005.

[9] Esuli, A. and Sebastiani, F., "Determining Term Subjectivity and Term Orientation for Opinion Mining," In Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 193-200, 2006.

[10] Esuli, A. and Sebastiani, F., "SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining," In Proceedings of the 5th Conference on Language Resources and Evaluation(LREC'06), pp. 417-422, 2006.

[11] Esuli, A. and Sebastiani, F., "Random-Walk Models of Term Semantics : An Application to Opinion-Related Properties," In Proceedings of the 3rd language Technology Conference(LTI '07), pp. 221-225, 2007.

[12] Gim, E. Y., "A Study on the Korean Emotion Verbs," PhD thesis, Chonnam National University, 2004.

[13] Hamouda, A. and Rohaim, M., "Reviews Classification Using SentiWordNet Lexicon," The Online Journal on Computer Science and information Technology(OJCSIT), Vol. 2, No. 1, pp. 120-123, 2011.

[14] Hatzivassiloglou, V. and Katheleen R. M., "Predicting the Semantic Orientation of Adjectives," In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, pp. 174-181, 1997.

[15] Hwang, J. W. and Ko, Y. J., "A Korean Sentence and Document Sentiment Classification System Using Sentiment Features," Journal of KISS : computing practices, Vol. 14, No. 3, pp. 336-340, 2008.

[16] Kamps, J., Marx, M., Mokken, R. J., and

Rijke, M. D., “Using WordNet to Measure Semantic Orientation of Adjectives,” In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, Vol. Ⅳ, pp. 1115-1118, 2004.

[17] Lyons, W., Emotion, Cambridge University Press, London, 1980.

[18] Ohana, B. and Tierney, B., “Sentiment Classification of Reviews Using SentiWordNet,” Proceedings of the 9th IT&T Conference, 2009.

[19] Rao, D., Lewis, S., and Reichenbach, C., “Automatic Opinion Poloarity Classification of Movie Reviews,” Colorado Research in Linguistics, Vol. 17, No. 1, 2004.

[20] Roh, J. H., Kim, H. J., and Chang, J. Y., “Improving Hypertext Classification Systems through WordNet-based Feature Abstraction,” Journal of Society for e-Business Studies, Vol. 18, No. 2, pp. 95-110, 2013.

[21] Rohracher, H., Einführung in die psychologie, Urban und Schwarzenberg, München, Berlin, Wien, 1976(윤흥섭 역. 심리학 개론, 성원사, 1990).

[22] Shaver, P., Schwarth, J., Kirson, D., and O’Connor, C., “Emotion Knowledge : Further Exploration of a Prototype Approach,” Journal of Personality and Social Psychology, Vol. 52, No. 6, pp. 1061-1086, 1987.

[23] Su, Q., Xiang, Kun., Wang, H., Sun, B., and Yu, S., “Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews,” International Conference on the Computer Processing of Oriental Languages, pp. 22-30, 2006.

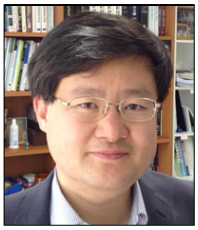
[24] Turney, P. D. and Littman, M. T., “Measuring Praise and Criticism : Inference of Semantic Orientation from Association,” ACM Transactions on Information Systems, Vol. 21, No. 4, pp. 315-346, 2003.

[25] Yeon, J., Shim, J., and Lee, S. G., “Outlier Detection Techniques for Biased Opinion Discovery,” Journal of Society for e-Business Studies, Vol. 18, No. 4, pp. 315-326, 2013.

[26] Yoon, A. S. and Kwon, H. C., “Compononet Analysis for Constructing an Emotion Ontology,” Korean Journal of Cognitive Science, Vol. 21, No. 1, pp. 157-175, 2010.

저 자 소 개

최석재	(E-mail : sjchoi@khu.ac.kr)
1999년	고려대학교 국어국문학과 (학사)
2001년	고려대학교 대학원 국어국문학과 (석사)
2008년	고려대학교 대학원 국어국문학과 (박사)
2001년~2002년	카네기멜론 대학 전산학부 방문연구원
2003년~2005년	연변과학기술대학 언어공학연구소 실장
2008년~2010년	고려대학교 BK21 연구교수
2011년~2014년	성신여자대학교 국어국문학과 초빙교수
2014년~현재	경희대학교 경영대학 학술연구교수
관심분야	한국어 정보화, 빅데이터분석, 자연언어처리



권오병	(E-mail : obkwon@khu.ac.kr)
1988년	서울대학교 경영학과 (경영학사)
1990년	한국과학기술원 경영과학과 (공학석사)
1995년	한국과학기술원 경영과학과 (공학박사)
2001년~2002년	카네기멜론대학 전산학부 방문연구원
2009년~2011년	샌디에고주립대학 경영정보학과 방문교수
2004년~현재	경희대학교 경영대학 교수
관심분야	빅데이터분석, 사물인터넷, 소셜미디어, 유비쿼터스 컴퓨팅