공학석사 학위논문

영어 감정사전의 감정 점수 전파를 통한 한국어 감정사전 제작

Developing a Korean sentiment lexicon through sentiment score propagation of English sentiment lexicon

지도교수 김재훈

2019 년 2 월

한국해양대학교 대학원

컴퓨터공학과
박 호 민

본 논문을 박호민의 공학석사 학위논문으로 인준함.

위 원 장 박 휴 찬 (인)

위 원 이 장 세 (인)

지도교수 김 재 훈 (인)

2019년 01월 03일

한국해양대학교 대학원

목 차

List of Tables ····· iv
List of Figures ····· vii
Abstract ····· iv
초록 ····· vi

제 1 장 서 론 ····· 1

제 2 장 관련 연구 ····· 4
2.1 감정분석 ····· 4
2.1.1 데이터 수집 ····· 4
2.1.2 주관성 탐지 ····· 5
2.1.3 극성 탐지 ····· 6
2.2 감정사전 ····· 7
2.2.1 사전 기반 감정사전 ····· 7
2.2.2 말뭉치 기반 감정사전 ····· 9
2.2.3 집단지성 기반 감정사전 ····· 12
2.3 VADER 감정사전 ····· 14

제 3 장 감정 점수 전파를 통한 감정사전 제작 ····· 18
3.1 한영 이중언어사전 제작 ····· 19
3.1.1 한영 병렬 말뭉치 토큰화 ····· 19
3.1.2 상호정보량 행렬 제작 ····· 20
3.1.3 코사인 유사도를 통한 이중언어사전 제작 ····· 24
3.2 한국어 fastText 표상 모델 제작 ····· 26
3.3 한영 이중언어그래프 제작 ····· 27

3.4 감정 점수 전파 ····· 31

제 4 장 실험 및 평가 ····· 37

4.1 제작 과정의 발견법적(heuristic) 접근의 검증 ····· 37

4.2 제작된 감정사전의 검증 ····· 38

4.2.1 감정분석 시스템 ····· 39

4.2.2 감정분석 시스템을 활용한 감정 말뭉치 감정분석 ····· 41

제 5 장 결론 및 향후 연구 ····· 45

참고문헌 ····· 47

감사의 글 ····· 55

List of Tables

Table 3.1 Statistics of the Korean-English parallel corpus ····· 19
Table 3.2 Statistics of the VADER sentiment lexicon ····· 21
Table 3.3 Derived words from 'cute' with various suffixes in the VADER sentiment lexicon ····· 22
Table 3.4 Statistics of a Korean-English bilingual lexicon ····· 25
Table 3.5 Statistics of Korean corpus by POS ····· 26
Table 3.6 Statistics of a Korean-English bilingual graph ····· 31
Table 3.7 Statistics of the Korean sentiment lexicon by polarity and POS ·· 36
Table 3.8 Statistics of the Korean sentiment lexicon by strength of value ··· 36
Table 4.1 The comparison of correction rate of a Korean-English bilingual lexicon based on extraction limit of cosine similarity ····· 38
Table 4.2 Statistics of KMU sentiment corpus and NAVER sentiment movie corpus ····· 38
Table 4.3 The correct answer range of a sentiment score of each polarity ·· 42
Table 4.4 The table of a evaluation result ····· 42
Table 4.5 The confusion matrix ····· 42
Table 4.6 The accuracy result of KMU sentiment corpus based on a extraction limit of cosine similarity ····· 43
Table 4.7 The precision, recall and result of NAVER sentiment movie corpus based on a extraction limit of cosine similarity ····· 43
Table 4.8 The accuracy result of KMU sentiment corpus based on addition of language conventions ····· 44
Table 4.9 The precision, recall and result of NAVER sentiment movie corpus based on addition of language conventions ····· 44

List of Figures

Figure 2.1 The processing of creating a sentiment lexicon using a lexicon-based method ····· 8
Figure 2.2 The processing of creating a sentiment lexicon using a corpus-based method ····· 10
Figure 2.3 The processing of creating a sentiment lexicon using a collective intelligence-based method ····· 13
Figure 2.4 The processing of creating the VADER sentiment lexicon ····· 16
Figure 3.1 The processing of creating a Korean sentiment lexicon by the proposed method ····· 19
Figure 3.2 An example of tokenization of Korean and English sentences ···· 20
Figure 3.3 The processing of creating PMI matrices ····· 24
Figure 3.4 Examples of two generated PMI matrices ····· 24
Figure 3.5 The extraction process of Korean morphemes with the top 10 cosine similarity of 'cute' ····· 25
Figure 3.6 An example of fastText embedding of Korean morphemes and English words ····· 27
Figure 3.7 The processing of creating a Korean-English bilingual graph ····· 29
Figure 3.8 Extracting words and morphemes for vertex of a Korean-English bilingual graph ····· 30
Figure 3.9 Adding edges in a Korean-English bilingual graph ····· 30
Figure 3.10 Initializing weights on edges in a Korean-English bilingual graph ····· 31
Figure 3.11 A flow diagram for the label propagation algorithm on a Korean-English bilingual graph ····· 33
Figure 3.12 The initialization step of the label propagation algorithm on a Korean-English bilingual graph ····· 34

Figure 3.13 After first propagation step on a Korean-English bilingual graph ····· 35
Figure 3.14 After second propagation step on a Korean-English bilingual graph ····· 35
Figure 4.1 The processing of sentiment scoring about text data in the VADER sentiment analysis system ····· 39
Figure 4.2 The processing of sentiment scoring about documents in the proposed sentiment analysis system ····· 41

Developing a Korean sentiment lexicon through sentiment score propagation of English sentiment lexicon

Park, Ho-Min

Department of Computer Engineering
Graduate School of Korea Maritime and Ocean University

Abstract

Nowadays, people express their personal feelings and opinions on social media, and such the posts or reviews are frequently used as the data for the sentiment analysis to order to identify public opinions, market trends, and so on. Sentiment analysis is the automated process of understanding an attitudes and opinion about a given topic from written or spoken text. One of the sentiment analysis approaches is a dictionary-based approach, in which a sentiment dictionary plays an important role. However, many posts on the social media cannot be analyzed by dictionary-based approach due to the absence of sentiment words in the dictionary. Therefore the sentiment dictionary should be expanded or built in totally new domains.

In this paper, we propose a method to automatically create a Korean sentiment lexicon from the verified English sentiment lexicon called VADER sentiment lexicon. The proposed method consists of three steps. The first step is to produce a Korean-English bilingual lexicon using the Korean-English

parallel corpus. The bilingual lexicon is a set of pairs between VADER sentiment words and Korean morphemes. The second step is to generate a bilingual graph using the bilingual lexicon. The vertex on the graph is a word (VADER sentiment words or Korean morphemes), and the edge is a pair of words, which are in the bilingual lexicon or belongs to synonyms for the same language. The third step is to run the label propagation algorithm throughout the bilingual graph. Finally a new Korean sentiment lexicon is created by repeatedly applying the propagation algorithm until the values of all vertices converge.

To validate the sentiment lexicon generated by the proposed method, we made a dictionary-based Korean sentiment classifier with some heuristic rules, which is quite similar to the VADER sentiment classifier in English, but most of its rules have been specially adapted to suit Korean characteristics. The resources used for evaluating the classifier are two Korean sentiment corpus: news article and movie review. The accuracy of 81% and the F-score of 72% for the news article corpus and the movie review corpus are achieved, respectively. Through the evaluation, we have observed that the proposed method is pretty good and very effective. In the future, we will have more experiments for comparing the performance of various approaches like a machine learning-based approach, a deep learning-based approach, and so on.

KEY WORDS: Sentiment lexicon, Sentiment analysis, PMI (Point-wise mutual information), Word embedding, Cosine similarity, label propagation

영어 감정사전의 감정 점수 전파를 통한 한국어 감정사전 제작

박호민

컴퓨터공학과
한국해양대학교 대학원

초록

요즘 사람들은 자신의 개인적인 감정과 의견을 표현하기 위해 소셜 네트워크 서비스를 주로 이용한다. 따라서 여론 조사나 시장 동향 등을 파악하기 위해 감정분석을 위한 데이터로 자주 사용된다. 감정분석은 문서 또는 대화 상에서 주어진 주제에 대한 태도와 의견을 이해하는 자동화된 프로세스이다. 감정분석의 다양한 접근법 중 하나는 감정사전을 이용하는 사전기반 접근법이다. 그러나 소셜 네트워크 서비스에서의 많은 게시물들에는 감정사전에 존재하지 않는 단어가 많아 사전기반 방식으로 분석하기 어렵다. 따라서 감정분석을 효과적으로 수행하기 위하여, 감정사전의 확장 또는 새로운 감정사전 제작이 요구된다.

본 논문에서는 검증된 영어 감정사전인 VADER의 감정사전을 활용하여

한국어 감정사전을 자동으로 생성하는 방법을 제안한다. 제안하는 방법은 세 단계로 구성된다. 첫 번째 단계는 한영 병렬 말뭉치를 사용하여 한영 이중언어사전을 제작한다. 이중언어사전은 VADER 감정어와 한국어 형태소 쌍들의 집합이다. 두 번째 단계는 이중언어사전을 사용하여 이중언어 그래프를 생성한다. 그래프의 정점은 VADER 감정어와 한국어 형태소를 사용하고, 간선 연결은 이중언어사전 및 동일 언어의 동의어 순으로 구성된다. 세 번째 단계는 이중언어그래프 상에서 레이블 전파 알고리즘을 실행한다. 그래프 상의 모든 정점들의 값이 수렴될 때까지 레이블 전파 알고리즘을 반복적으로 적용하여 끝으로 새로운 감정사전이 제작된다.

제안하는 방법으로 제작된 감정사전을 검증하기 위하여 사전기반의 한국어 감정분석 시스템을 구축하였다. VADER 감정분석 시스템에서의 발견법적 접근을 한국어의 특성에 맞춰 변화하여 적용시켰다. 평가 자료로는 뉴스 기사의 댓글을 모아놓은 KMU 감정 말뭉치, 영화평을 모아놓은 네이버 감정 영화 말뭉치 두 개를 사용하였다. 평가 결과, KMU 감정 말뭉치에서는 81%의 정확도를 보였으며 네이버 감정 영화 말뭉치에서는 72%의 $F_1 Score$를 달성하였다. 이와 같은 결과를 통해 제안하는 방법이 새로운 감정사전 제작과 감정분석에 있어서 효과적임을 알 수 있다. 향후에는 기계학습, 심층학습을 적용하여 연구를 진행할 예정이다.

KEY WORDS: 감정사전, 감정분석, PMI, 단어 표상, 코사인 유사도, 레이블 전파

제 1 장 서 론

발전된 정보통신기술(information communication technology)과 대량으로 보급된 스마트폰(smartphone)같은 통신기기의 보급으로 사람들은 시공간을 뛰어넘어 데이터와 정보를 생산하고 공유한다. 대표적인 소셜 네트워크 서비스(social network service) 사이트인 페이스북(Facebook)에서는 1분당 24만개 이상의 사진이 올라오고, 트위터(Tweeter)에서는 1분당 35만개 이상의 트윗(tweet)이 게시된다. 또한 유튜브(Youtube)에는 1분당 총합 400시간의 새로운 동영상들이 올라오며 동시에 70만 시간의 동영상들이 소비된다[1]. 바야흐로 빅데이터(big data)의 시대라고 할 수 있다.

이러한 빅데이터의 시대에서 가장 주요한 관심사는 넘쳐나는 데이터 속에서 의미 있는 정보를 추출하는 것이다. 단순히 빅데이터로는 그저 거대한 데이터의 집합일 뿐 정보는 되지 못한다. 의미 있고 가치 있는 정보를 빅데이터에서 가려내야 실질적으로 활용하고자 하는 분야에서 사용할 수 있기 때문이다. 따라서 빅데이터 못지않게 그 빅데이터를 정제하고 분석하는 기술 역시 중요하다. 본 논문에서는 여러 정보 중에서 감정(sentiment) 정보를 다루며 그것을 분석하는 기술인 감정분석(sentiment analysis)에 필수적인 감정사전 제작에 대해 다룬다.

감정분석은 텍스트에서 나타난 저자 혹은 발화자의 태도, 의견 등과 같은 주관적인 정보를 추출하는 기술이다. 사용자의 진심이나 숨겨진 의도를 정확히 파악하기 위함으로 여론 분석이나 시장 동향 파악 등 다양한 분야에 두루 사용된다. 사용자의 감정을 제대로 파악할 수 있다면 그에 대한 알맞은 대응을 할 수 있으며 동시에 사용자의 시스템에 대한 만족도

1) http://www.go-globe.com/blog/things-that-happen-every-60-seconds, 2017

를 높일 수 있기 때문이다.

영어의 감정분석에는 크게 두 가지로, 사전 기반 방법과 기계학습 기반 방법이 있다(Gilbert & Hutto, 2014). 본 논문은 사전 기반 방법에 관련한다. 사전 기반 방법은 감정사전(sentiment lexicon)을 이용하여 극성(polarity) 혹은 점수를 단어에 부여하는 방식이다. 단순하게는 긍정, 부정의 극성이나 점수를 각 단어마다 부여하거나(Gilbert & Hutto, 2014; Nielsen 2011; Hansen, *et al.*, 2011; Thelwall, *et al.*, 2010) 각성(arousal)과 지배(dominance)의 심리학적인 측면을 추가하여 다각적인 관점에서 감정을 분석하기도 한다(Bradley & Lang, 1999; Warriner, *et al.*, 2013).

이러한 감정사전을 제작하는 방법은 집단지성을 통한 사전 제작 방법과 자동 생성 방법이 있다(Gilbert & Hutto, 2014). 자동으로 생성하는 방식으로는 초기 사전을 사용하여 동의어(synonym), 반의어(antonym) 등을 추출하여 사전을 확장하거나, 이중언어 사전, 말뭉치 등을 이용해 기계학습을 진행하여 생성한다(허찬, 2017). 집단지성을 통해 제작한 사전은 각 단어별 부여된 극성과 점수 등이 검증되었다는 장점이 있는 반면에 제작하는데 매우 큰 노력과 시간이 든다는 단점이 존재한다. 자동 생성을 통해 제작한 사전은 비교적 쉽고 빠르게 제작할 수 있지만 그에 대한 검증이 부족하다.

본 논문은 영어의 검증된 감정사전으로부터 한국어 감정사전을 자동으로 생성하는 방법을 제안한다. 대표적인 영어의 검증된 감정사전으로는 VADER 감정분석 시스템(Gilbert & Hutto, 2014)에서 사용하는 감정사전이 있다. VADER 감정분석 시스템은 집단지성을 통해 구축한 감정사전과 더불어 통사적인 언어적 규칙을 추가적으로 활용한다. 이를 통해 다른 감정사전들보다 높은 정확도를 보이며 심지어 소셜 네트워크 서비스의 마이크로 블로그 글들에 대해서는 사람보다 높은 수준으로 감정을 분석했다.

본 논문에서는 VADER의 검증된 영어 감정사전을 토대로 한영 병렬 말뭉치 기반의 이중언어 그래프를 제작하고, 레이블 전파(label propagation)

알고리즘(Xiaojin & Zoubin, 2002)을 활용하여 한국어 감정사전을 제작하는 방법을 제안한다. 제안하는 방법은 두 자료 사이의 연관정도를 계산하는 상호정보량(point-wise mutual information)(Church & Hanks, 1990)을 사용하여 한영 병렬 말뭉치에서 한영 이중언어사전(bilingual lexicon)을 생성한다. 이와 같이 제작된 이중언어사전 자료와 Facebook AI research에서 공개한 fastText(Bojanowski, *et al.*, 2017) 모형을 한국어 말뭉치에 적용시켜 한국어 형태소들의 단어표상(word embedding)을 생성한다. 단어표상 자료를 각 형태소들의 코사인 유사도를 그래프 내 간선(edge)값으로 활용하여 그래프를 만든다. 레이블 전파 알고리즘을 통해 그래프 상에서 VADER 감정사전의 점수들을 점수가 매겨지지 않은 한국어 형태소로 전파하여 한국어 감정사전을 제작한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 감정분석 연구의 전반적인 내용과 감정사전 제작과 관련된 기존 연구들을 설명한다. 주요한 범용 감정사전과 영역별 감정사전들, 그리고 각각의 제작 방법에 대해 자세히 설명한다. 3장에서는 본 논문에서 제안하는 병렬 말뭉치를 기반으로 제작된 이중언어그래프 상의 레이블 전파 방법으로 감정사전을 제작하는 방법을 설명한다. 4장에서 간단한 감정분석 시스템을 구현하여 다양한 감정분석 말뭉치를 통하여 제작된 감정사전을 평가하고, 마지막으로 5장에서 결론을 맺고 향후 연구 방향을 제시한다.

제 2 장 관련 연구

2.1 감정분석

오피니언 마이닝이란 “텍스트에 나타난 여론과 의견을 분석하여 유용한 정보로 재가공하는 기술”, “텍스트를 분석하여 사람들의 감정과 의견을 통계화, 수치화하여 객관적인 정보로 바꿀 수 있는 기술”, “사용자들의 생각과 표현의 파편을 모아 일정한 법칙성을 찾아내어 새로운 의견 형성을 발굴하고 탐사하는 방식” 이라고 정의된다(TTA정보통신용어사전, 2012). 그러므로 감정분석은 텍스트에 나타난 감정을 분류하거나 수치화하여 객관적인 정보로 바꾸는 기술을 뜻한다. 좁은 의미로 감정분석은 텍스트 상의 긍정적, 부정적 감정을 분석하는 것이다.

감정분석은 주로 세 단계의 과정으로 이루어진다. 첫 번째는 감정분석에 필요한 텍스트와 그 외의 데이터들을 수집하는 ‘데이터 수집’ 단계이다. 두 번째는 수집된 자료로부터 사람들의 주관적인 의견이나 감정이 드러난 부분만을 걸러내는 ‘주관성 탐지(subjectivity detection)’ 단계이다. 세 번째 단계는 두 번째에서 걸러낸 텍스트를 긍정이나 부정으로 분류하는 ‘극성 탐지(polarity detection)’ 단계이다(신수정, 2014).

2.1.1 데이터 수집

데이터 수집 단계는 감정분석에 필요한 자료를 찾아내고 모으는 단계이다. 자료는 감정문서와 감정어가 있다. 감정문서는 뉴스 기사 댓글이나 마이크로블로그 문서 등이 여기에 속한다. 감정어는 감정분석에 활용할 어구(phrase) 또는 단어이며 이들을 한데 모아 수록하면 감정사전이 된다. 이

러한 감정어들은 텍스트의 감정분석을 위한 단서가 되므로 매우 중요하다. 본 연구에서는 영어의 감정어들을 활용하여 대응되는 한국어 형태소를 수집하여 한국어 감정사전을 제작한다.

2.1.2 주관성 탐지

주관성 탐지는 초기 감정분석의 전처리(pre-processing) 과정으로 텍스트의 내용이 주관적(subjective)인지 객관적(objective)인지 파악하는 것이다. 기본적으로 주관적으로 분류되는 텍스트에 부정적이거나 긍정적인 감정이 표현되어 있다고 생각했기 때문이다. 초기의 주관성 탐지는 수동으로 텍스트의 주관성을 분석하였다(Bruce & Wiebe, 1999). 주관적인 문장은 개인적인 생각이나 평가, 감정 그리고 판단과 같은 정보를 포함하고 있으며, 객관적인 문장은 일반적인 사실이나 정보 전달만을 포함하고 있다. 하지만 실제 수동적 평가 결과, 주관적인 내용의 텍스트가 항상 감정을 포함하는 것이 아니고, 예상과는 달리 객관적인 내용의 텍스트에 감정이 포함되는 경우도 있다(Bruce & Wiebe, 1999). 그러므로 주관성 탐지를 감정분석의 전처리 과정이 아닌 독립적인 문제로 여기기도 한다. 그렇지만 상대적으로 주관적인 문장이 감정을 표현하는 경우가 많기 때문에 여전히 감정분석의 전처리 단계로 주관성 탐지를 수행하는 경우가 많다(김정호, 2015).

기존의 주관성 탐지는 주로 기계학습(machine learning)의 한 방법론인 지도학습(supervised learning)을 이용하였다. 형용사나 대명사 등과 같은 특정 품사의 단어들을 사용하여 문장의 주관성을 분류하였다(Bruce & Wiebe, 1999). 반대로 비지도학습(unsupervised learning)을 이용하여 주관성 탐지를 진행한 연구도 있다(Wiebe, 2000; Lin, 1998). 정답이 존재하지 않기 때문에 명시적으로 초기 감정어를 이용하였다. 해당 단어의 텍스트 내 분포를 이용해 이와 유사한 분포의 특성을 가지는 단어들을 찾아내어 주관성을 판단하였다. 또한 텍스트 내에 주관적이지 않은 문장들을 추려내어 정확도를 높이는 방법도 있다(Hatzivassiloglou & Mckeown, 1997;

Hatzivassiloglou & Wiebe, 2000).

또한 단어를 넘어 구문을 활용해 주관성 탐지를 수행한 연구도 있다(Hatzivassiloglou, *et al.*, 2001). 그리고 초기 단어와 구문을 이용해 텍스트 내 문장 간의 유사도를 측정하여 주관성을 탐지하기도 했다(Yu & Hatzivassiloglou, 2003). 이와 같은 연구들은 초기 단어와 구문의 품질과 양이 정확도와 밀접한 관계를 가지기 때문에, WordNet(George, 1995)과 같은 검증된 자료로부터 유의어를 추출하여 정확도를 높이려고 하였다.

그 외에도 각 문장을 그래프의 정점(vertex)으로 사용하여 정점 간 유사도를 계산하는 방법(Pang & Lee, 2004), 또는 주관적인 문장에서 주로 보이는 문장 구조 패턴을 통해 주관성을 탐지하는 규칙기반(rule-based) 방법(Wiebe & Riloff, 2005), 이모티콘(emoticon)과 같은 다른 요소들을 활용하는 방법(Barbosa & Feng, 2010) 등 주관성 탐지에 있어서 다양한 연구가 수행되어 왔다.

과거의 감정분석은 앞서 설명한 방법들로 텍스트 내의 주관적인 내용들만 찾아내는 전처리 과정을 거친다. 하지만 수동으로 주관성 여부 태깅을 진행했던 연구(Bruce & Wiebe, 1999)의 결론처럼 항상 주관적이어야만 감정을 포함하는 것이 아니므로 이 단계는 생략될 수도 있다(김정호, 2015). 본 연구에서는 주관성 탐지 단계는 생략한다.

2.1.3 극성 탐지

극성 탐지 단계는 텍스트에 담겨있는 긍정 또는 부정의 감정을 발견하여 해당 텍스트에 극성을 부여하는 단계이다. 텍스트의 주관성과 객관성을 탐지하는 이전 단계와 유사하다. 이전 단계가 주관성을 명확하게 띄는 단어와 구문을 활용한 것처럼 긍정 또는 부정의 성질을 명확하게 띄는 단어와 구문을 이용하여 텍스트의 감정을 파악한다.

대표적으로 긍정, 부정 단어의 초기 사전을 활용한다. 초기 감정어와 감정어 후보들과의 유사도를 측정하여 각 감정어 후보 별 긍정, 부정 점수

를 구한 뒤, 문장에 포함된 각 단어들이 가지는 긍정, 부정 점수의 평균으로 문장의 극성을 분류하기도 하였다(Turney, 2002; Yu & Hatzivassiloglou, 2003). 또한 단어들의 긍정, 부정 점수를 WordNet의 유의어와 반의어 관계를 통해 확장하여 텍스트의 감정 극성을 탐지하는데 사용한 연구도 있다(Hu & Liu, 2004).

서포트 벡터 머신(support vector machine, SVM), 신경망(Neural Network), 랜덤 포레스트(random forest), 나이브 베이지안 등과 같은 지도학습 모형을 사용하여 텍스트의 극성을 분류한 연구도 진행되었다(Nigam, *et al.*, 2000; Mullen & Collier, 2004; Ghiassi, *et al.*, 2013; Shah, *et al.*, 2013). 추가적으로 미리 극성이 부여된 텍스트를 사용하여 다른 텍스트들과의 유사도를 계산하여 극성을 분류하는 준지도학습(semi-supervised learning) 방법도 제안되었으며(Gamon, *et al.*, 2005), 문장 구조 패턴과 구문을 활용한 규칙기반의 방법도 사용되었다(Ding, *et al.*, 2008).

텍스트의 극성 탐지는 주관성 탐지와 같이 어떤 단어와 구문을 초기 사전으로 사용할 것인지가 중요하다. 적절한 초기 사전이 준비된다면 앞서 설명한 어떠한 방법론을 적용시켜도 탐지가 잘 이루어질 수 있기 때문이다. 그러므로 감정사전의 중요성이 매우 높아진다. 2.2절에서는 기존의 감정사전 제작에 대한 연구들과 방법론을 자세히 설명한다.

2.2 감정사전

이 절에서는 감정사전과 감정사전을 제작하는 방법에 대한 관련 연구들을 설명한다. 감정사전을 제작하는 방법은 크게 총 세 가지로 나뉘며 사전 기반(lexicon-based) 방법과 말뭉치 기반(corpus-based) 방법, 그리고 집단지성 기반(collective intelligence-based) 방법이다. 각각의 방법들에 대해 소개하고 그에 따른 주요 알고리즘에 관해 기술한다.

2.2.1 사전 기반 감정사전

사전 기반 방법의 특징은 감정사전을 제작하기 위한 감정어와 감정구를 사전에서 추출한다는 것이다. 보편적으로 사전에는 각 단어의 의미와 동의어와 반의어와 같은 단어 사이의 관계 정보가 수록되어 있다. 따라서 내용이 검증된 사전을 보유하고 있다면 비교적 쉽게 다양한 감정어들을 얻어낼 수 있다. Figure 2.1은 사전을 기반으로 감정사전을 제작하는 과정을 보인다. 이후에 설명할 사전 기반의 방법론들은 이와 같은 과정으로 감정사전을 생성하며, 단지 사용하는 사전의 종류나 각 감정어에 감정 극성과 점수를 부여하는 방법에서 차이가 있다.

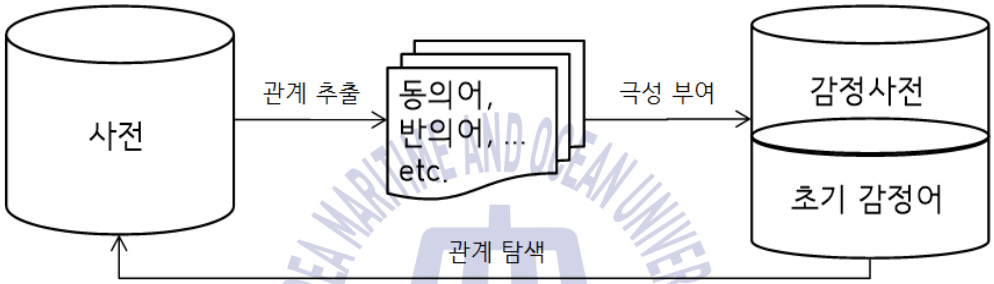


Figure 2.1 The processing of creating a sentiment lexicon using a lexicon-based method

비교적 적은 수의 초기 감정어를 시작으로 이들의 동의어와 반의어들을 WordNet에서 추출하여 동의어에는 동일한 극성을, 반의어에는 반대 극성을 부여하여 사전에 추가한다. 그것을 반복하여 감정사전의 크기를 점점 키워나가는 접근법이 있다(Hu & Liu, 2004; Esuli & Sebastiani, 2005; Blair-goldensohn, *et al.*, 2008; Rao & Ravichandran, 2009; Hassan, *et al.*, 2010; Dragut, *et al.*, 2010). 각 단어의 극성을 확률로 부여하는 연구도 있다(Kim & Hovy, 2004). 언어적 특징에 따라 접두사, 접미사 등을 활용하여 동의어와 반의어뿐만 아니라 더욱 다양한 관계어들을 찾아낸 방법론도 있다(Mohammad, *et al.*, 2009).

WordNet 상의 두 단어 사이에 의미상의 거리를 계산할 수 있다. 이러한 의미의 거리를 관계 추출 시에 활용하여 극성을 부여하거나(Kamps, et al.,

2004) 그에 따른 실질적인 수치를 부여하기도 했다(Williams & Anand, 2009). WordNet에서 추출한 동의어로 그래프를 생성한 뒤, 준지도학습 알고리즘을 사용하여 긍정과 부정의 두 극성으로 분리시키기도 했다(Rao and Ravichandran, 2009). 총 세 가지 알고리즘으로 결과를 평가했으며, 최소 절단 알고리즘(Stoer & Wagner, 1997)과, 무작위 최소 절단 알고리즘(Karger, 1993), 그리고 레이블 전파 알고리즘(Xiaojin & Zoubin, 2002)을 사용했다. 결과적으로 최소 절단 계열 알고리즘이 더 나은 F1 점수를 보였으나, 레이블 전파 알고리즘이 매우 높은 정밀도(precision)를 나타냈다. WordNet에서 추출한 동의어, 반의어로 그래프를 생성한 뒤, 레이블 전파 알고리즘을 사용하여 단어들의 극성을 구하기도 하였다(Blair-goldensohn, *et al*., 2008).

단어 간 관계의 다양성 측면에 있어서, 동의어와 반의어뿐만 아니라 상위어(hypernym)와 하위어(hyponym)도 WordNet에서 추출하였다. 추출한 단어에게 긍정, 부정, 중립의 세 가지로 분류하고 각 극성에 속하는 정도를 계산했다. 이렇게 제작된 대표적인 영어의 감정사전이 SentiWordNet(Esuli and Sebastiani, 2006)이다.

사전을 기반으로 감정사전을 제작할 때, 초기 단어와 더불어 탐색 시 발견되는 감정어들을 비교적 쉽게 구할 수 있다. 하지만 초기 단어의 종류와 범위를 특정하기가 방대하고 복잡하며 연구자 개개인의 노력도 필연적으로 들어갈 수밖에 없다. 더욱이 언어에 따라 단어 간 관계까지 정의되어 있는 사전을 구하는 그 자체가 난관일 수 있다. 그러므로 사전 기반 방법의 단점을 보완하기 위해 여러 말뭉치로부터 감정사전을 제작하는 말뭉치 기반 방법이 최근 주류를 이루고 있다.

2.2.2 말뭉치 기반 감정사전

말뭉치 기반 방법은 사전이 아닌 문서들의 집합인 말뭉치로부터 단어와 구문을 수집한다. 말뭉치를 수집한 영역에서 사용되는 단어와 구문을 위

주로 감정사전을 제작하기 때문에 말뭉치의 크기와 분야(domain)에 제한이 될 수 있다. Figure 2.2는 말뭉치를 기반으로 감정사전을 제작하는 과정을 보인다. Figure 2.1에서의 사전과 달리 말뭉치에서 단어와 구문을 추출해내 감정사전을 제작해 나간다. 대다수의 말뭉치 기반 방법들은 Figure 2.2와 같이 감정사전을 제작한다. 그러나 초기 감정어를 선정하는 것, 사용하는 말뭉치의 종류와 분야, 말뭉치 내에서 단어 간 관계의 종류와 파악 방법 등에 차이가 존재한다.

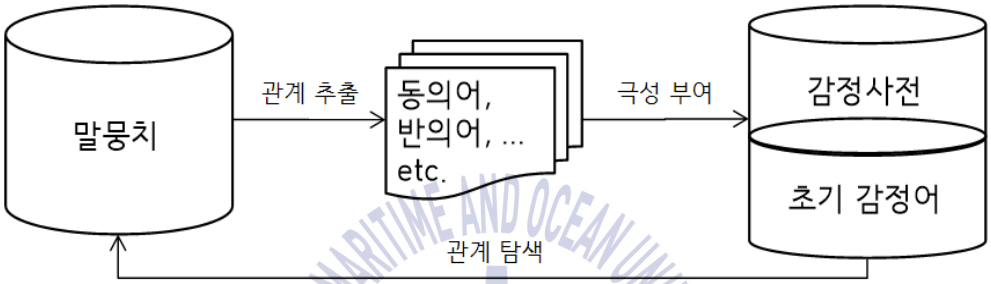


Figure 2.2 The processing of creating a sentiment lexicon using a corpus-based method

말뭉치 기반 방법의 주요 접근방법은 미리 선정된 형용사 집합을 통해 말뭉치 내에서 다른 감정 형용사를 관계적인 특성에 근거하여 찾아내는 것이다. 가장 기초적인 관계로서, 랜덤하게 설정된 긍정, 부정의 초기 형용사 집합과 접속사로 연결된 다른 형용사들을 추출한 연구가 있다(Hatzivassiloglou & McKeown, 1997). 해당 연구에서의 예시로 “simple and well-received”라는 구문에서 ‘and’와 같은 접속사로 연결된 형용사는 감정상의 동일한 극성을 가지는 특성을 이용했다. 그와 반대로 ‘but’이라는 접속사와 연결된 형용사는 서로 반대의 극성을 가지는 특성을 이용했다. 하지만 “bold but cautious”의 경우와 같이 ‘but’과 같은 접속사로 연결된 형용사라도 반대의 극성을 가지지 않는 예외가 있다고 판단하여, 접속사로 연결된 단어들에 클러스터링을 수행한 뒤, 그 결과에 따라 각 형용사에 감정의 극성을 할당하였다(Hatzivassiloglou & McKeown, 1997).

또한 문장 내에서의 접속사만 사용하지 않고 문장과 문장을 연결하는 접속사로까지 확장하기도 하였다(Kanayama & Nasukawa, 2006). 어떤 문장의 다음에 나오는 문장은 앞선 문장과 동일한 감정의 극성을 가지는 경우가 많다는 감정의 일관성(coherency)을 기초로 접근하였다. 일본어 말뭉치를 활용하여 문장 내·외의 접속사로 연결된 형용사들을 추출하고 감정을 할당하는 식으로 영역별 감정사전을 제작했다.

동일한 영역 내에서도 하나 이상의 감정을 가지는 단어를 처리하기 위해 구문을 추가하여 감정사전을 제작하기도 하였다(Ding, *et al.*, 2008). 카메라의 상품평에서 "카메라 배터리의 수명이 매우 길다"의 '길다'와 "프로그램 구동시간이 너무 길다"의 '길다'는 같은 단어이지만 서로 다른 감정의 극성을 가진다. 그렇기 때문에 두 단어를 결합한 ('카메라 배터리', '길다')와 ('프로그램 구동시간', '길다')인 감정 구문을 사용하여 감정사전을 제작했다.

웹 검색 엔진을 사용하여 웹 페이지 내 특정 단어와 초기 감정어의 동시등장횟수로 감정어들을 추출하여 감정사전을 제작한 연구 사례도 있다(Turney, 2002; Turney & Littman, 2003). 동일한 방법으로 중국어 감정사전을 제작하기도 하였다(Wiebe, *et al.*, 1999).

어떤 한 분야에서 제작된 감정사전을 다른 분야에 적용시키려는 연구도 진행되었다. 범용적인 감정어 집합을 특정 분야의 감정분석에 적합하도록 수정하는 방법이 제안되었다(Choi & Cardie, 2009). 특정 분야의 감정어와 범용 감정어 집합 내의 관계를 통해 해당 분야에 적합한 감정어 극성을 추측해 나가는 방식으로 해당 분야에 특화된 감정어를 확장해나갔다. 또한 '긍정(부정) 감정어가 많이 사용된 문서는 긍정(부정)적인 문서인 것'과 '긍정(부정) 문서에서 자주 등장하는 단어는 긍정(부정)인 감정어'라는 가정을 사용하여 타 분야에 적합한 감정어를 제작한 연구도 있다(Du, *et al.*, 2010).

주관성, 객관성의 분류를 통해 감정사전을 제작한 연구도 있다(Wiebe &

Mihalcea). 감정어를 제작하기 이전에 해당 단어의 주관성을 판단한다. 해당 연구에서는 단어가 감정을 지니고 있음을 판단할 때 그 단어의 주관성이 직접적인 영향을 미친다는 것을 실험을 통해 나타내고 있다. 그래서 말뭉치로부터 추출한 단어들의 주관성을 분류한 뒤, 주관적인 단어들만 사용하여 감정사전을 제작했다.

특정 분야의 말뭉치를 활용하여 감정어 후보들을 추출하여 레이블 전파 알고리즘을 적용해 해당 분야에 특화된 감정사전을 제작한 연구도 있다 (김정호, 2015). 이는 사전 기반 방법에서 분야에 따른 감정어 극성의 변화에 대한 모호성을 해결하기 위한 것으로 해당 연구에서는 영화평, 상품평에 특화된 감정사전을 제작하였다.

기존에 제안되었던 말뭉치 기반 방법들은 초기 감정어와의 동시등장 정보를 활용하여 추가적인 감정어들을 말뭉치에서 추출했다. 하지만 단순히 동시에 등장한다는 정보만으로 초기 감정어와 같은 감정의 극성을 지닌다고 보기엔 근거가 부족하다고 할 수 있다. 동일한 극성을 가지는 단어들의 동시등장 횟수가 그 반대 단어와의 동시등장 횟수보다 항상 많지 않기 때문이다. 따라서 기존의 동시등장 정보만을 활용하는 방식으로는 두 단어 간의 관계로 정확한 단어의 감정을 부여하기가 어렵다고 할 수 있다.

2.2.3 집단지성 기반 감정사전

집단지성 기반 방법은 사전이나 말뭉치가 아닌 어떠한 방법으로 미리 구성한 감정어 후보들을 이용한다. 이와 같이 구성된 감정어 후보를 여러 사람들이 감정의 극성 또는 극성과 강도를 동시에 매긴 뒤, 감정사전을 확장해 나간다. 사람들의 보편적인 인식을 이용해 감정사전을 제작하는 방식이기 때문에 말뭉치 기반 방식과는 달리 해당 영역에 국한되지 않고 사전 기반 방식과 비슷하게 범용적인 의미를 활용하여 감정이 부여된다. 특정한 알고리즘이나 매커니즘을 통해 감정사전을 제작했던 앞선 두 방식과 달리 직접적인 사람의 수고와 노력이 들어가는 방식으로, 제작 시간과

비용이 상대적으로 높다. 하지만 평가하는 인적 집단이 커질수록, 평가된 감정 점수의 신빙성이 높아지므로 앞선 두 방식보다 확실히 검증되고 정확한 감정사전을 제작할 수 있는 방법이기도 하다.

Figure 2.3은 집단지성 기반으로 감정사전을 제작하는 과정을 보인다. 이전 두 방법과 달리 사전이나 말뭉치에서 관계를 추출해내는 과정이 없다. 집단지성을 통한 극성 부여를 통해 후보 감정어의 감정 극성을 바로 부여하는 것이다. 대다수의 집단지성 기반 방법들은 Figure 2.3의 방법대로 감정사전을 제작한다. 그러나 극성 부여 방법, 집단지성의 구성 인원, 극성 부여 후 감정사전에 추가하는 절차 등에서 차이가 존재한다.

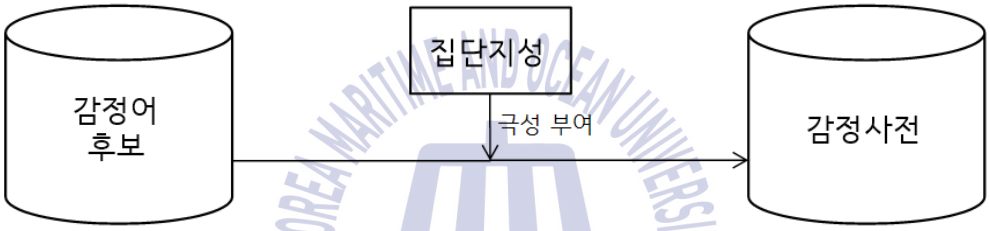


Figure 2.3 The processing of creating a sentiment lexicon using a collective intelligence-based method

집단지성 기반 방법의 주된 접근방법은 극성의 부여 방식보다는 극성의 종류에 초점을 맞춘다. 가장 대표적인 극성의 종류는 긍정과 부정으로 초기에는 감정사전이란 특정한 분야라기보다 여러 카테고리의 어휘사전이 제작되었다(Stone, *et al.*, 1962; Pennebaker, *et al.*, 1993). ‘general inquirer’는 어휘사전으로, 처음 배포된 1962년 이래로 지속적으로 카테고리나 단어가 추가되어 현재는 183개의 카테고리, 11,000개 이상의 단어를 보유하고 있다. 그 중에서 긍정 카테고리의 단어는 1,915개, 부정 카테고리의 단어는 2291개가 존재한다. ‘Linguistic Inquiry Word Count(LIWC)’는 처음 제작된 1993년 이래로 지속적으로 카테고리나 단어가 추가되어 LIWC-2001(Pennebaker, *et al.*, 2001), LIWC-2007(Pennebaker, *et al.*, 2007), LIWC-2015(Pennebaker, *et al.*, 2015) 총 세 가지 버전이 존재한다. 영리단

체 LIWC[2]에서 판매하고 있으며 총 76개의 카테고리, 4,500개 이상의 단어를 보유하고 있다. 긍·부정 카테고리에 포함된 단어는 905개가 존재한다.

단순히 감정의 긍·부정이라는 극성만 부여하지 않고 해당 극성의 강한 정도를 점수로 부여한 연구도 있다(Bradley & Lang, 1999; Nielsen, 2011). ‘Affective Norms for English Words(ANEW)’는 1,034개 단어에 대한 감정사전이다. 5를 중립으로하여 1~9사이의 정수 값을 각 단어의 감정에 맞게 부여했다. ‘AFINN’은 0을 중립으로 −5~5사이의 정수 값을 각 단어의 감정에 맞게 부여한 감정사전이다. 2009년에서 2011년까지 저자인 Finn Arup Nielsen 혼자서 직접 제작한 것으로 AFINN-96, AFINN-111의 총 두 가지 버전이 존재한다. 각각 1,468개, 2,477개의 긍·부정 감정어를 보유하고 있다.

또한 단순히 사전의 형태로 제작하는데 그치지 않고 온라인 API로 공개한 연구도 있다(안정국 & 김희웅, 2015). ‘openhangul(오픈한글)’은 국내 대학생 소셜 네트워크 사이트에서 투표를 진행하여 감정사전을 제작하였다. 총 35,000번 투표가 진행됐으며, 투표인에게 10가지 단어를 1분 내에 중립, 긍정, 부정의 극성을 선택하게 하였다. 투표가 진행될수록 쌓인 각 단어별 극성의 빈도수가 해당 감정어의 극성과 점수가 되는 방식이다. 이렇게 제작한 감정사전은 온라인 API 형식으로 “http://openhangul.com”에서 제공된다.

감정분석의 분야는 매우 다양하기 때문에 모든 분야마다 그에 맞는 감정사전을 만들기는 불가능에 가깝다. 따라서 일반적인 감정사전은 대부분이 범용적이다. 범용적인 목적의 감정사전만을 통한 감정분석은 중의적인 단어나 분석할 문서의 분야에 따라 정확하게 분석이 어려울 수 있다.

2.3 VADER 감정사전

2) http://liwc.wpengine.com

2.2.3절에서 언급한 감정분석이 정확하게 수행되기 어려운 점을 해결하기 위해 언어적인 규칙을 추가한 연구도 있다(Gilbert, & Hutto, 2014). Valence Aware Dictionary and sEntiment Reasoner(VADER)는 집단지성 기반 방법으로 제작한 감정사전과 더불어 언어적인 규칙 다섯 가지를 활용해 감정분석을 수행하는 시스템이다. VADER의 감정사전에는 긍정 감정어 3,345개, 부정 감정어 4,172개의 총 7,517개의 단어가 포함되어 있으며 각 감정어는 0을 중립으로 -4.0~4.0 사이의 실수로 값이 부여되어있다.

감정어 후보 집합은 앞서 설명한 LIWC, ANEW, general inquirer를 검토하여 이모티콘(emoticons), 축약어(abbreviation), 속어(slang) 등을 추출하여 통합해 만들었다. 이와 같이 제작한 감정어 후보를 Amazon Mechanical Turk[3]에서 10명의 평가자를 고용하여 감정 점수를 부여하였다. 의미있는 데이터 확보와 품질관리를 위해 네 가지의 과정을 진행했다. 첫 번째로 영어독해능력 평가를 실시하여 정답률 80% 이상이어야 하며, 두 번째로 온라인 감정평가 교육을 이수해야 했다. 세 번째로 평가 데이터 25개 마다 감정 점수가 확정된 다섯 개의 단어를 집어넣어 평가하게 한 뒤, 평가자가 매긴 점수가 정해진 점수의 표준편차의 두 배 이상 크게 평가된 단어가 세 개 이상일 경우에 해당 25개 데이터를 평가자의 데이터에서 삭제시켰다. 네 번째로 다른 평가자들이 가장 많이 매길 것 같은 점수도 체크하게 하여 맞출 경우 인센티브를 지급하였다. 또한 점수 부여 과정의 신빙성을 높이기 위해 각 단어마다 평가자들이 매긴 점수의 표준편차가 2.5 미만일 경우에만 사전에 포함시켰다. Figure 2.4는 VADER 감정사전의 제작 과정을 보인다.

3) https://www.mturk.com

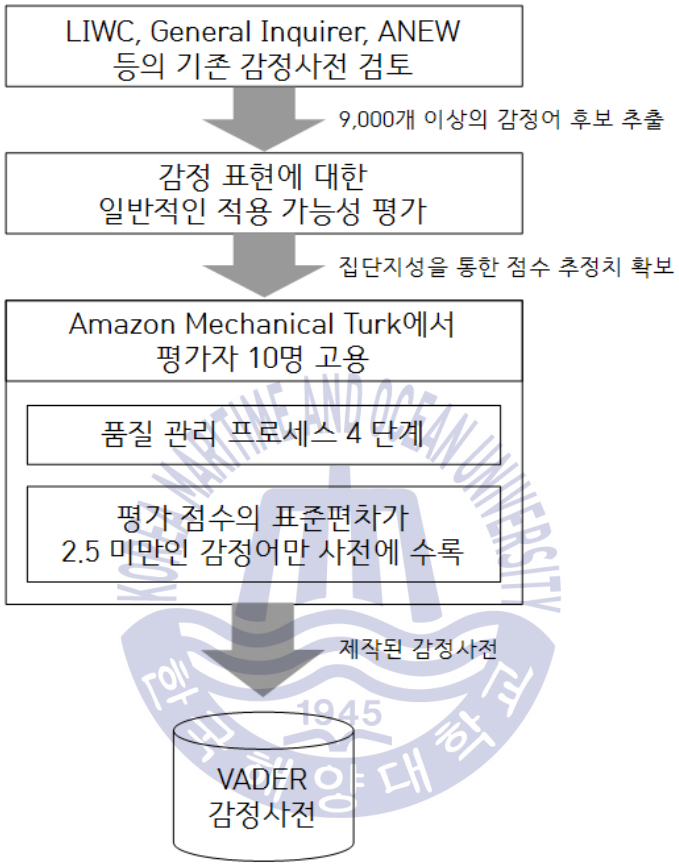


Figure 2.4 The processing of creating the VADER sentiment lexicon

이와 같이 제작한 감정사전을 검증하기 위해 무작위로 추출한 10,000개의 트윗(tweets)을 사용했다. Python의 웹 마이닝 모듈인 Pattern(Smedt & Daelemans, 2012)을 사용하여 10,000개 트윗 중 긍·부정 상위 400개를 추출한 뒤 두 명의 전문가에게 추출한 트윗 문서의 감정 점수를 -4~4 사이의 정수로 매기게 하였다. 이러한 과정에서 앞서 설명한 다섯 가지의 언어적 규칙을 발견하였다. 첫 번째로 대문자로 사용된 감정어의 경우에는 소문자로 사용된 감정어보다 추가적인 점수를 부여하는 것과 두 번째로 느낌표, 물음표와 같은

문장부호(punctuation)를 활용하였다. 세 번째로 부정어(negation) 앞에 온 감정어의 극성을 정반대로 변화시켜주거나 네 번째로 이모티콘과 축약어, 속어 활용, 다섯 번째로 ‘extremely’, ‘hardly’와 같은 감정의 증폭어(booster words)가 나타날 경우 뒤따르는 감정어에 추가적인 점수를 부여하였다.

이와 같이 제작한 감정분석 시스템을 다른 감정사전 및 기계학습 모델과 비교하였다. 4,200개의 트윗을 사용한 소셜 미디어 분야, 10,605개의 로튼토마토[4] 영화평을 사용한 영화평론 분야, 3,708개의 아마존[5] 상품평을 사용한 상품평론 분야, 5,190개의 뉴욕타임즈[6] 기사를 사용한 사설 분야의 총 네 가지 분야에서 실험을 진행하였고 소셜 미디어 분야에선 0.96의 F1 점수로 사람(0.84)보다 높은 정확도를 보였으며 나머지 세 분야에서는 사람(0.92, 0.85, 0.65)보다는 낮았으나 0.61, 0.63, 0.55의 F1 점수로 여타 방법론보다 높은 정확도를 보였다.

본 논문에서는 이와 같이 검증되고 정확도 높은 VADER 감정사전을 기준으로 활용하여 한영 이중언어사전에서의 한국어 형태소 감정어 후보 선정과 한영 이중언어그래프 상의 초기 정점의 레이블 값 부여를 진행한다.

4) https://www.rottentomatoes.com
5) https://www.amazon.com
6) https://www.nytimes.com

제 3 장 감정 점수 전파를 통한 감정사전 제작

본 논문은 그래프 기반의 영어 감정사전의 감정 점수 전파를 통한 한국어 감정사전 제작 방법을 제안한다. 제안하는 방법은 한국어 감정어 후보를 선정하기 위해 한영 병렬 말뭉치와 VADER 감정사전을 이용하여 한영 이중언어사전과 한영 이중언어그래프를 생성한다. 이와 같이 생성한 그래프 상에서 레이블 전파 알고리즘을 수행하여 VADER 감정사전 내 감정어들의 감정점수를 한국어 감정어 후보로 전파하여 한국어 감정사전을 제작한다.

Figure 3.1은 본 논문에서 감정사전을 제작하는 전체적인 과정과 각 과정의 방법을 보인다. 첫 번째로 한영 병렬 말뭉치의 정렬된 문장 자료를 활용하여 '영어 단어 - 영어 단어' 쌍의 상호정보량(이하 PMI) 행렬, '한국어 형태소 - 영어 단어' 쌍의 PMI 행렬을 제작한다. 제작된 PMI 행렬의 영어 단어 중 VADER 감정어들을 기준으로, 영어 단어와 한국어 형태소 PMI 벡터의 코사인 유사도를 측정한다. 이와 같이 가장 유사도가 높은 '한국어 형태소 - VADER 감정어' 대역어 쌍을 추출하여 한영 이중언어사전을 제작한다. 두 번째로, 추출한 이중언어사전과 사전에 포함되지 않은 말뭉치 내의 한국어 형태소들로 한영 이중언어그래프를 제작한다. 마지막으로 제작된 그래프 상에서 레이블 전파 알고리즘을 적용하여 그래프 내 한국어 형태소들의 감정 극성과 점수를 부여하여 한국어 감정사전을 제작한다. 각 과정의 자세한 내용은 순서에 맞게 각 절에서 설명한다.

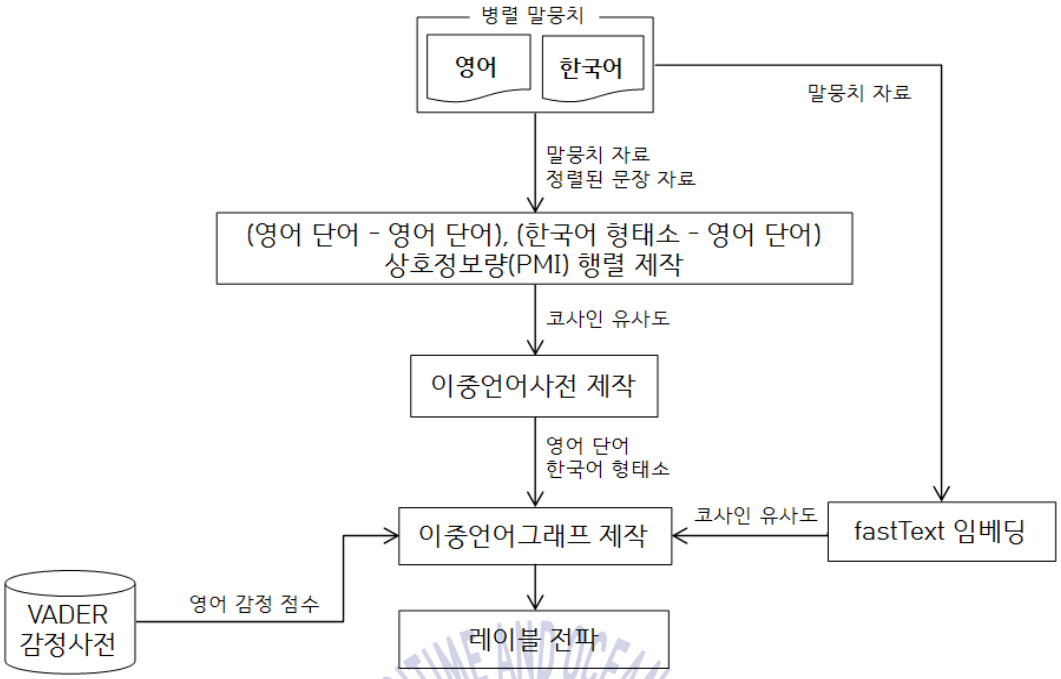


Figure 3.1 The processing of creating a Korean sentiment lexicon by the proposed method

3.1 한영 이중언어사전 제작

3.1.1 한영 병렬 말뭉치 토큰화

한영 이중언어사전은 '한국어 형태소 - VADER 감정어' 쌍의 집합이다. 사용되는 한국어 형태소와 영어 단어는 뉴스 기사 자료를 기반으로 자체 제작한 한영 병렬 말뭉치에서 추출된다. Table 3.1은 해당 말뭉치의 자세한 특징을 나타낸다.

Table 3.1 Statistics of the Korean-English parallel corpus

구분	한국어 말뭉치	영어 말뭉치
뉴스 출처	동아일보, 중앙일보, ETnews, YBM, CNN, Time, VOA	
문서 수	80,503	
문장 수	421,445	421,998
형태소(단어) 수	37,608,538	27,938,222
단일 형태소(단어) 수	193,916	475,955

이중언어사전을 제작하는 이유는 영어 단어의 의미에 부합되는 한국어 형태소 집합을 찾기 위함이다. 말뭉치에서 문서와 문장을 각각 영어 단어와 한국어 형태소로 토큰화(tokenization)를 수행한다. 영어에서 단어 단위를 선택한 이유는 VADER 감정사전의 감정어가 단어 단위이기 때문이다. 그에 비해 한국어 역시 단어 단위가 아닌 형태소 단위인 이유는 의미로서의 최소 단위라는 점과 한국어라는 언어의 특성 상 어떠한 단어는 기능어와 함께 매우 다양한 형태로 쓰인다는 점, 그리고 제작한 감정사전으로 감정분석 시 중의성과 모호성을 가능한 줄이기 위함이다. 영어 말뭉치의 단어 토큰화는 python의 NLTK 모듈(Edward & Steven, 2002)을 이용한다. 한국어 말뭉치의 형태소 토큰화는 utagger 형태소 분석기(신준철, 옥철영, 2012)를 사용한다. Figure 3.2는 토큰화 결과에 대한 예제를 보인다.

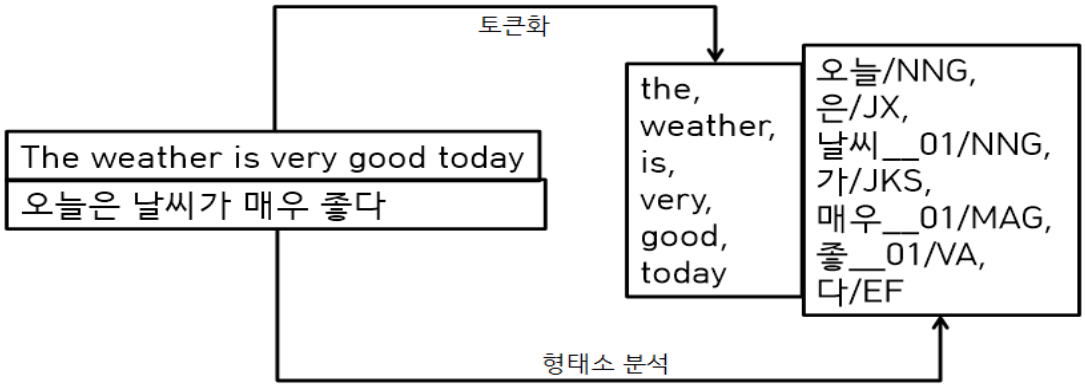


Figure 3.2 An example of tokenization of Korean and English sentences

3.1.2 상호정보량(PMI) 행렬 제작

Table 3.2는 VADER 감정사전 내 감정어들의 상세한 특징과 영어 말뭉치와의 상관성을 보인다.

Table 3.2 Statistics of the VADER sentiment lexicon

구분	긍정	부정	총합
전체 감정어	3,345	4,172	7,517
이모티콘	140	157	297
축약어, 속어	57	101	158
영어 말뭉치 내 존재	2,636	3,411	6,047

VADER 감정사전의 감정어 개수는 긍정 3,345개, 부정 4,172개로 총 7,517개가 존재한다. 그 중에서 이모티콘이 긍정 140개, 부정 157개로 총 297개가 있지만 한영 이중언어사전 제작에 사용하기엔 말뭉치에 등장하지 않는다. 뉴스 기사 기반의 말뭉치이므로 이모티콘, 축약어, 속어가 등장하지 않는다. 그러나 기사 내 문장에서 빈도가 매우 낮게 사용되는 단어가 존재한다. 긍정 감정어는 612개, 부정 감정어는 858개 총 1,470개가 등장하지 않는다.

그 이유로 VADER 감정사전의 감정어 선정 방식을 들 수 있다. VADER 감정사전에 존재하는 감정어들은 어간 추출(stemming)이나 원형화(lemmatization)을 진행하지 않은 상태로 수록되어 있기 때문이다. 이러한 과정을 진행하지 않는 이유는 같은 어간을 지닌 단어라도 어떠한 접미사가 붙는지에 따라 감정의 극성 또는 강도가 변화할 수 있기 때문이다. Table 3.3은 VADER 감정사전에 존재하는, ‘cute’ 감정어의 다양한 접미사가 붙어서 파생된 단어들을 보인다.

Table 3.3 Derived words from 'cute' with various suffixes in the VADER sentiment lexicon

단어	접미사	기능	감정 점수	영어 말뭉치 내 등장횟수
'cutely'	'-ly'	'cute' 의 부사화	1.3	1
'cuteness'	'-ness'	'cute' 의 명사화	2.3	9
'cutenesses'	'-nesses'	'cute' 의 명사화	1.9	0
'cuter'	'-r'	'cute' 의 비교급	2.3	7
'cutesie'	'-sie'	'cute' 의 인격화	1.0	0
'cutesier'	'-sier'	'cutesie' 의 비교급	1.5	0
'cutesiest'	'-siest'	'cutesie' 의 최상급	2.2	2
'cutest'	'-st'	'cute' 의 최상급	2.0	5
'cutesy'	'-sy'	'cute' 의 인격화	2.1	0
'cutey'	'-y'	'cute' 의 인격화	2.1	4
'cutie'	'-ie'	'cute' 의 인격화	1.5	0
'cutiepie'	'-iepie'	'cute' 의 인격화	2.0	0

'-ly', '-ness'와 같은 다른 품사로 변형시키는 접미사가 붙거나 '-r', '-st'와 같은 비교급, 최상급 접미사가 붙기도 한다. 이렇게 단어의 의미적 증폭을 뜻하는 비교급, 최상급 접미사가 붙으면 기본형 단어보다 감정 점수가 0.5~1.2점 강해진 것을 볼 수 있다.

이와 같이 기본 어간에 여러 접미사가 붙은 파생형 단어들이 감정사전에 추가되어 있는 것은 좀 더 정확하고 폭넓은 감정분석을 위해 필수적이다. 하지만 그러한 단어들이 실제 문서에서 비슷한 비율로 등장하지 않는다. 기본 어간이 되는 단어 'cute'가 말뭉치에서 215번 등장하는 반면 그것의 파생어들은 아무리 많이 등장해도 10번이 채 등장하지 않고 오히려 아예 등장하지 않는 단어들이 대다수이다. 이러한 예시로 말뭉치의 크기에 관계없이 등장하지 않는 감정어들이 존재할 수밖에 없으며, 또한 등장하지 않는 감정어들은 기본 어간이 되는 단어에서 파생된 단어들이 대다수이기에 배제하더라도 VADER 감정사전의 감정 점수들을 크게 대표하지 않음을 알 수 있다. 그러므로 본 연구에서는 VADER 감정사전에 존재하고 이모티콘, 축약어, 속어가 아닌 7,062개의 단어 중에서 자체 제작한 뉴스 기사 기반 말뭉치에 등장하는 6,047개의 감정어만 이중언어사전 제작

에 사용한다.

VADER 감정사전은 영어 단어로 구성되어 있기 때문에 감정 점수를 한국어로 전파하여 한국어 감정사전을 만들기 위해서 합리적이고 근거있게 전파 대상을 규정해야 한다. 따라서 한국어 형태소를 추출하기 위해 아래의 사항을 가정한다.

가정. 어떤 한국어 형태소와 VADER 감정어 사이의 모든 영어 단어와의 PMI 값 분포가 유사할수록 둘의 의미 역시 유사하다.

위 가정 1은 동시등장 빈도에 근거한 것이다. 문장 단위로 정렬(alignment)되어 있는 병렬 말뭉치로서, 문장 단위의 동시등장 횟수를 계산할 수 있다. 이러한 동시등장 횟수를 말뭉치 내 등장 확률로 활용하여 각 '한국어 형태소 - 영어 단어' 쌍, 'VADER 감정어 - 영어 단어' 쌍의 PMI 값들을 저장한 두 개의 PMI 행렬을 제작한다.

Figure 3.3은 PMI 행렬들을 생성하는 과정을 보인다. Figure 3.4는 제작된 PMI 행렬의 예제를 보인다.

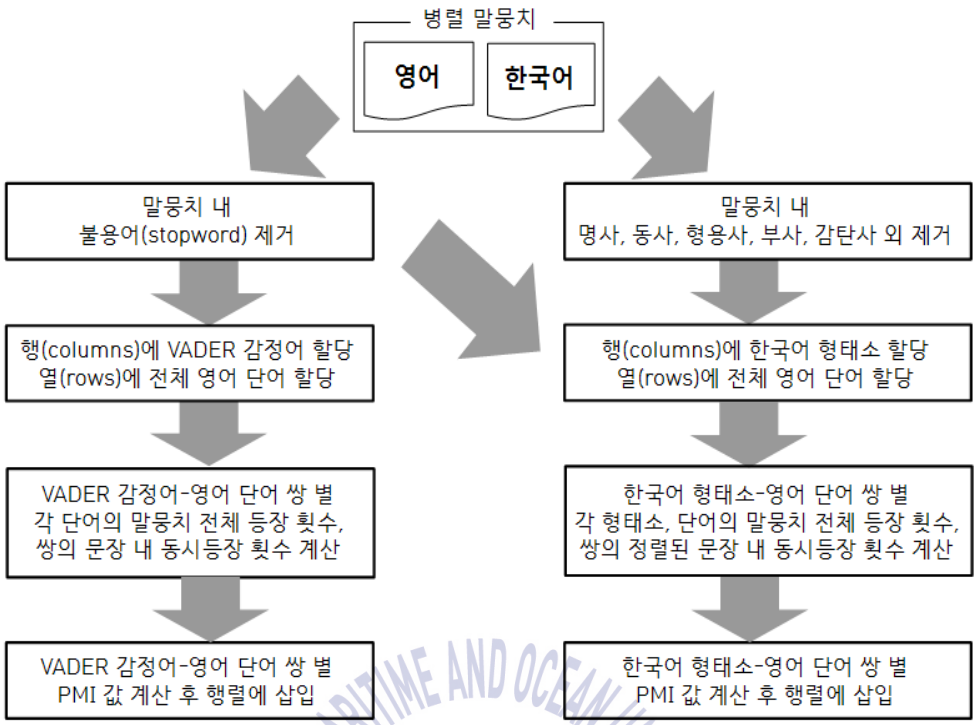


Figure 3.3 The processing of creating PMI matrices

	cute	pretty	love	hate	...
cute	26.6	25.3	18.1	3.7	...
pretty	25.3	28.7	31.9	8.8	...
love	18.1	31.9	27.3	5.5	...
hate	3.7	8.8	5.5	29.2	...
...	...	...	...	...	...

	cute	pretty	love	hate	...
귀엽/VV	27.8	22.5	15.3	8.3	...
예쁘/VA	23.2	26.9	18.1	6.8	...
사랑/NNG	18.7	15.4	12.5	5.1	...
증오/NNG	8	6.3	5.7	32.1	...
...	...	...	...	...	...

Figure 3.4 Examples of two generated PMI matrices

Figure 3.4와 같이 제작된 PMI 행렬로 각 VADER 감정어와 한국어 형태소에 벡터를 부여한다.

3.1.3 코사인 유사도를 통한 이중언어사전 제작

생성된 VADER 감정어들의 벡터와 한국어 형태소들의 벡터 간의 코사인 유사도를 계산하여 해당 영어 단어와 한국어 형태소가 얼마나 유사한지 측정한다. 측정한 값은 0~1 사이의 실수를 가지며 한국어 형태소 - VADER 감정어 쌍을 추출하기 위해 유사도의 상위 10위까지 사용한다.

차용된 쌍은 이중언어사전에 추가된다. 모든 VADER 감정어에 대해 상위 10위의 대역어를 추출하면 이중언어사전이 완성된다.

Figure 3.5은 PMI 행렬 내 코사인 유사도 계산을 통해 ‘cute’의 상위 10위 대역어가 추출되어 이중언어사전에 삽입되는 과정을 보인다.

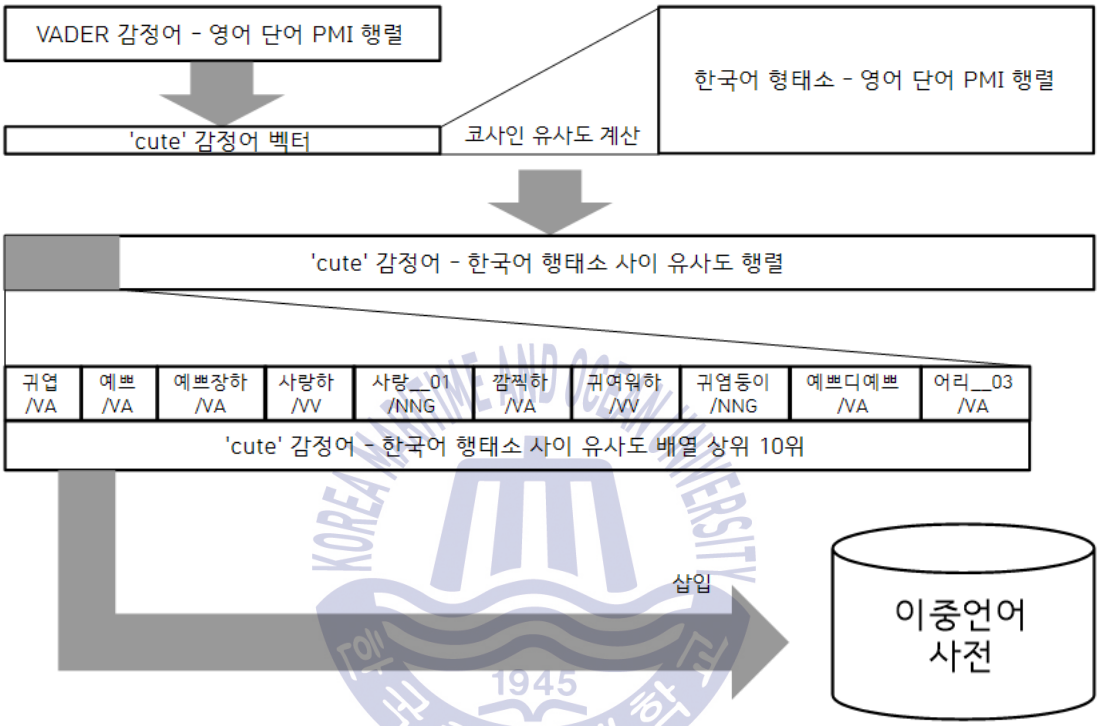


Figure 3.5 The extraction process of Korean morphemes with the top 10 cosine similarity of ‘cute’

Table 3.4는 제작된 한영 이중언어사전의 자세한 특징을 나타낸다.

Table 3.4 Statistics of a Korean-English bilingual lexicon

VADER 감정어	6,047
단일 형태소	19,832
명사	11,493
동사	6,126
형용사	1,392
부사	789
감탄사	32

3.2 한국어 fastText 표상 모델 제작

한국어 fastText 표상 모델은 한국어 말뭉치에 포함된 명사, 동사, 형용사, 부사, 감탄사 형태소들을 학습하여 표상 벡터로 변환한다. 텍스트 데이터에서 벡터 데이터로 변형되면 데이터 사이의 코사인 유사도를 측정할 수 있다. 이는 텍스트 데이터, 즉 한국어 형태소끼리 얼마나 유사한지 파악할 수 있다. 단순히 자소나 음절이 비슷하다는 의미가 아니라 fastText 모델은 학습 시 창 크기(window size)를 통해 주변 문맥까지 함께 표상화를 진행한다. 따라서 유사도가 높을수록 형태소 자체의 말뭉치 내에서의 의미나 쓰임새가 비슷하다는 뜻이 된다.

표상 학습에 사용한 한국어 말뭉치는 Table 3.1의 내용대로 뉴스 기사 기반이며 포함된 품사별 정보는 Table 3.5와 같다.

Table 3.5 Statistics of Korean corpus by POS

구분	전체 토큰 수	단일 타입 수
형태소	37,608,538	194,116
명사	9,990,219	73,187
동사	3,350,365	11,484
형용사	778,593	2,655
부사	603,537	2,116
감탄사	3,822	454

약 20여만 개의 형태소를 사용하여 학습을 진행한다. fastText는 기계학습 시 언어의 가장 작은 단위로 학습을 진행하기 때문에 한국어의 경우

음절 단위로 학습한다. 좀 더 작은 단위로 자소가 있으나 python 텍스트 데이터 상의 가장 작은 단위이기 때문에 영어는 알파벳, 한국어는 음절이 사용되는 것이다. 따라서 학습이 진행된 후 '사랑'이란 단어의 표상 벡터는 '사'와 '랑' 음절의 표상 벡터 두 개를 더한 결과가 출력된다.

본 연구에서는 python의 gensim에 구현된 fastText를 통해 한국어 형태소들을 표상화한다. Table 3.5의 내용대로 약 3천 7백만여 개의 형태소를 학습한다. fastText 모델의 하이퍼파라미터는 gensim에 구현되어있는 초기값을 사용한다. 표상 차원의 초기값은 100, 창 크기는 5, 반복 횟수는 5이다. 그렇게 학습이 진행된 fastText 모델은 한국어 말뭉치 내에 존재하는 모든 음절에 대한 표상 벡터를 보유하게 된다. Figure 3.6은 표상 벡터에 대한 예제를 보인다.

Figure 3.6 An example of fastText embedding of Korean morphemes and English words

3.3 한영 이중언어그래프 제작

한영 이중언어그래프는 이전 단계에서 생성한 한영 이중언어사전과 이중언어사전에 포함되지 않은 나머지 명사, 동사, 형용사, 부사, 감탄사들로 구성된다. VADER 감정어와 한국어 형태소들을 그래프의 정점으로 활용한다. VADER 감정어는 이중언어사전을 기반으로 한국어 형태소들과 간선을 연결하고 한국어 형태소, 즉 한국어 감정어 후보들은 자기 자신을 제외한 모든 다른 형태소들과 간선을 연결한다(fully connected). 연결된 간선의 가중치는 이중언어사전과 학습된 fastText 모델을 활용한다. VADER

감정어와 연결된 한국어 형태소 사이의 간선에는 가중치 1.0을 부여한다. VADER 감정어와 연결된 한국어 형태소들은 VADER의 감정 정보를 한국어 형태소들로 전파할 때, 다리(bridge) 역할을 해야 하므로 가능한 VADER의 감정 정보를 살리기 위함이다. 한국어 형태소 사이 연결된 간선의 가중치는 각 형태소의 표상 벡터를 활용한 코사인 유사도가 사용된다. Figure 3.7은 한영 이중언어그래프 제작 과정을 보인다. Figure 3.8~3.10은 과정에 대한 간단한 예시를 보인다.

Figure 3.8은 이중언어그래프의 정점으로 활용하기 위해 한국어 말뭉치와 이중언어사전에서 한국어 형태소와 영어 단어를 추출하는 과정을 보인다. 한국어 말뭉치에서 예시 형태소 '사랑/NNG', '시샘하/VV', '더없이/MAG' 세 가지를 추출하고 이중언어사전에서는 'cute'-'귀엽/VA', 'hate'-'증오/NNG'의 두 가지 쌍을 추출한다.

Figure 3.9은 Figure 3.8에서 추출한 한국어 형태소와 영어 단어들을 이중언어그래프의 정점으로 추가하고 간선을 연결하는 과정을 보인다. 이중언어사전에서 추출된 쌍은 서로 연결된 쌍끼리 간선을 연결한다. 한국어 형태소들은 출처에 상관없이 모두 빠짐없이 서로를 연결한다.

Figure 3.10은 Figure 3.9에서 추가된 정점의 레이블, 간선의 가중치를 초기화하는 과정을 보인다. 이중언어사전에서 추출된 영어 단어는 VADER 감정어이므로 VADER 감정사전의 감정값을 그대로 부여한다. 그에 비해 한국어 형태소들은 출처에 상관없이 아직 감정값이 전파되지 않았으므로 0.0을 '레이블이 없음'의 의미로 부여한다. 이중언어사전에서 추출된 쌍끼리 연결된 간선엔 가중치 1.0을 부여한다. VADER 감정어와 연결된 한국어 형태소들은 VADER의 감정 정보를 한국어 형태소들로 전파할 때, 다리(bridge) 역할을 하기 때문이다. 한국어 형태소 사이의 간선에는 각 형태소의 표상 벡터를 이용한 코사인 유사도를 가중치로 사용한다. 형태소가 표상된 벡터들 사이의 코사인 유사도를 구하는 것은 형태소들 사이의 의미의 유사함을 수치화시키는 것이다. 다음 절의 감정 점수 전파에 있어서

유사도가 높은 정점끼리는 전파 받는 점수에 큰 영향을 주고, 유사도가 낮은 정점끼리는 전파 받는 점수에 크게 영향을 주지 못한다.

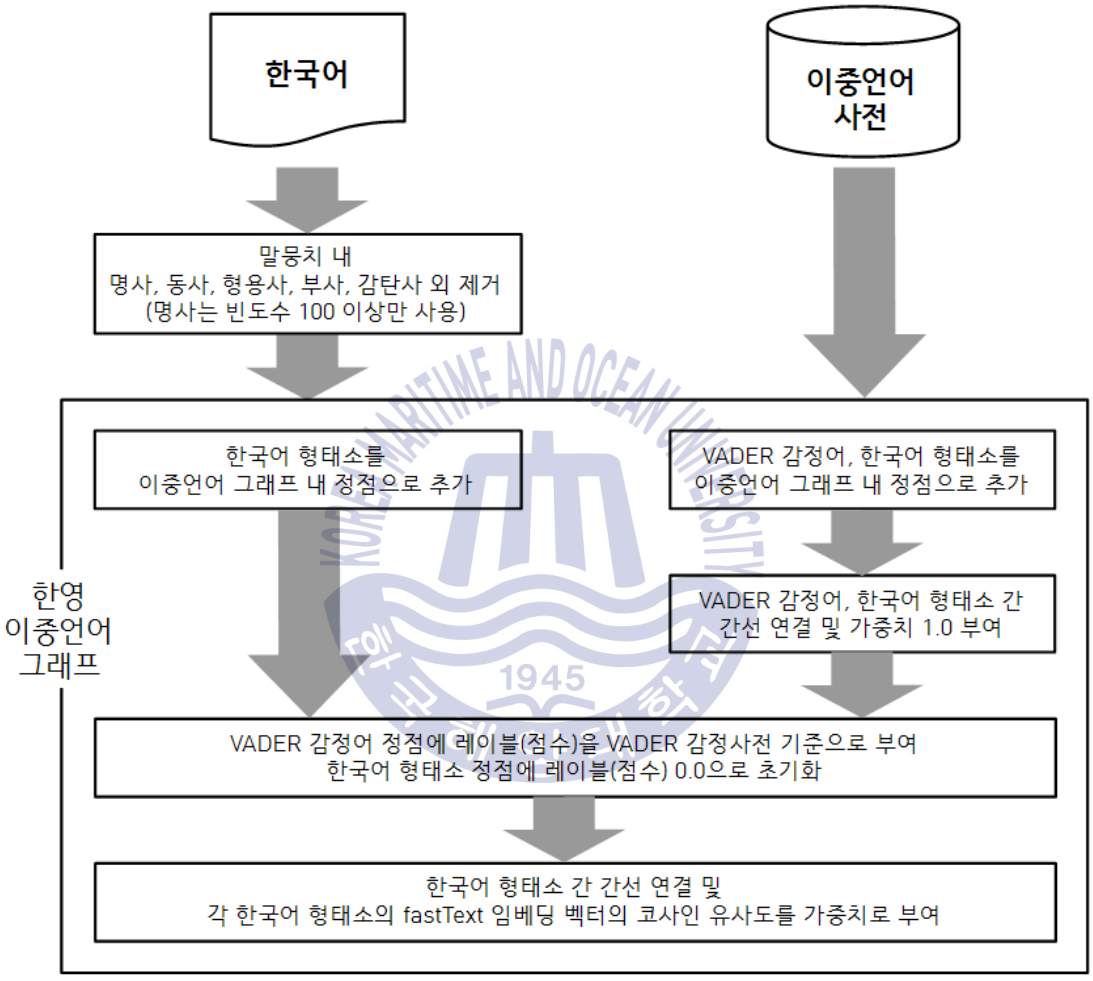


Figure 3.7 The processing of creating a Korean-English bilingual graph

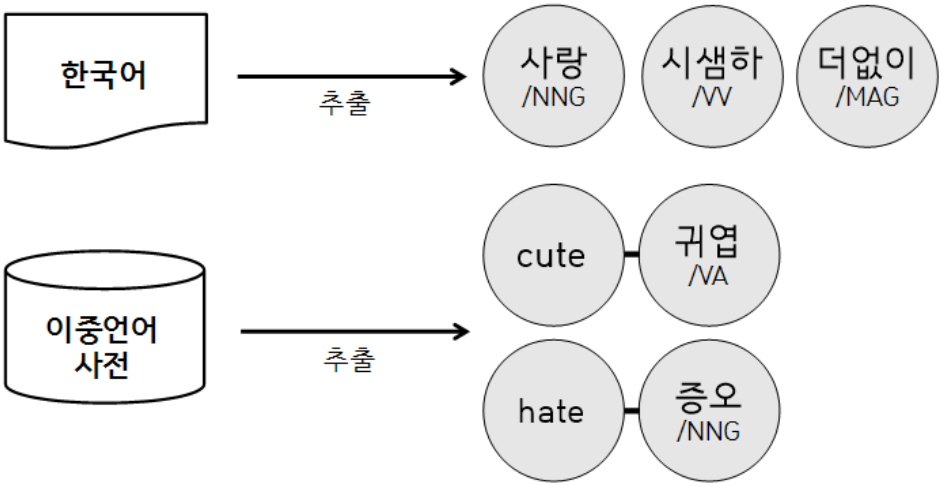


Figure 3.8 Extracting words and morphemes for vertex of a Korean-English bilingual graph

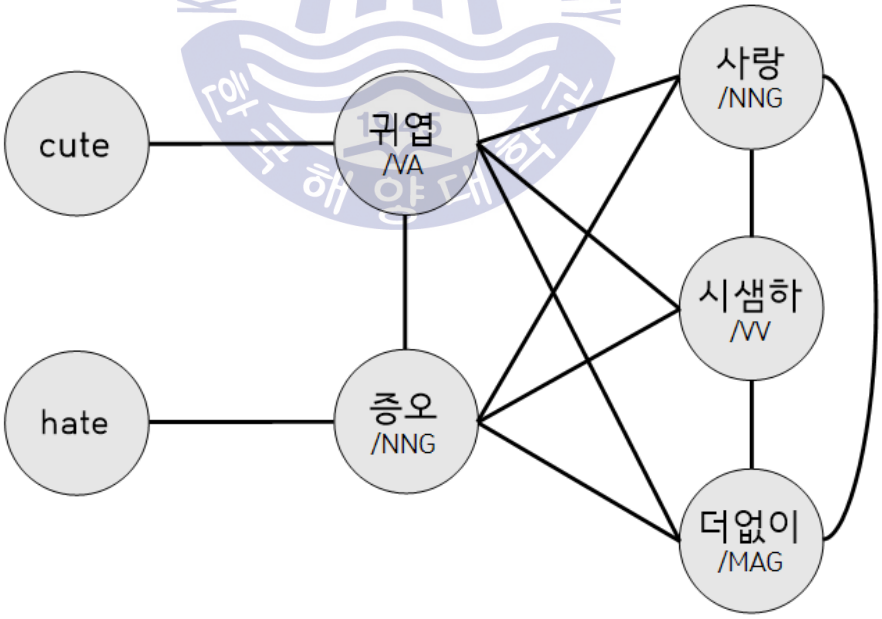


Figure 3.9 Adding edges in a Korean-English bilingual graph

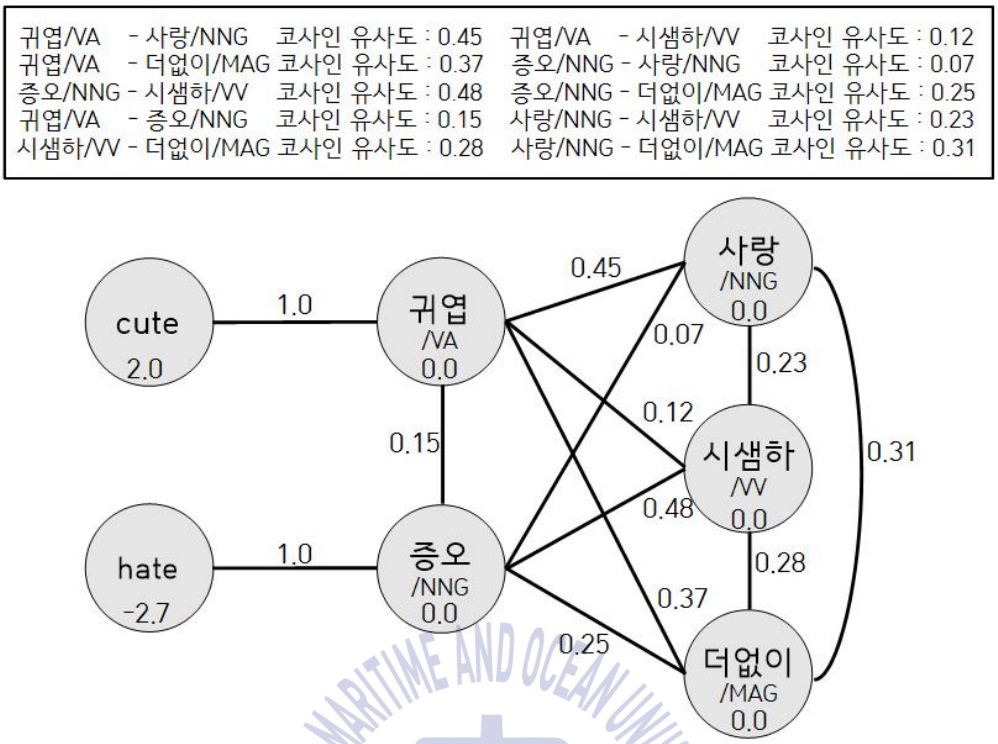


Figure 3.10 Initializing weights on edges in a Korean-English bilingual graph

Table 3.6은 제작된 한영 이중언어그래프의 자세한 특징을 나타낸다.

Table 3.6 Statistics of a Korean-English bilingual graph

정점 수	55,880
간선 수	2,506,812,676
정점 당 간선 보유 평균	44,861
초기 레이블 수	71 (-3.6~3.4)

3.4 감정 점수 전파

레이블 전파 알고리즘은 결과적으로 그래프 내에서 레이블이 존재하지 않는 정점에게 레이블이 존재하는 정점들의 레이블 중 하나를 부여하는

알고리즘이다. 앞서 제작한 한영 이중언어그래프 상에서 레이블이란 그래프에 포함된 VADER 감정어들의 감정 점수들이다.

그러나 ‘cute’라는 VADER 감정어가 2.0의 감정 점수를 지니고 있어도 ‘귀엽/VA’라는 형태소의 레이블이 2.0에 딱 맞게 수렴하기는 어려울 것이다. 사전적 의미는 동일하나 영어와 한국어라는 언어적 차이, 말뭉치의 크기로 인한 ‘cute’와 ‘귀엽/VA’의 언어적 사용 예시를 포함하는 한계 등등의 이유가 있을 것이다. 물론 레이블 전파 알고리즘의 마지막 단계인 레이블 할당 단계를 거치고 나면 높은 확률로 ‘귀엽/VA’는 그래프의 초기 단계에 존재했던 레이블 중 하나인 2.0을 받게 될 것이라고 예상할 수 있다. 사전적으로 의미가 동일한 단어이므로 결과적으로 레이블도 동일하게 부여될 것이라는 예상이다.

하지만 본 연구에서는 레이블 전파 알고리즘의 마지막 단계인 레이블 할당 단계를 수행하지 않는다. 오직 전파 반복 단계에서 마지막으로 수렴한 각 정점의 수치를 한글 감정어 후보의 감정 점수로 삼는다. 그 이유는 앞서 설명한 영어와 한국어라는 언어적, 문화적 차이를 감안하기 때문이다. Figure 3.11은 이중언어그래프에서의 레이블 전파 알고리즘 흐름도를 보인다. Figure 3.12~3.14은 Figure 3.10에서의 예시 그래프를 활용하여 초기화 단계, 전파 반복 1단계, 전파 반복 2단계까지의 간단한 예시를 보인다.

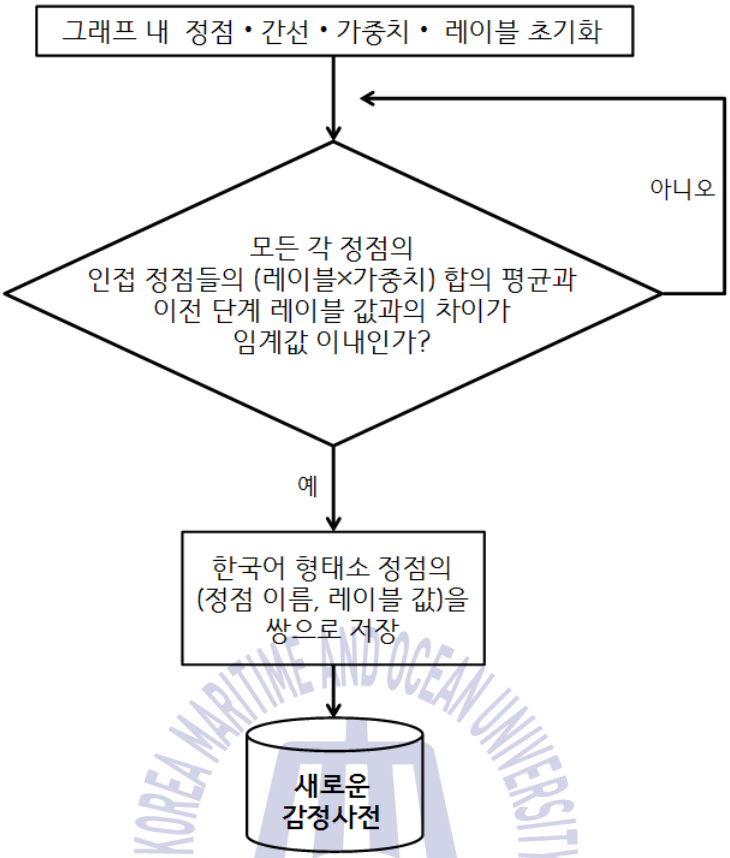


Figure 3.11 A flow diagram for the label propagation algorithm on a Korean-English bilingual graph

Figure 3.12는 한영 이중언어그래프 상에서 감정 점수 전파의 첫 번째 단계인 초기화 단계를 보여준다. 그래프를 생성하고 초기화하는 단계로, VADER 감정어와 한국어 형태소를 정점으로 선별한다. 그렇게 선별된 정점들은 Figure 3.7의 과정처럼 간선으로 연결이 되고 가중치가 부여된다. 그리고 마지막으로 각 정점에 레이블 값이 부여된다.

Figure 3.13~3.14은 각각 감정 점수 전파의 두 번째 단계인 전파 반복 단계의 첫 번째와 두 번째 반복이 끝난 직후를 보여준다. 이와 같이 그래프 상에서 감정 점수 전파가 진행되며, 각 반복 단계 마다 레이블 값 차이가 미리 정해둔 임계값보다 크다면 해당 임계값보다 작아질 때까지 수

렴하도록 반복을 진행한다. 그렇게 레이블 값 차이가 해당 임계값보다 작아지게되면, 전파 반복 단계를 벗어나 마지막 단계인 레이블 할당 단계대신 새로운 감정사전을 제작하게 된다. 모든 한국어 형태소 정점이 마지막으로 지니고 있는 레이블 값을 새로운 감정 점수로 설정하여 감정사전에 저장하는 것이다. 이에 대한 결과로 한국어 형태소 50,068개의 새로운 감정사전이 제작된다.

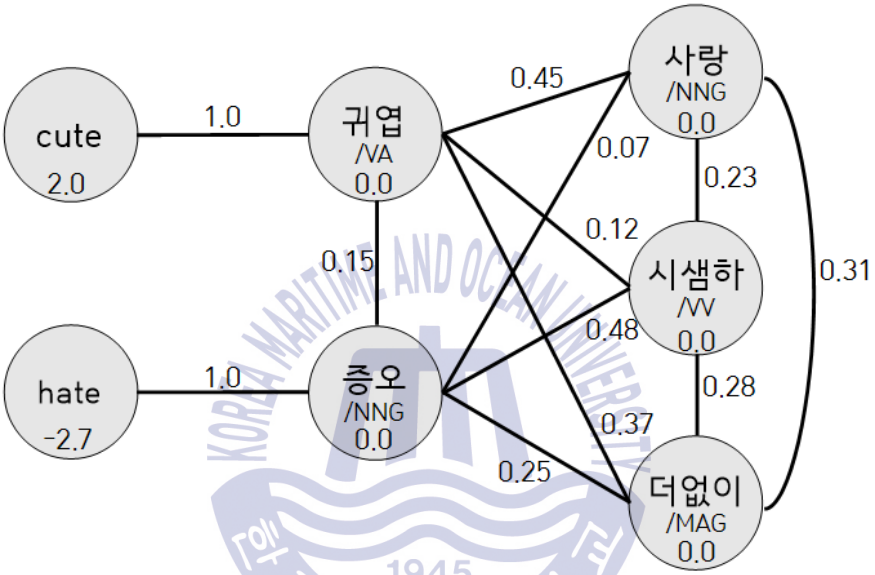
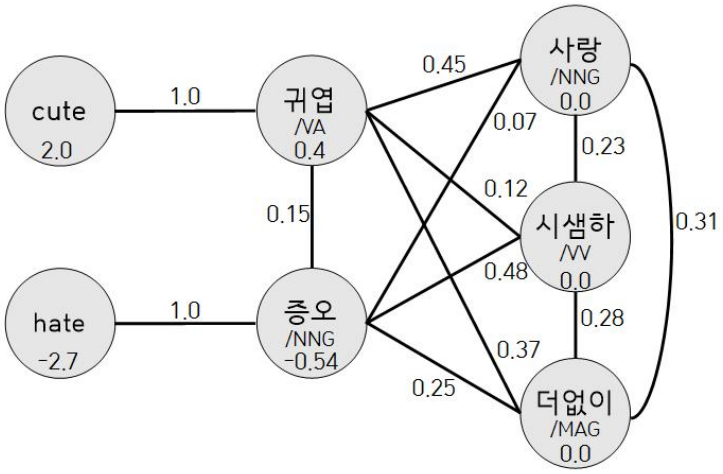
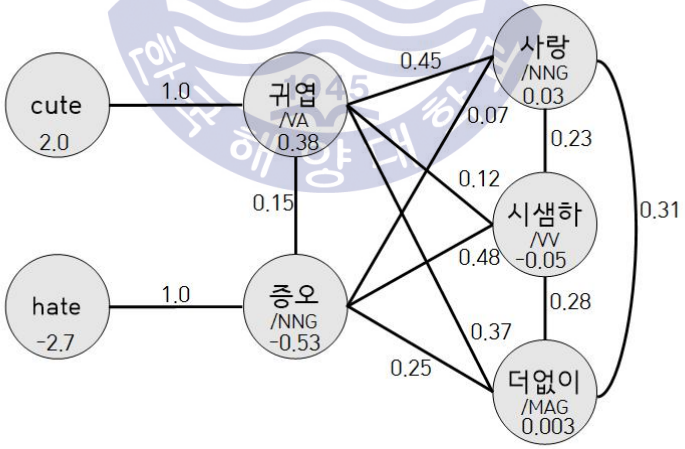


Figure 3.12 The initialization step of the label propagation algorithm on a Korean-English bilingual graph



귀엽/VA	(2.0×1.0 + 0.0×0.15 + 0.0×0.45 + 0.0×0.12 + 0.0×0.37) ÷ 5	0.4
증오/NNG	(-2.7×1.0 + 0.0×0.15 + 0.0×0.07 + 0.0×0.48 + 0.0×0.25) ÷ 5	-0.54
사랑/NNG	(0.0×0.45 + 0.0×0.07 + 0.0×0.23 + 0.0×0.31) ÷ 4	0.0
시샘하/VV	(0.0×0.12 + 0.0×0.48 + 0.0×0.12 + 0.0×0.28) ÷ 4	0.0
더없이/MAG	(0.0×0.37 + 0.0×0.25 + 0.0×0.31 + 0.0×0.28) ÷ 4	0.0

Figure 3.13 After first propagation step on a Korean-English bilingual graph



귀엽/VA	(2.0×1.0 - 0.54×0.15 + 0.0×0.45 + 0.0×0.12 + 0.0×0.37) ÷ 5	0.38
증오/NNG	(-2.7×1.0 + 0.4×0.15 + 0.0×0.07 + 0.0×0.48 + 0.0×0.25) ÷ 5	-0.53
사랑/NNG	(0.4×0.45 - 0.54×0.07 + 0.0×0.23 + 0.0×0.31) ÷ 4	0.03
시샘하/VV	(0.4×0.12 - 0.54×0.48 + 0.0×0.12 + 0.0×0.28) ÷ 4	-0.05
더없이/MAG	(0.4×0.37 - 0.54×0.25 + 0.0×0.31 + 0.0×0.28) ÷ 4	0.003

Figure 3.14 After second propagation step on a Korean-English bilingual graph

제작된 감정사전에 대한 자세한 내용은 Table 3.7~3.8에 나타난다.

Table 3.7 Statistics of the Korean sentiment lexicon by polarity and POS

감정어 수	50,068
긍정 감정어 수	20,759
부정 감정어 수	29,309
명사	38,073
동사	7,923
형용사	2,171
부사	1,557
감탄사	343

Table 3.8 Statistics of the Korean sentiment lexicon by strength of value

구분	긍정	부정
(+-) 0.0이상 1.0미만	2,246	3,646
(+-) 1.0이상 2.0미만	4,586	6,410
(+-) 2.0이상 3.0미만	3,390	4,877
(+-) 3.0이상 4.0미만	10,537	14,376

제 4 장 실험 및 평가

이 장에서는 제작한 감정사전의 제작 과정과 제작된 감정사전에 대한 검증을 진행한다. 제작 과정의 검증 내용으로는 한영 이중언어사전 제작에 있어서 VADER 감정어의 대역어 추출의 기준이 있다. 제작된 감정사전의 검증 내용으로는 KMU 감정 말뭉치(서형원 외, 2010)와 네이버 감정 영화 말뭉치(박은정, 2015)로 두 가지 분야에서 감정사전의 정확성을 검증해본다.

4.1 제작 과정의 발견법적(heuristic) 접근의 검증

이 절에서는 한영 이중언어사전 제작, 한영 이중언어그래프 제작, 감정점수 전파로 이어지는 제작 과정에 있어서 발견법적인 접근을 하였던 과정의 적합성을 검증한다.

한영 이중언어사전 제작 시, 이전 단계에서 제작된 PMI 행렬을 이용하였다. 각 VADER 감정어의 벡터와 유사도가 높게 나오는 상위 10위의 한국어 형태소를 차용하여 이중언어사전을 제작하였다. 이 방법의 적합성을 밝히기 위해 한영사전을 이용해 정답률을 측정해본다. 여기서 정답이란, 해당 영어 단어의 한국어 의미로 사전에 기재되어 있는 것을 말한다. Table 4.1은 상위 1, 5, 10, 20위를 차용할 때, 제작된 한영 이중언어사전에 나타나는 사전적인 정답의 단어 별 비율을 나타낸다.

Table 4.1 The Comparison of correction rate of a Korean-English bilingual lexicon based on extraction limit of cosine similarity

구분	상위 1위	상위 5위	상위 10위	상위 20위
이중언어사전 크기	7,703	13,215	25,679	38,511
정답 수	749	2,246	9,511	15,404
오답 수	6,954	10,969	16,168	23,107
정답률	약 10%	약 17%	약 37%	약 40%

추출하는 범위를 늘릴수록 조금 더 높은 확률로 정답이 존재함을 확인할 수 있다. 그러나 상위 20위는 추출 범위가 두 배로 늘어났음에도 이중언어사전의 크기의 증가율이 감소하였고, 정답률이 10위와 비교할 때 확연히 높아지지 않았다. 물론 정답이란 기준이 사전에 기재된 여부를 따지는 것이므로 Table 4.1에 오답으로 측정된 형태소들도 어느 정도의 의미적 유사성이 존재할 수도 있다. 하지만 정답으로 측정된 형태소들은 VADER 감정어와 쌍으로 한영사전에 등장하는 것이기 때문에 정답의 개수와 정답률로 추출 한계를 지정하는 것이 타당하다고 할 수 있다.

4.2 제작된 감정사전의 검증

이 절에서는 두 분야의 감정분석을 위한 말뭉치를 활용하여 제작된 감정사전의 정확성과 적용 범위를 검증한다. 각각 분야는 뉴스 기사의 댓글과 영화평으로, 뉴스 기사의 댓글을 모아놓은 KMU 감정 말뭉치와 영화평을 모아놓은 네이버 감정 영화 말뭉치를 사용한다. Table 4.2는 두 말뭉치를 자세히 보인다.

Table 4.2 Statistics of KMU sentiment corpus and NAVER sentiment movie corpus

KMU 감정 말뭉치			네이버 감정 영화 말뭉치	
긍정	중립	부정	긍정	부정
788	301	7,739	90,449	92,094

KMU 감정 말뭉치는 뉴스 기사의 댓글 8,828개의 집합으로 긍정적인 댓글은 788개, 중립적인 댓글은 301개, 부정적인 댓글은 7,739개가 존재한

다. 감정의 강도나 점수는 매겨져 있지 않으며 극성만 부여되어 있다. 네이버 감정 영화 말뭉치는 영화평 182,543개의 집합으로 긍정적인 영화평은 90,449개, 부정적인 영화평은 92,094개가 존재한다. 감정의 강도나 점수는 매겨져 있지 않으며 극성만 부여되어 있다. 하지만 (박은정, 2015)에서 밝혔듯, 1에서 10점까지 평가되어있는 영화평에서 1~4점은 부정, 5~8점은 중립, 9~10점은 긍정으로 보고 중립적 영화평은 배제했다.

4.2.1 감정분석 시스템

말뭉치 내 문서들의 감정 점수 도출 방법은 VADER 감정분석 시스템의 방법을 차용한다. Figure 4.1은 VADER 감정분석 시스템이 텍스트의 감정 값을 도출하는 과정을 보인다.

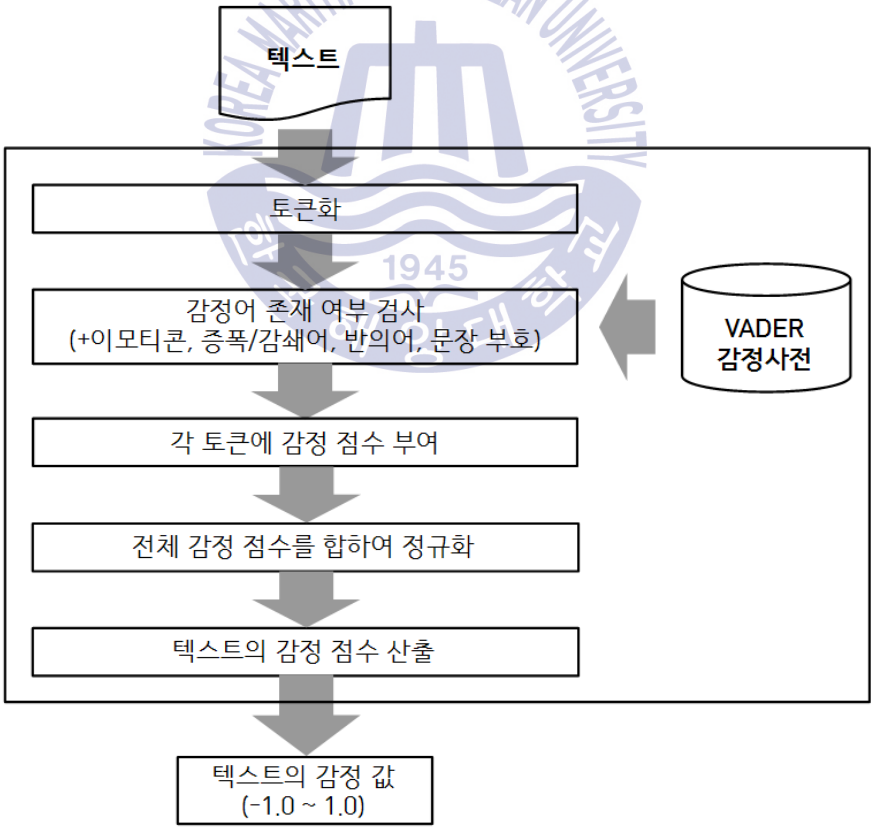


Figure 4.1 The processing of sentiment scoring about text data in the VADER sentiment analysis system

VADER 감정분석 시스템은 영문 텍스트가 입력되면 토큰화 하여 VADER 감정사전 내 등장 여부를 검사한다. 등장하는 단어가 존재할 경우, VADER 감정분석 시스템은 설정된 언어적 규칙을 점수에 적용시킨다. 그렇게 텍스트 내 등장한 각 감정어들은 최종적인 감정값을 부여받고 시스템은 그러한 감정어들의 값을 전부 합친 뒤 정규화 하여 출력한다. 정규화를 위하여 식 (4.1)을 사용한다. s는 텍스트 내 토큰 전체의 감정 점수 합을 뜻하고 n은 정규화된 결과 값을 뜻한다.

$$n = \frac{s}{\sqrt{s^2 + 15}} \tag{4.1}$$

정규화된 감정 점수는 -1.0에서 1.0 사이 값을 지니므로 해당 연구에서는 식 (4.1)의 값이 -1.0 미만일 경우 -1.0으로 출력, 1.0 이상일 경우 1.0으로 출력한다. 따라서 제작한 본 논문의 감정분석 시스템도 동일하게 적용하였다.

Figure 4.2는 본 연구의 말뭉치 내 문서들의 감정 점수 도출을 위한 감정분석 시스템을 보인다. 제안된 감정분석 시스템은 한국어 문서를 입력으로 사용한다. 입력된 문서의 형태소 분석을 진행하여 제작된 감정사전 내 등장 여부를 검사한다. 등장하는 형태소가 존재할 경우, 제안된 감정분석 시스템은 설정된 언어적 규칙을 점수에 적용시킨다. 그렇게 문서 내 등장한 각 감정어들은 최종적인 감정값을 부여받고 시스템은 그러한 감정어들의 값을 전부 합친 뒤 정규화 하여 출력한다. 정규화 공식은 VADER 감정분석 시스템과 같이 식 (4.1)을 사용한다.

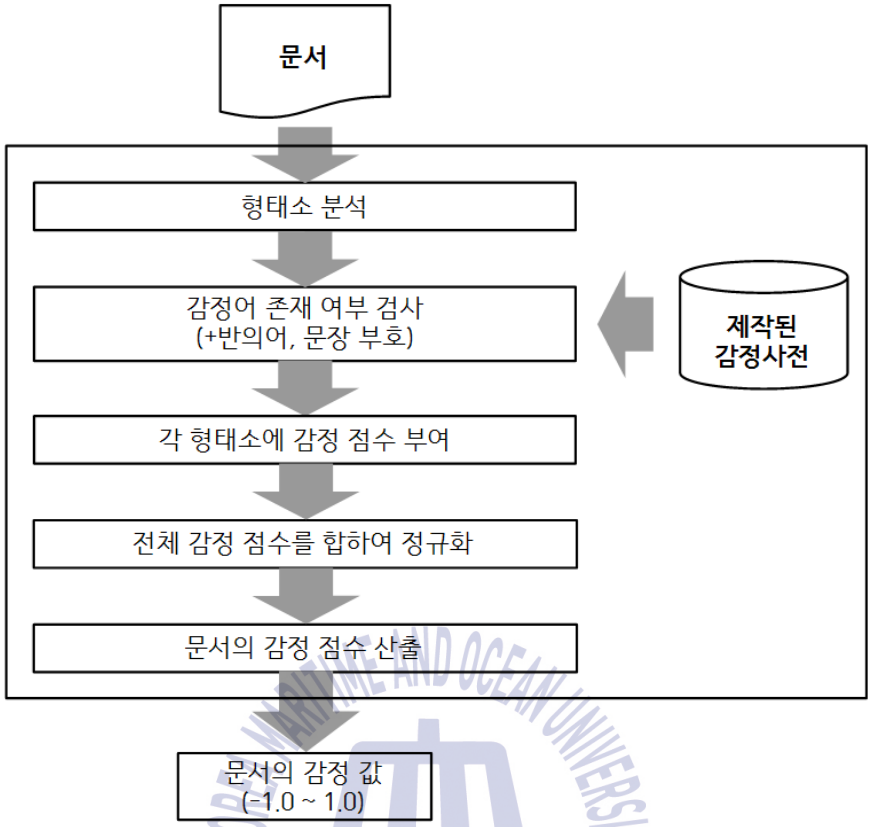


Figure 4.2 The processing of sentiment scoring about documents in the proposed sentiment analysis system

4.2.2 감정분석 시스템을 활용한 감정 말뭉치의 감정분석

제작된 감정사전과 감정분석 시스템을 활용하여 감정 말뭉치 두 개의 감정분석을 진행하였다. KMU 감정 말뭉치는 뉴스 기사 기반의 말뭉치로 각 문서마다 긍정, 부정, 중립 극성이 부여되어 있다. 네이버 감정 영화 말뭉치는 각 영화평마다 긍정, 부정 극성이 부여되어 있다.

(Gilbert & Hutto, 2014)에서는 VADER 감정분석 시스템을 사용하여 소셜 네트워크 서비스, 영화평, 상품평, 뉴스 기사 댓글의 총 네 가지 분야에서 감정분석을 진행하였다. 해당 각 텍스트의 극성을 -1.0~-0.5의 범위를 부정, -0.5~0.5의 범위를 중립, 0.5~1.0의 범위를 긍정으로 분석하여 정

답과 비교하는 방식이다. 이러한 감정 극성에 따른 점수 범위를 KMU 감정 말뭉치에 그대로 적용해 실험을 진행하였다. 그러나 네이버 감정 영화 말뭉치에는 적용하지 않는다. (박은정, 2015)은 1에서 10점까지 평가되어 있는 영화평에서 1~4점은 부정, 5~8점은 중립, 9~10점은 긍정으로 보고 중립적 영화평은 배제하였기 때문이다. 따라서 -1.0~1.0의 정규화된 감정 값을 -1.0~-0.2까지 부정, 0.6~1.0까지 긍정으로 적용하여 실험을 진행한다. Table 4.3은 각 감정 말뭉치의 극성에 따른 정답 범위를 밝힌다.

Table 4.3 The correct answer range of a sentiment score of each polarity

KMU 감정 말뭉치			네이버 감정 영화 말뭉치	
긍정	중립	부정	긍정	부정
0.5~1.0	-0.5~0.5	-0.5~-1.0	0.6~1.0	-1.0~-0.2

본 연구의 실험에서는 제작한 감정사전과 감정분석 시스템을 평가하기 위해 정밀도(precision), 재현율(recall), $F_1\,Score$(Yutaka, 2007), 정확도(accuracy)를 사용한다. 정밀도는 식 (4.2), 재현율은 식 (4.3), $F_1\,Score$는 식 (4.4), 정확도는 식 (4.5)로 계산한다. 계산된 세 가지 지표는 Table 4.4~4.5의 평가결과표와 교차표(confusion matrix)로 나타낸다.

Table 4.4 The table of a evaluation result

긍정		중립		부정		정확도
맞춤	틀림	맞춤	틀림	맞춤	틀림	
a	b	c	d	e	f	

Table 4.5 The confusion matrix

		실험 결과	
		긍정	부정
정답	긍정	g	h
	부정	i	j

$$\text{정밀도} = \frac{g}{g+i} \tag{4.2}$$

$$\text{재현율} = \frac{g}{g+h} \tag{4.3}$$

$$F_1\,Score = \frac{2\times(\text{정밀도}\times\text{재현율})}{(\text{정밀도}+\text{재현율})} \tag{4.4}$$

$$\text{정확도} = \frac{a+c+e}{a+b+c+d+e+f}\times 100 \tag{4.5}$$

제안한 평가 지표들을 사용하여 실험을 진행한다. 한영 이중언어사전을 제작하는 과정에서 추출하는 범위에 따른 실험 결과는 Table 4.6~4.7에 나타난다.

Table 4.6 The accuracy result of KMU sentiment corpus based on a extraction limit of cosine similarity

	상위 1위	상위 5위	상위 10위	상위 20위
정확도	56.84	70.81	80.75	74.16

Table 4.7 The precision, recall and $F_1\,Score$ result of NAVER sentiment movie corpus based on a extraction limit of cosine similarity

	상위 1위	상위 5위	상위 10위	상위 20위
정밀도	0.57	0.62	0.70	0.67
재현율	0.59	0.66	0.75	0.71
$F_1\,Score$	0.58	0.64	0.72	0.69

4.1절의 결과에 근거하여, 한영 이중언어사전의 추출 범위가 늘어날수록 사전적인 정답은 증가하는 반면에 정확도, 정밀도, 재현율, $F_1\,Score$는 상위 10위에서 가장 높은 결과를 보였다. 추출 범위의 증가량 당 평가 수치가 가장 크게 증가한 경우에 정확도에서는 1위에서 5위까지 차용했을 때 13.97%의 증가를 보였고 정밀도와 재현율, $F_1\,Score$ 상에서는 5위에서 10위까지 차용했을 때 가장 크게 증가했다.

다음은 제작한 감정분석 시스템 상에서 언어적 특징을 부여 안했을 때의 결과와 부여했을 때 결과를 비교하였다. 부여한 언어적 특징은 총 세 가지로 첫 번째는 '안, 않, 못, 아니'등과 같은 부정어, 두 번째는 '매우, 너무, 굉장히'등과 같은 증폭/감쇄어, 세 번째는 '!, ?'등과 같은 문장부호이다. 첫 번째의 부정어가 사용되었을 경우, 문장 내 특정 범위에 감정어가 존재한다면 점수를 반대 극성으로 치환한다. 두 번째의 증폭/감쇄어가 사용되었을 경우, 문장 내 특정 범위에 감정어가 존재한다면 점수를 2만큼 더 강하게 해준다. 세 번째의 문장 부호가 사용되었을 경우엔 VADER 감정분석 시스템과 동일하게 적용한다. 예를 들면 증폭/감쇄어와 같이 '!'가 사용된 개수만큼 점수를 더 강하게 해주는 방식이다. Table 4.8~4.9는 해당 결과를 나타낸다.

Table 4.8 The accuracy result of KMU sentiment corpus based on addition of language conventions

	언어적 특징 미추가	언어적 특징 추가
정확도	77.22	80.75

Table 4.9 The precision, recall and $F_1 Score$ result of NAVER sentiment movie corpus based on addition of language conventions

	언어적 특징 미추가	언어적 특징 추가
정밀도	0.65	0.70
재현율	0.68	0.75
$F_1 Score$	0.66	0.72

언어적 특징을 추가함에 따라 정확도는 3.53% 증가하였다. 정밀도와 재현율, $F_1 Score$는 각각 5%, 7%, 6%의 증가를 보였다. 이는 언어적 특징을 반영하여 감정 점수를 부여하는 것이 정확한 감정분석을 위해 필요함을 알 수 있다.

제 5 장 결론 및 향후 연구

어떠한 문서의 감정을 정확하게 읽어내기 위해, 정확하고 방대한 감정 사전의 필요성은 매우 높다. 감정사전을 제작하기 위해서는 크게 사전 기반 제작 방식, 말뭉치 기반 제작 방식, 집단지성 기반 제작 방식이 존재한다. 하지만 각각의 방식은 장단점이 존재한다. 정확한 감정사전 제작을 위해서는 전문가의 검증이 필수적인 반면, 방대한 감정사전 제작을 위해서는 사람의 노력이 적게 들어가거나 아예 포함되지 않는 자동 제작 방법이 불가피하다.

따라서 본 연구에서는 기존에 집단지성 제작 방식으로 제작되어 전문가에게 검증된 VADER 감정사전을 사용했다. VADER 감정사전과 한영 병렬 말뭉치를 활용하여 한영 이중언어사전과 한영 이중언어그래프를 제작하였다. 그렇게 제작된 이중언어그래프 상에서 레이블 전파 알고리즘을 수행하여 새로운 한국어 감정사전을 제작하는 방법을 제안하였다. 제안하는 방법의 핵심은 VADER 감정사전의 영어 감정어와 한국어 형태소 간의 관계를 수치화 한 PMI 행렬을 제작하였다는 것이다. 제작한 PMI 행렬을 통해 생성된 한영 이중언어사전에 등재된 한국어 형태소는 영어에서 한국어로의 감정 점수 전파 단계에서 다리 역할을 하였다. 이러한 과정을 통해 한국어 감정사전을 제작하였고 검증을 위해 두 분야의 감정 말뭉치를 사용하여 감정분석을 진행하였다.

한영 이중언어사전에서 상위 10위까지의 대역어를 추출했을 때와 언어적 특징을 사용했을 때 감정분석의 결과로, 뉴스 기사의 댓글 기반의 1만여 개의 문서를 모은 KMU 말뭉치에서는 최대 80.75%의 정확도를 보였고, 네이버 영화평 댓글 기반의 20만여 개의 댓글을 모은 네이버 감정 영

화 말뭉치에서는 최대 70%의 정밀도, 75%의 재현율, 72%의 $F_1 Score$를 보였다.

언어적 특징은 부정어, 증폭/감쇄어, 문장부호를 사용한 것으로 사용하지 않았을 때보다 최소 3% 이상의 성능 증가를 보였기 때문에 추가적인 언어적 특징을 적용한 연구가 수행되어야 할 것으로 보인다. 본 연구에서 사용된 한국어 언어 단위인 형태소를 넘어서서 단어, 어절, 문장 단위의 언어적 특징을 부여할 수 있을 것이다. 부정어, 증폭/감쇄어, 문장부호 적용의 정확한 대상을 파악하기 위해 의존 구문 분석을 진행하거나, 감정어의 정확한 파악을 위해 개체명 인식 등을 진행할 수 있을 것이다.

감정사전이나 감정분석 시스템에서 다루는 감정 극성의 종류, 점수의 범위 역시 추가적인 연구가 수행되어야 할 것으로 보인다. 본 연구에서는 긍정과 부정의 두 가지 감정 극성에 대해 감정사전을 제작하였으나 단순 긍·부정을 넘어선 인간의 다양한 감정 종류 모델에 따른 다각적인 분석 역시 정확한 감정분석에 필요할 것이라 예상된다. 또한 한국어 단어나 형태소 등 다양한 언어 단위에 따라, 모든 극성에 따른 의미를 수치화할 수 있는 언어 모델 구성 역시 선행되어야 할 과제라고 생각한다.

참고문헌

Barbosa, L., and Feng, J. (2010). “Robust sentiment detection on Twitter from biased and noisy data”, *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. pp. 36-44.

Blair-goldensohn, S. Neylon, T., Hannan, K., Reis, G. A., Mcdonald, R., and Reynar, J., (2008). “Building a sentiment summarizer for local service reviews”, *Proceedings of the WWW Workshop on NLP in the Information Explosion Era.* pp. 339-348.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). “Enriching word vectors with subword information”, *Facebook AI Research*.

Bradley, M. M. and Lang, P. J. (1999). Affective Norms for English Words(ANEW): Instruction Manual and Affective Ratings, *Technical Report C-1, the Center for Research in Psychophysiology, University of Florida.*

Bruce, R. F. and Wiebe, J. M. (1999). “Recognizing subjectivity: A case study in manual tagging”, *Natural Language Engineering.* 5(2):187-205.

Choi, Y., and Cardie, C., (2009). “Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification”, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* pp. 590-598.

Church, K. W. and Hanks, P. (1990). “Word association norms, mutual information, and lexicography”, *Computational Linguistics*. 16(1):22-29.

Ding, X., Liu, B., and Yu, P. S. (2008). “A holistic lexicon-based approach to opinion mining”, *Proceedings of the 2008 International Conference on Web*

Search and Data Mining. pp. 231-240.

Dragut, E. C., Yu, C., Sistla, P., and Meng, W. (2010). "Construction of a sentimental word dictionary", *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. pp. 1761-1764.

Du, W., Tan, S., Cheng, X., and Yun, X. (2010). "Adating information bottleneck method for automatic construction of domain-oriented sentiment lexicon", *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. pp. 111-120.

Edward, L. and Steven, B. (2002). "NLTK: the natural language toolkit", *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 1:63-70.

Esuli, A. and Sebastiani, F. (2005). "Determining the semantic orientation of terms through gloss classification", *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. pp. 617-624.

Esuli, A. and Sebastiani, F. (2006). "SENTIWORDNET: A publicly available lexical resource for opinion mining", *Proceedings of the 5th Conference on Language Resources and Evaluation*. pp. 417-422.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). "Pulse: Mining customer opinions from free text", *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*. pp. 121-132.

George, A. M. (1995). "WordNet: A lexical database for english", *Communications of the ACM*. 38(11):39-41.

Ghiassi, M., Skinner, J., and Zimbra, D. (2013). "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network", *Expert Systems with Applications*. 40(16):6266-6282.

Gilbert, E. and Hutto, C. J. (2014). "VADER : A parsimonious rule-based model for sentiment analysis of social media text", *Eighth International Conference on Weblogs and Social Media*. pp. 216-225.

Harris, Z. S., (1954). “Distributional structure”, *Word.* 10(23):146-162.

Hassan, A., Qazvinian, V., and Radev, D. (2010). “What’s with the attitude? identifying sentences with attitude in online discussion”, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 1245-1255.

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., and Etter, M. (2011). “Good friends, bad news affect and virality in Twitter”, *Future information technology*, Springer, Berlin, Heidelberg. pp. 34-43.

Hatzivassiloglou, V. and McKeown, K. R. (1997). “Predicting the semantic orientation of adjectives”, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* pp. 174-181.

Hatzivassiloglou, V. and Wiebe, J. M. (2000). “Effects of adjective orientation and gradability on sentence subjectivity”, *Proceedings of International Conference on Computational Linguistics (COLING-2000).* pp. 299-305.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M., and McKeown, K. R. (2001). “SIMFINDER: A flexible clustering tool for summarization”, *Proceedings of the NAACL Workshop on Automatic Summarization.* pp. 41-49.

Hu, M. and Liu, B. (2004). “Mining and summarizing customer reviews”, *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 168-177.

Kamps, J., Marx, M., Mokken, R. J., and Rijke, M. D. (2004). “Using WordNet to measure semantic orientation of adjectives”, *Proceedings of LREC-2004.* pp. 1115-1118.

Kanayama, H. and Nasukawa, T. (2006). “Fully automatic lexicon expansion for domain-oriented sentiment analysis”, *Proceedings of the 2006 Conference on*

Empirical Methods in Natural Language Processing. pp. 355-363.

Karger, D. (1993). "Global min-cuts in RNC and other ramifications of a simple mincut algorithm", *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms*. pp. 21-30.

Kim, S. M. and Hovy, E. (2004). "Determining the sentiment of opinions", *Proceedings of the 20th International Conference on Computational Linguistics*. pp. 1367-1373.

Kwon, H. S., Seo, H. W., and Kim, J. H. (2015). "Evaluating a pivot-based approach for bilingual lexicon extraction", *Computational Intelligence and Neuroscience*. 2015:1-13.

Lin, D. (1998). "Automatic retrieval and clustering of similar words", *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. pp. 768-774.

Mohammad, S., Dunne, C., and Dorr, B. (2009). "Generationg high-coverage semantic orientation lexicons from overtly marked words and a thesaurus", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. pp. 599-608.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient estimation of word representations in vector space", *Journal of Artificial Intelligence Research*. 37:141-188.

Mullen, T. and Collier, N. (2004). "Sentiment analysis using support vector machines with diverse information sources", *Proceedings of Conference on Empirical Methods in Natural Language Processing*. pp. 412-418.

Nielsen, F. A. (2011). "A new ANEW: Evolution of a word list for sentiment analysis in microblogs", *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. pp. 33-38.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). "Text classification from labeled and unlabeled documents using EM", *Machine*

learning. 39(2-3):103-134.

Pang, B., and Lee, L. (2004). "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. pp. 271-278.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (1993). Linguistic inquiry and word count, *Technical Report*, Dallas, TX: Southern Methodist University.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). "Linguistic inquiry and word count: LIWC 2001", *Mahway: Lawrence Eribaum Associates*. 71(2001):2001.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). "LIWC 2007: Linguistic inquiry and word count", *Austin, Texas: liwc.net*.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). "Linguistic inquiry and word count: LIWC 2015", Pennebaker Conglomerates.

Rao, D., and Ravichandran, D. (2009). "Semi-supervised polarity lexicon induction", *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 675-682.

Rehurek, R., and Sojka, P. (2010). "Software framework for topic modeling with large corpora", *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. pp. 45-50.

Shah, K., Munshi, N., and Reddy, P. (2013). "Sentiment analysis and opinion mining of microblogs", *University of Illinois at Chicago, Course CS 583 – Data Mining and Text Mining*.

Smedt, T. D., and Daelemans, W. (2012). "Pattern for python", *Journal of Machine Learning Research*. 13(6):2063-2067.

Stoer, M., and Wagner, F., (1997). "A simple min-cut algorithm", *Journal of the ACM*. 44(4):585-591.

Stone, P. J., Bales, R. F. Namenwirth, J. Z., and Ogilvie, D. M. (1962). "The general inquirer: A computer system for content analysis and retrieval based on

the sentence as a unit of information", *Computers in Behavioral Science*. 7:484-498.

Thelwall, M., Buckley, K., Paltoglou, G., and Cai, D. (2010). "Sentiment strength detection in short informal text", *The Journal of the American Society for Information Science and Technology*. 61(12):2544-2558.

Turney, P. D. (2002). "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pp. 417-424.

Turney, P. D., and Littman, M. L., (2003). "Measuring praise and criticism: Inference of semantic orientation from association", *ACM Transactions on Information Systems*. 21(4):315-346.

Warriner, A. B., Kuperman, V., and Brysbaert, M., (2013). "Norms of valence, arousal, and dominance for 13,915 English lemmas", *Behavior Research Methods*. 45(4):1191-1207.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P., (1999). "Development and use of a gold-standard data set for subjectivity classifications", *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. pp. 246-253.

Wiebe, J. M., (2000). "Learning subjective adjectives from corpora". *Proceedings of the Seventeenth Natural Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. pp. 735-740.

Wiebe, J. M., and Riloff, E., (2005). "Creating subjective and objective sentence classifiers from unannotated texts", *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*. pp. 486-497.

Wiebe, J., and Mihalcea, R., (2006). "Word sense and subjectivity", *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. pp.

1065-1072.

Williams, G. K., and Anand, S. S., (2009). "Predicting the polarity strength of adjectives using WordNet", *Proceedings of AAAI International Conference on Weblogs and Social Media*. pp. 346-349.

Xiaojin, Z., and Zoubin, G., (2002). Learning from Labeled and Unlabeled Data from Label Propagation, *Technical Report CMU-CALD-02-107, Carnegie Mellon University*.

Yu, H., and Hatzivassiloglou, V., (2003). "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences", *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. pp. 129-136.

Yutaka, S., (2007). "The truth of the F-measure", *Teach Tutor mater*. 1(5):1-5.

김정호 (2015). 영역별 감성사전 구축을 위한 그래프 기반 방법, 한국항공대학교 컴퓨터공학과 응용소프트웨어 박사학위 논문.

김흥규, 강범모, 홍정하 (2007). "21세기 세종계획 현대국어 기초말뭉치: 성과와 전망", *한글 및 한국어 정보처리 학술대회 발표 논문집*. pp. 311-316.

박은정 (2015). "한국어와 NLTK, Gensim의 만남 (https://www.lucypark.kr/docs/2015-pyconkr/#20)", PyCon Korea 2015.

서형원, 이공주, 류길수, 김재훈 (2010). "뉴스 댓글의 감정 분류를 위한 자질 가중치 설정", *한국마린엔지니어링 학회지*. 34(6):871-879.

신수정 (2014). "글에서 감정을 읽다 감정 분석의 이해", IDG Korea IT World.

신준철, 옥철영 (2012). "기분석 부분 어절 사전을 활용한 한국어 형태소 분석기", *정보과학회 논문지, 소프트웨어 및 응용*. 39(5):415-424.

안정국, 김희웅 (2015). "집단지성을 이용한 한글 감성어 사전 구축", *지능정보연구 학회지*. 21(2):49-67.

안주영, 배정환, 한남기, 송민 (2015). "텍스트 마이닝을 이용한 감정 유발 요인 'Emotional Trigger'에 대한 연구", *지능정보연구회 논문지*.

21(2):69-92.

한국정보통신기술협회 (2012). TTA정보통신용어사전, http://terms.tta.or.kr.

허찬, 온승엽 (2017). “Word2vec와 Label Propagation을 이용한 감성사전 구축 방법”, *한국차세대컴퓨팅학회 논문지*. 13(2):93-101.

감사의 글

2019년 2월, 2년간의 석사 과정을 마무리하며, 창밖의 바다만큼 무수히 펼쳐진 감사함을 여기에 올립니다.

제 학업과 연구를 진행하며 수없이 많았던 크고 작은 벽들을 이렇게 무사히 넘도록 지도하고 아껴주신 김재훈 교수님께 감사드립니다. 교수님의 나침반과 같은 조언과 가르침을 통해 본 논문이 잘 마무리 될 수 있었습니다. 앞으로도 많은 가르침으로 지도 부탁드립니다.

본 논문의 심사를 맡아주신 박휴찬 교수님과 이장세 교수님께도 감사드립니다. 두 분의 사려깊은 배려로 많은 심사를 거쳐 본 논문의 완성도를 높일 수 있었습니다. 중요하고 뜻깊은 경험이었습니다. 학업을 이어가며 큰 밑거름으로 삼겠습니다.

자연언어처리실험실의 제 동료들에게도 감사를 전합니다. 때로는 친근한 동기이자 든든한 선배이신 천민아님, 어느새 실험실의 빛과 소금이자 기둥이 되어있는 윤호님, 모든 일에 신중하고 열정을 쏟으시는 남궁영님, 자기자신보다 늘 남을 챙기며 이제는 친동생같은 최민석님, 막내이지만 누구보다도 믿음직하고 항상 긍정적인 김재균님. 보기만해도 언제나 즐겁습니다. 진심어린 감사를 전합니다.

한국해양대학교 컴퓨터공학과 동료들과 신도고등학교 동창들에게도 감사를 전합니다. 당연한 듯 저에게 다가와 물들이듯 아름다운 추억을 나누고, 지금은 각자의 삶을 별처럼 살아가는 여러분이 저의 자랑입니다. 모두들 본연의 빛을 세상에 비추기 위해 바쁘지만 힘차게 살아가고 있습니다. 어떤 별은 멀어 손이 닿기 어렵지만, 그 빛만은 저에게 늘 태양처럼 세상

을 비춰주고 있습니다. 감사합니다.

“심연속의 진주” 동아리 여러분들에게도 감사를 전합니다. 정상을 향했던 시절부터 저와 함께해주신 고마운 오랜 친구분들, 서리어린 들판 위를 불꽃처럼 뜨겁게 달리는 늑대같은 동료들. 여러분과 함께하는 이 즐거운 시간이 언제까지고 계속되길 바랍니다.

마지막으로 제 가족들에게 감사를 전합니다. 제 평생동안 무한한 사랑과 배려, 지원을 아끼지 않아주신 분들. 언제나 부족한 저를 믿어주시고 응원해주셔서 감사합니다. 여러분이 곧 ‘저’이자 저의 자랑입니다. 항상 건강하시고 저로인해 많이 웃으시고 행복하셨으면 좋겠습니다. 이번 해에 대학에 진학하는 사랑스런 동생, 선유에게도 그동안 고생한 만큼 기쁜 일이 가득한 대학 생활이 되길 기원합니다.

어느덧 떠오른 해가 어두웠던 하늘과 바다에 스며듭니다. 미처 다 표현하지 못한 감사함을 되새기며, 제 자신은 또 하나의 물방울이 되어 새로운 바다를 향해 흘러갑니다.

- 박호민 올림 -