

Bi-LSTM 기반의 한국어 감성사전 구축 방안*

박상민

군산대학교 소프트웨어융합공학과
(b1162@kunsan.ac.kr)

나철원

군산대학교 소프트웨어융합공학과
(ncw0034@kunsan.ac.kr)

최민성

군산대학교 소프트웨어융합공학과
(alstjd517@kunsan.ac.kr)

이다희

군산대학교 소프트웨어융합공학과
(dahee@kunsan.ac.kr)

온병원

군산대학교 소프트웨어융합공학과
(bwon@kunsan.ac.kr)

.....

감성사전은 감성 어휘에 대한 사전으로 감성 분석(Sentiment Analysis)을 위한 기초 자료로 활용된다. 이와 같은 감성사전을 구성하는 감성 어휘는 특정 도메인에 따라 감성의 종류나 정도가 달라질 수 있다. 예를 들면, ‘슬프다’라는 감성 어휘는 일반적으로 부정의 의미를 나타내지만 영화 도메인에 적용되었을 경우 부정의 의미를 나타내지 않는다. 그렇기 때문에 정확한 감성 분석을 수행하기 위해서는 특정 도메인에 알맞은 감성사전을 구축하는 것이 중요하다. 최근 특정 도메인에 알맞은 감성사전을 구축하기 위해 범용 감성 사전인 오픈한글, SentiWordNet 등을 활용한 연구가 진행되어 왔으나 오픈한글은 현재 서비스가 종료되어 활용이 불가능하며, SentiWordNet은 번역 간에 한국 감성 어휘들의 특징이 잘 반영되지 않는다는 문제점으로 인해 특정 도메인의 감성사전 구축을 위한 기초 자료로써 제약이 존재한다. 이 논문에서는 기존의 범용 감성사전의 문제점을 해결하기 위해 한국어 기반의 새로운 범용 감성사전을 구축하고 이를 KNU 한국어 감성사전이라 명명한다. KNU 한국어 감성사전은 표준국어대사전의 뜻풀이의 감성을 Bi-LSTM을 활용하여 89.45%의 정확도로 분류하였으며 긍정으로 분류된 뜻풀이에서는 긍정에 대한 감성 어휘를, 부정으로 분류된 뜻풀이에서는 부정에 대한 감성 어휘를 1-gram, 2-gram, 어구 그리고 문형 등 다양한 형태로 추출한다. 또한 다양한 외부 소스(SentiWordNet, SenticNet, 감정동사, 감성사전0603)를 활용하여 감성 어휘를 확장하였으며 온라인 텍스트 데이터에서 사용되는 신조어, 이모티콘에 대한 감성 어휘도 포함하고 있다. 이 논문에서 구축한 KNU 한국어 감성사전은 특정 도메인에 영향을 받지 않는 14,843개의 감성 어휘로 구성되어 있으며 특정 도메인에 대한 감성사전을 효율적이고 빠르게 구축하기 위한 기초 자료로 활용될 수 있다. 또한 딥러닝의 성능을 높이기 위한 입력 자료로써 활용될 수 있으며, 기본적인 감성 분석의 수행이나 기계 학습을 위한 대량의 학습 데이터 세트를 빠르게 구축에 활용될 수 있다.

주제어 : Sentiment Lexicon, Sentiment Analysis, Deep Learning, Text Mining, Bi-LSTM

.....

논문접수일 : 2018년 9월 17일 논문수정일 : 2018년 12월 18일 게재확정일 : 2018년 12월 24일

원고유형 : 일반논문 교신저자 : 온병원

* 이 논문은 2016년 정부(미래창조과학부)의 재원으로 한국연구재단의 중견연구자지원사업의 지원을 받아 수행된 연구임(No. NRF-2016R1A2B1014843)

이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068188)

1. 서론

비즈니스 인텔리전스(Business Intelligence) 책임자들은 갈수록 확대되고 있는 정보 자산을 활용할 수 있어야 하고 기업들은 조직 내·외부의 다양한 소스에서 발생하는 데이터에서 실행 가능한 통찰력을 발견하여 이를 의사 결정에 활용하려는 바람을 가지고 있다고 가트너는 언급하였다. 특히 빅데이터에 대한 시장의 관심이 큰 가운데, 새로운 데이터 소스로부터 통찰력을 추출, 활용하고 이를 기반으로 행동하는 것의 잠재적 가치가 증대되었다고 언급하였다(CIO, 2013). 이와 같이 최근 빅데이터에서 통찰력을 추출하고 활용하는 다양한 연구들이 진행되고 있으며 반정형 또는 비정형 텍스트 데이터를 분석하여 의미있는 정보를 발견하는 텍스트 마이닝(Text Mining)의 중요성이 대두되고 있다(Wikipedia, 2018e). 텍스트 마이닝 기법 중 하나인 감성 분석(Sentiment Analysis)은 최근 다양한 분야에서 활발한 연구가 진행되고 있다(Wikipedia 2018d). 대표적으로 사용자들이 게시한 텍스트 데이터의 감성을 분석하여 여론 조사를 수행하는 연구가 진행되고 있으며, 제품에 대한 사용자들의 후기 데이터 분석을 통해 제품에 대한 사용자들의 평판을 분석, 이를 마케팅에 활용하는 연구가 활발히 진행되고 있다(Park S. M., 2017). 감성 분석의 가장 기초적인 방법은 긍정, 부정, 중립으로 구축된 감성 어휘 목록을 통해 어휘 수준의 감성 분석을 수행하는 것이며 이와 같은 감성 어휘 목록을 감성사전(Sentiment Lexicon)이라고 한다. 감성 어휘는 특정 도메인에 따라 감성의 종류나 정도가 달라지는 특징을 가지고 있다. 예를 들면, ‘슬프다’라는 감성 어휘는 일반적으로 부정의 감성을 지니고 있지만 영화 도메인에서의 ‘슬프다’

라는 감성 어휘는 부정의 감성으로 사용되지 않는다. 이와 같이 특정 도메인에 대한 감성 어휘의 성향이 고려되지 않은 감성 사전을 감성 분석에 활용하는 경우 정확한 분석이 수행되기 어렵다. 그렇기 때문에 대부분의 감성사전은 감성 분석을 하고자 하는 대상 텍스트 데이터를 기준으로 자체적으로 구축하여 활용한다. 하지만 특정 도메인에 대한 감성사전을 자체적으로 구축하게 되면 시간이 많이 소요될 뿐만 아니라 기준이 되는 감성 어휘 없이 감성사전을 구축할 경우 다양한 감성 어휘가 포함될 수 없으며 주관적인 감성 어휘들이 추출될 수 있다는 문제점이 존재한다. 최근 이와 같은 문제를 해결하기 위해 오픈 한글에서 제공하는 한국어 범용 사전을 확장하고 특정 도메인에 대한 감성사전을 구축하는 연구가 진행되었으나 현재 오픈 한글의 서비스 중단으로 이를 활용하기 어렵다. 영어 감성 어휘 사전인 SentiWordNet이나 SenticNet을 기초 자료로 활용하여 특정 도메인에 대한 감성사전을 구축하는 연구가 수행되었다. 하지만 SentiWordNet이나 SenticNet은 한국어로 번역할 경우 한국 감성 어휘의 특징이 잘 반영되지 않는 문제가 존재하여 특정 도메인의 감성사전 구축을 위한 기초 자료로 활용하는데 제약이 발생한다.

이 논문에서는 기존의 범용 감성 사전이 갖는 문제점을 해결하기 위해 새로운 범용 한국어 감성사전을 구축하고 이를 ‘KNU 한국어 감성사전’이라고 명명한다. KNU 한국어 감성사전은 특정 도메인에 대한 감성사전을 빠르고 효율적으로 구축하기 위해 구축되었으며, 특정 도메인에 영향을 받지 않는 도메인에 독립적인 감성 어휘로 구성되어 있다. 예를 들면, 인간의 보편적인 기본 감정 표현을 나타내는 긍·부정 어휘인 ‘감동받다’, ‘감사하다’, ‘가치있다’, ‘그저 그렇다’

등이 있다. KNU 한국어 감성사전은 표준국어대 사전에 수록된 뜻을풀이를 활용하여 감성 어휘를 추출하고 구축한다(NIKL, 2018). 뜻을풀이를 활용한 감성 어휘 추출 방안은 다음과 같다. 첫째, 뜻을풀이를 분류하기 위한 뜻을풀이 감성 분류 모델을 Bi-LSTM(Bidirectional Long-Short Term Memory) 기법을 사용하여 구축한다. 둘째, 구축한 뜻을풀이 감성 분류 모델을 통해 뜻을풀이를 긍정, 부정으로 분류한다. 셋째, 긍정으로 분류된 뜻을풀이에서는 긍정에 관련된 감성 어휘를, 부정으로 분류된 뜻을풀이에서는 부정에 관련된 감성 어휘를 추출한다. 이 논문에서 제안한 뜻을풀이 감성 분류 모델은 89.45%의 정확도로 뜻을풀이의 감성을 분류하였다. 이 외에도 다양한 외부 소스를 활용하여 감성 어휘를 확장하였으며 온라인 텍스트 데이터에서 사용되는 신조어, 이모티콘 감성 정보를 추가하였다.

KNU 한국어 감성사전은 1-gram, 2-gram, 어구(n-gram) 그리고 문형 등 다양한 형태의 14,843개의 감성 어휘로 구성되어 있다. 기존에 구축된 감성사전들과 다르게 도메인에 영향을 받지 않는 단어들로 구성 되었으며, 다양한 형태의 어휘 정보를 가진다는 점에서 차별성을 둘 수 있다.

최근 감성사전을 활용하지 않고 딥러닝 기법을 통해 감성 분석을 수행하는 연구가 유의미한 결과를 보였고 이와 같은 이유로 일부 연구자들은 감성사전 구축 및 활용에 대하여 의문을 가지고 있다. 하지만 최근 딥러닝과 감성사전을 동시에 활용한 감성 분석 연구에 따르면 감성사전의 감성 어휘를 딥러닝 입력의 자질로 활용할 경우 더 높은 정확도로 감성 분석을 수행한다는 결과가 도출되었다(Teng, Z., 2016). 이와 같은 연구 결과를 통해 감성사전은 단순히 감성 분석을 위해 활용되는 것 외에도 감성 분석의 성능을 높이기

위한 딥러닝 입력의 자질로 활용될 수 있는 점에서 그 중요성을 높이 평가할 수 있다.

이 논문에서 제안한 KNU 한국어 감성사전 또한 특정 도메인의 감성사전 구축을 위한 기초 자료로 활용되는 것 외에도 딥러닝을 통한 감성 분석에 있어 입력의 자질로 활용될 수 있다. 또한 기본적인 감성 분석의 수행이나 기계 학습을 위한 대량의 학습 데이터 세트를 빠르게 구축하는데 활용될 수 있다.

이 논문의 구성은 다음과 같다. 2장에서는 이 논문의 기초가 되거나 관련 있는 연구를 정리하여 소개한다. 3장에서는 KNU 한국어 감성사전 구축 방안에 대하여 자세히 설명한다. 4장에서는 실험 환경 및 결과에 대해 자세히 논의 하고, 5장에서는 결론 및 향후 연구의 방향을 다룬다.

2. 관련 연구

2.1 기존의 감성사전

기존에 구축되어 있는 한국어 기반의 감성사전은 다음과 같다. DecoSelex라는 한국어 감성사전은 오피니언 마이닝(Opinion Mining)을 위해 구축한 감성사전으로 SentiWordNet을 통하여 감성 어휘를 추출하여 한국어 Deco 사전을 통해 확장하는 방식으로 감성사전을 구축하였다(Shin et al., 2016). 하지만 현재 구축된 DecoSelex 감성사전은 제공되고 있지 않다.

오픈 한글은 단어에 대한 원형과 품사 그리고 감성에 대한 정보를 제공해주는 오픈 서비스이다(An et al., 2015 ; OpenHangul, 2018). 감성사전에 수록된 단어는 집단지성의 참여자가 긍정, 부정, 중립에 대하여 투표하고 누적됨에 따라 신뢰

도가 높아지도록 설계하였다. 하지만 오픈 한글은 현재 오픈 서비스의 한계적인 문제로 인하여 잠정적으로 서비스가 중단되어 사용할 수 없다.

서울대에서는 KOSAC 말뭉치를 사용하여 한국어 감정 어휘 목록(감성사전)을 구축하였다. 이 감정 어휘 목록은 한국어 감정 분석 연구에 폭 넓게 활용할 수 있도록 형태소 단위의 감정 특성을 제공한 어휘 목록이다(Shin et al., 2016 ; KOSAC, 2018). 하지만 이 연구에서 구축한 감정 어휘 목록은 도메인에 따라 감정의 정도가 달라질 수 있는 어휘에 대하여 고려하지 않은 채 감정의 정도를 부여하였다.

K-LICW는 글의 언어학, 심리학적 특징을 분석하고자 개발된 한국어 글 분석 프로그램이다. K-LIWC는 기존에 구축된 단어를 기반으로 글을 분석하여 해당 글의 언어적 특징을 분석할 수 있다(Lee et al., 2005). 하지만 현재 K-LIWC는 제공되고 있지 않아 활용하기 어렵다.

감정과 감정동사에 대한 정의를 내리고 474개의 감정동사 목록을 제시, 해당 감정동사들의 특징을 정리한 연구가 수행되었다(Kim, 2004).

SentiWordNet은 워드넷(WordNet)의 synset이라는 유의어 집단에 포함되어 있는 단어들을 유의어, 반의어 관계를 통해 확장하고 이를 분류기로 학습하여 긍정, 부정, 객관성에 대한 값을 부여한 감성사전이다(Baccianella et al., 2010). SentiWordNet은 단순히 단어 관계 확장과 분류기 학습을 통해 감성 정도가 부여되었기 때문에 일부 어휘에 대한 감성 정도가 올바르지 않다. 예를 들면, ‘연뇌막’이라는 단어는 긍정적인 감성을 가지고, ‘비난하다’라는 단어는 객관적인 감성을 가진다. 또한 부정 감성 어휘 계산에 큰 영향을 미치는 요소를 충분히 고려하지 않아 부정확한 감성 정도를 제공한다. 또

한 SentiWordNet의 감성 어휘를 한국어로 번역하여 사용할 경우 다음과 같은 문제점이 발생한다. 첫째, 한국어에서 단어인 것이 영어에서는 구로 존재한다. 대표적으로 ‘신물나다’라는 단어는 ‘sick of’라는 구로 존재한다. 둘째, 두 언어의 감성 정도 값이 일치하지 않는다. 한국어에서 ‘노발대발하다’라는 단어의 감성 정도는 7.7점이지만 이와 동일한 의미인 ‘infuriate’라는 단어는 2.5점이다. 마지막으로 한국어에서는 서로 다른 어휘로 존재하는 어휘들이 영어로는 같은 어휘로 대역된다. 예를 들면, ‘역정나다’, ‘성질나다’, ‘화나다’, ‘노하다’, ‘분하다’, ‘성나다’, ‘약오르다’, ‘꼴나다’라는 단어는 모두 ‘angry’라는 하나의 단어로 번역된다.

2.2 감성사전 구축 방안

최근 한국어 감성사전 구축을 위한 다양한 연구가 진행되었다. 분석하고자 하는 문서에 최적화된 감성사전을 구축하기 위해 키워드를 선정하고 Word2Vec을 활용하여 후보 키워드를 추출, 추출된 후보 키워드를 통해 긍·부정 감성 어휘를 추출하여 감성사전을 구축하는 연구가 수행되었다(Jang et al., 2017).

특정 도메인을 잘 대표하는 감성사전을 구축하기 위해 회귀 분석 기법인 엘라스틱 넷(ElasticNet)을 사용하여 각 단어의 회귀 계수를 구하고 이를 활용하여 감성사전을 자동으로 구축한 연구가 수행되었다(Kim et al., 2015).

언어의 기본 단위인 단어에 대하여 감정 정보를 부여하고 이를 통해 한국어 감정 어휘 사전을 구축하는 방안을 제안하는 연구가 수행되었다. 이 연구에서는 기초 감정 어휘에 대해 설문을 하여 정도 값을 구하고, 나머지 감정 어휘에 대해

여 사전의 표제어 설명부(Gloss)를 이용해 정도 값을 추론하였다(Choi et al., 2014).

개인이 그 동안 겪어온 경험 등에 근거하여 자신의 생애에 대한 주관적인 평가를 주관적 웰빙 상태라고 한다. 이러한 주관적 웰빙 상태를 측정하기 위해 SentiWordNet을 번역하여 기본 감정 어휘 사전을 구축하였고, 온라인 뉴스 기사의 댓글을 수집하여 댓글의 긍·부정성 파악에 도움이 되는 감정 어휘를 추출하여 상황적 감정어 목록을 구축, 이를 통해 주관적 웰빙 상태를 측정하는 연구가 수행되었다(Choi et al., 2016).

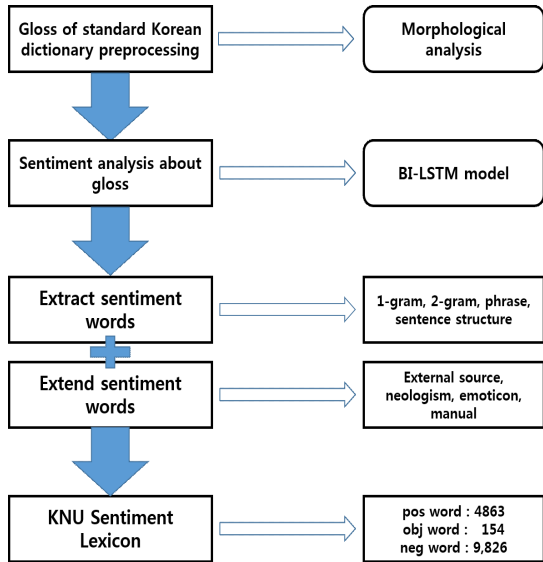
오픈 한글 서비스에서 제공하는 범용 감성사전을 통해 도메인의 특징을 나타내지 않는 단어들을 제외하고 특정 도메인에 대한 단어들의 중요도를 빈도수를 통해 측정하여 도메인 감정 어휘 목록을 구축, 이를 통해 도메인 감성지수를 산출하여 도메인 감정 어휘를 활용한 도메인 맞춤형 감성사전을 구축하는 연구가 수행되었다(Kim et al., 2015). 하지만 이 연구에서 활용하였던 범용 감성사전은 오픈 한글의 서비스 중단으로 현재는 활용할 수 없다.

감성 분석의 성능을 향상시키기 위해 데이터 특성에 맞는 맞춤형 감성사전 구축을 위한 연구가 수행되었다. 영화의 장르에 따라 감정 어휘가 차이가 나는 영화 리뷰 데이터를 대상으로 형용사만을 추출하고 PMI(Pointwise Mutual Information)를 활용하여 감성사전을 구축하였다(Lee et al., 2016). 하지만 단순히 형용사에 대한 어휘만을 통해 감성사전을 구축하였다는 점에서 풍부한 감정 어휘를 지니는 감성사전이라고 볼 수 없다.

3. 제안 방안

그림 1은 이 논문에서 제안한 KNU 한국어 감성사전 구축 알고리즘을 도식화하여 보여준다. KNU 한국어 감성사전 구축을 위해 국립국어원에서 발행하는 표준국어대사전을 구성하는 모든 단어와 뜻풀이(Gloss)를 수집하고 정제한다(Park, 2017).

수집된 모든 단어에 대하여 형태소 분석을 수행하고 형용사, 부사, 동사, 명사를 품사로 하는 단어의 뜻풀이를 감성사전 구축을 위해 추출한다. 이 논문에서는 뜻풀이가 긍정을 나타내면 해당 뜻풀이에서는 긍정에 관한 감정 어휘를, 부정을 나타내면 해당 뜻풀이에서는 부정에 관한 감정 어휘를 추출한다. 뜻풀이의 감성 분류를 위해 딥러닝 기법 중 하나인 Bidirectional Long Short Term Memory(Bi-LSTM)을 사용하여 뜻풀이 감성 분류 모델을 구축한다. 뜻풀이 감성 분류 모델을 통해 뜻풀이를 긍·부정으로 분류하고 분류된 각각의 뜻풀이를 활용하여 1-gram, 2-gram, 어구, 문형에 해당하는 감정 어휘를 수작업으로 추출한다. 감정 어휘를 추출한 후 3명의 투표자는 각각의 감정 어휘의 감성 정도와 도메인에 독립적인 감정 어휘에 대한 여부를 판단한다. 감정 어휘 목록을 확장하기 위해 표준국어대사전에서 추출한 감정 어휘 이외에도 감정 동사 목록, SentiWordNet, SenticNet, 감정단어사전0603(Kim, 2015), 신조어(Wikipedia, 2018c), 이모티콘(Wikipedia, 2018b) 그리고 수작업으로 수집한 감정 어휘들을 추가하여 KNU 한국어 감성사전을 확장한다.



〈Figure 1〉 Flow chart of the proposal 'KNU Sentiment Lexicon' construct algorithm

3.1 감성 어휘 추출 방안 : 표준국어대사전 뜻풀이

이 논문에서는 표준국어대사전의 뜻풀이를 활용한 감성 어휘 추출을 제안한다. 뜻풀이가 감성 어휘 추출을 위한 자원으로 활용된 이유는 다음과 같다. 첫째, 뜻풀이는 단어의 의미를 표현하는 것이 목적이다. 이는 단어가 긍정의 의미를 갖는 경우 뜻풀이는 긍정을 표현하는 어휘들로 구성된다. 이를 통해 긍정의 의미를 표현하는 뜻풀이에서는 긍정과 관련된 감성 어휘들을 추출할 수 있다. 둘째, 뜻풀이는 문장으로 이루어져 있으며 1-gram, 2-gram, 어구, 문형 등으로 구성되어 있다. 이를 통하여 다양한 형태의 감성 어휘를 추출할 수 있다. 셋째, 뜻풀이는 단어의 의미를 설명하는 것이 목적이기 때문에 기본적인 어휘로 구성되어 있다. 시, 수필 등을 제외한 일

반적인 경우 대부분 감성이 들어있는 문장을 작성 또는 표현할 때 기본적인 어휘를 주로 사용한다. 이를 다르게 해석하면, 뜻풀이에 포함되어 있는 감성 어휘는 기본적인 어휘이기 때문에 실제 사용되는 감성 어휘 대부분을 포함한다.

이 절에서는 표준국어대사전을 통한 KNU 한국어 감성사전 구축 방안에 대하여 설명한다.

3.1.1 표준국어대사전

표준국어대사전은 국립국어원에서 발행하는 한국어 사전이며 511,160(+)개의 어휘가 수록되어 있다. 표준국어대사전은 각 단어에 대해 품사, 뜻풀이, 활용 정보 등에 대하여 제공한다.

이 논문에서는 표준국어대사전에 수록되어 있는 단어들 중 형용사, 부사, 동사, 명사를 품사로 갖는 모든 단어들과 뜻풀이를 수집하여 감성 어휘를 추출한다.

3.1.2 뜻풀이 감성 분류

표준국어대사전에서 수록된 단어들은 해당 단어의 의미를 설명하는 문장을 가지고 있으며 이를 '뜻풀이'라고 한다. 이 논문에서는 KNU 한국어 감성사전 구축을 위해 뜻풀이를 활용한다.

각 단어는 쓰임에 따라 다양한 감성과 의미를 가질 수 있기 때문에 하나의 단어에 대한 뜻풀이는 하나 이상이 존재한다. 표 1은 '좋아하다'라는 단어에 대한 뜻풀이다. '좋아하다'라는 단어의 뜻풀이는 4가지 뜻풀이로 구성되어 있다. 뜻풀이 1, 2, 3은 우리가 일반적으로 알고 있는 긍정적인 감성을 나타내는 '좋아하다'로써 긍정적인 감성을 표현하는 어휘는 '좋은 느낌', '잘 먹거나', '잘 마시다', '즐겁게' 등이 있다. 뜻풀이 4는 부정적인 감성을 나타내는 '좋아하다'로써 부정적

〈Table 1〉 Sample of gloss

gloss	좋아하다
1	어떤 일이나 사물 따위에 대하여 좋은 느낌을 가지다. (pos)
2	특정한 음식따위를 특별히 잘 먹거나 마시다. (pos)
3	특정한 운동이나 놀이, 행동 따위를 즐겁게 하거나 하고 싶어하다. (pos)
4	남의 어리석은 말이나 행동을 비웃거나 빈정거릴 때 하는 말. (neg)

인 감성을 표현하는 어휘는 ‘어리석은’, ‘비웃거나’, ‘빈정거릴때’ 등이 있다. 그렇기 때문에 ‘좋아하다’라는 단어 대한 뜻풀이를 일반적으로 우리가 알고 있는 긍정으로 모두 분류하게 된다면 ‘좋아하다’라는 단어에 포함되어 있는 부정적인 뜻풀이 또한 긍정의 감성으로 분류되는 문제가 존재하게 된다. 이와 같은 문제를 해결하기 위해 단어의 감성을 기준으로 모든 뜻풀이의 감성을 분류하지 않고 각각의 뜻풀이에 대한 감성을 확인 후 분류를 수행한다. 각각의 뜻풀이에 대하여 감성 분류를 수행한 후 감성이 분류된 뜻풀이를 활용하여 긍정의 의미를 지니는 뜻풀이에서는 긍정에 관한 감성 어휘를, 부정의 의미를 지니는 뜻풀이에서는 부정에 관한 감성 어휘를 추출한다. 예를 들면, ‘좋아하다’라는 뜻풀이에 의한 긍정 감성 어휘는 뜻풀이 1, 2, 3에 의해 ‘좋은’, ‘좋은 느낌’, ‘잘 먹거나’, ‘잘 마시다’, ‘즐겁게’ 등이 추출될 수 있다. ‘좋은(좋아하다)’이라는 단어는 뜻풀이의 의미가 긍정과 부정의 감성 두 가지를 모두 갖지만 긍정의 감성 어휘로써 추출한다. 이와 같은 이유는 긍정의 감성을 나타내는 뜻풀이에서만 ‘좋은(좋아하다)’이라는 단어가 사용되었고, 이를 통해 ‘좋은(좋아하다)’의 기본 감성은

긍정을 나타낸다고 할 수 있기 때문이다. 부정 감성 어휘는 뜻풀이 4에 의해 ‘어리석은’, ‘비웃거나’, ‘빈정거릴때’ 등이 추출될 수 있다.

뜻풀이 감성 분류를 위한 학습 데이터 구축은 3.1.3절에서, 감성 분류를 위한 제안 방안은 3.1.4절에서 소개한다.

3.1.3 학습 데이터 구축

뜻풀이 감성 분류 모델 구축을 위해 분류 문제 해결에 있어 높은 성능을 보이는 딥러닝 기법을 사용한다. 딥러닝은 학습 데이터를 통해 모델을 학습시켜 모델을 구축하고 이를 통해 새로운 데이터를 분류하는 기법이다(Wikipedia, 2018a). 뜻풀이 감성 분류 모델을 학습시키기 위한 학습 데이터의 구축은 다음과 같이 이루어진다. 첫째, 형용사, 부사 그리고 일부 동사를 품사로 하는 단어들의 뜻풀이를 추출한다. 둘째, 추출된 뜻풀이들에 대하여 3명의 투표자가 각 뜻풀이에 대한 감성을 측정한다. 뜻풀이의 감성은 표 2와 같이 표기한다. 셋째, 3명의 투표자가 뜻풀이에 대하여 측정한 감성이 모두 일치하는 경우 해당 감성을 뜻풀이의 감성으로 부여한다. 3명의 감성이 모두 일치하지 않는 경우에는 3명의 투표자가 토론을 하여 감성이 만장일치가 될 경우에만 감성을 부여한다. 만장일치가 되지 않는 경우에는 해당 뜻풀이를 학습 데이터에서 제외시킨다. 구축된 학습 데이터는 표 3과 같이 뜻풀이와 감성으로 구성되어 있다.

〈Table 2〉 Sentiment marks of gloss

	neg	obj	pos
Sentiment degree	-1	0	1

〈Table 3〉 Training data set

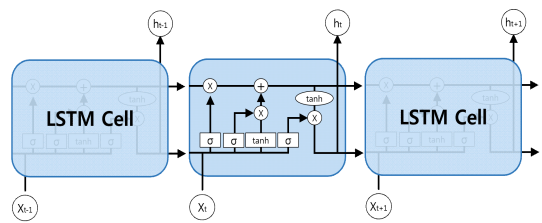
gloss	label
어떤 일이나 사물 따위에 대하여 좋은 느낌을 가진다.	1
특정한 음식따위를 특별히 잘 먹거나 마시다.	1
특정한 운동이나 놀이, 행동 따위를 즐겁게 하거나 하고 싶어하다.	1
남의 어리석은 말이나 행동을 비웃거나 빈정거릴 때 하는 말.	-1

3.1.4 뜻풀이 감성 분류 모델 구축

이 절에서는 뜻풀이 감성 분류 모델 구축 방안 에 대해 설명한다. 뜻풀이를 구성하는 문장 g_1, g_2, g_3, g_4, g_5 가 있다고 가정하고 긍정 뜻풀이 의 집합을 G_1 , 부정 뜻풀이의 집합을 G_{-1} 그리고 중립 뜻풀이의 집합을 G_0 이라고 가정하자. $G_1 = \{g_1, g_2\}$, $G_{-1} = \{g_4, g_5\}$ 그리고 $G_0 = \{g_3\}$ 인 경우, G_1 에 포함되어 있는 g_1 과 g_2 는 긍정의 성 향을 띄는 1-gram, 2-gram, 어구, 문형이 적어도 하나 이상은 포함되어 있어야 한다. G_{-1} 은 부정 의 성향을 띄는 1-gram, 2-gram, 어구, 문형이 적 어도 하나 포함되어 있어야 하며, G_0 는 감성을 띄지 않는 객관적인 1-gram, 2-gram, 어구, 문형 으로서 이루어져 있어야 한다. 따라서 각 G 에 포 함되어 있는 g 를 통하여 해당 감성을 표현하는 감성 어휘를 추출할 수 있다. 즉, $g_n = \{w_{n1}, w_{n2}, w_{n3}, \dots, w_{nk}\}$ 이고 w_{n2}, w_{n3} 이 감성 어 휘일 경우, $G_{sentivord} = \{w_{n2}, w_{n3}\}$ 이다(w_{nk} 는 n 번째 뜻풀이의 k번째 단어).

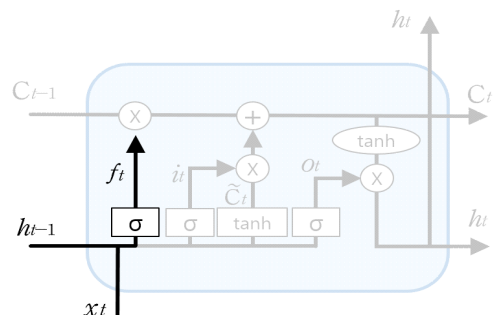
이 논문에서는 뜻풀이 분류 모델 구축을 위해 딥러닝 기법 중에서 문맥의 특징을 잘 압축하 는 Bi-LSTM 모델을 사용한다. Bi-LSTM은 입 력을 양방향으로 받는 것 외에 전체적인 셀 스

테이트(Cell State)의 구성은 LSTM과 같다. 예 를 들면, LSTM은 입력을 $[x_0, x_1, \dots, x_{n-1}]$ 으로 받지만 Bi-LSTM은 입력을 $[x_0, x_1, \dots, x_{n-1}]$ 과 $[x_{n-1}, \dots, x_1, x_0]$ 으로 받는다. 이는 중요한 정 보가 주로 뒤에 있는 한국어 문맥 탐지에 적합한 모델이다. LSTM은 RNN 모델이 가지고 있는 가 장 큰 단점인 시퀀스(Sequence)의 길이가 길어질 수록, 역전파시 기울기가 완만해지기 때문에 정 보가 손실되는 단점(Vanishing Gradient Problem) 을 해결한 모델이다(Christopher, 2015). 그림 2는 LSTM 모델을 나타낸다.

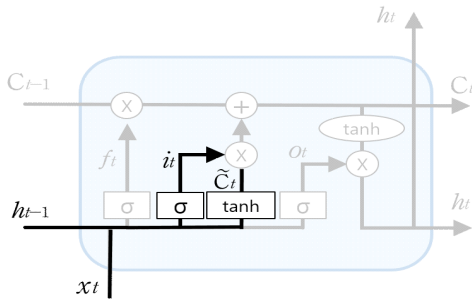


〈Figure 2〉 LSTM model

LSTM의 핵심은 셀 스테이트(Cell State)이다. 셀 스테이트는 포겟 게이트 레이어(Forget Gate Layer), 인풋 게이트 레이어(Input Gate Layer), 하이퍼볼릭 탄젠트 레이어(Tanh Layer)로 구성되어 있다. 그림 3은 포겟 게이트 레이어, 그림 4는 인



〈Figure 3〉 Forget gate layer



<Figure 4> Input gate, tanh layer

포켓 게이트 레이어와 하이퍼볼릭 탄젠트 레이어에 대한 그림이다.

포켓 게이트 레이어 f_t 는 셀 스테이트에서 어떤 정보를 버릴지 선택하는 과정이다. 이전의 히든 스테이트인 h_{t-1} 과 현재의 입력인 x_t 를 받아 시그모이드 함수를 취해준다. 출력 값이 0이면 완전히 이 값을 버린다는 의미이고, 1이면 완전히 이 값을 유지하라는 의미이다. f_t 는 아래와 같은 수식으로 표현된다.

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f)$$

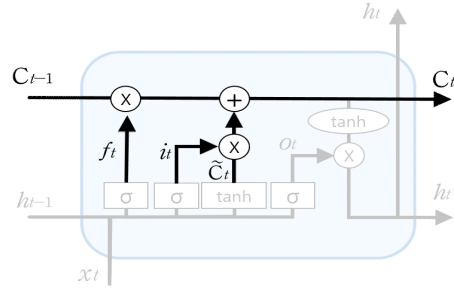
인풋 게이트 레이어 i_t 는 입력으로 h_{t-1} 과 x_t 를 받아 어떤 값을 업데이트할 것인가에 대하여 결정한다. 다음으로 하이퍼볼릭 탄젠트 레이어 \tilde{C}_t 는 셀 스테이트에 더해질 수 있는 값을 만들어 낸다. 두 레이어를 통해 나온 i_t 와 \tilde{C}_t 를 곱하여 다음 스테이트에 영향을 준다. i_t 와 \tilde{C}_t 의 수식은 아래와 같다.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C)$$

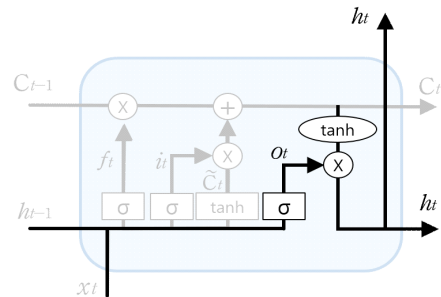
다음으로 이전의 셀 스테이트 C_{t-1} 를 현재 셀 스테이트인 C_t 로 업데이트 한다. C_t 는 포켓 레이어에서 나온 f_t 와 이전 셀 스테이트 C_{t-1} 의 곱을 인풋 게이트 레이어에서 나온 i_t 와 하이퍼볼릭 탄젠트 레이어에서 나온 \tilde{C}_t 의 곱의 합과 합한 값이다. 이와 같은 과정은 그림 5로 표현되며 C_t 의 수식은 다음과 같다.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



<Figure 5> Update step of C_t

마지막으로 그림 6에서는 h_t 의 출력에 대하여 결정한다. h_{t-1} 과 x_t 를 입력 받아 시그모이드 레이어를 통해 o_t 를 구한다. 그 다음 C_t 를 하이퍼볼릭 탄젠트 함수를 통해 -1과 1 사이의 값을



<Figure 6> Output step of h_t

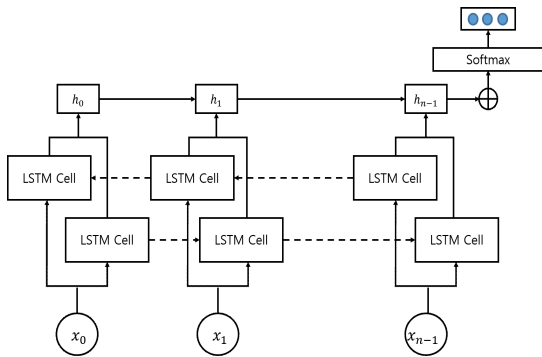
추출한 후 o_t 와 $\tanh(C_t)$ 를 곱해준다. 그림 6의 수식은 아래와 같다.

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

그림 7은 이 논문에서 뜻풀이 분류를 위해 사용한 Bi-LSTM 모델이다.

x_{n-1} 은 문장을 구성하는 n번째 단어 벡터의 입력을 의미한다. h_{n-1} 은 n번째까지의 입력을 통하여 나온 히든 스테이트(Hidden State)이다. 최종적으로 나온 h 을 소프트맥스(Softmax)라는 활성화 함수를 통해 확률로 변환 후 입력으로 주어진 뜻풀이에 대한 감성을 예측한다.



<Figure 7> Bi-LSTM model

3.1.5 감성 어휘 추출

뜻풀이 감성 분류 모델에 의해 분류된 긍정 뜻풀이의 집합을 G_1 , 부정 뜻풀이에 대한 집합을 G_{-1} 그리고 중립 뜻풀이에 대한 집합을 G_0 이라고 가정하자. 3.1.4절에서 설명한 것과 같이 감성 어휘 추출을 위해 G_1 에서는 긍정의 성향을 띄는 감성 어휘를 G_{-1} 에서는 부정의 성향을 띄

는 감성 어휘를 추출할 것이다. 만약 G_1 에 ‘어떤 일이나 사물 따위에 좋은 느낌을 가지다.’라는 뜻풀이 g_n 이 있다고 가정하자. g_n 은 긍정 뜻풀이 집합에 포함된 문장이기 때문에 해당 문장은 긍정과 관련된 1-gram, 2-gram, 어구, 문형을 적어도 1개 이상 포함하고 있을 것이다. 표 4는 g_n 에서 감성 어휘 w_{nk} 에 대하여 추출한 예제이다. 이와 같은 방법으로 긍정의 성향을 띄는 감성 어휘와 부정의 성향을 띄는 감성 어휘를 긍정, 부정으로 분류된 뜻풀이에서 추출한다.

<Table 4> Example of extracting sentiment words

gloss	“어떤 일이나 사물 따위에 대하여 좋은 느낌을 가지다.”	
Senti word	1-gram	좋은
	2-gram	좋은 느낌
	pattern	좋은 느낌을 가지다

3.2 감성 어휘 추출 방안 : SentiWordNet, SenticNet, 감정동사, 감성단어사전0603

이 논문에서는 뜻풀이에서 추출한 감성 어휘 이외에도 다양한 외부 소스를 활용해 KNU 한국어 감성사전의 감성 어휘를 확장한다.

3.2.1 SentiWordNet, SenticNet

SentiWordNet은 워드넷의 synset이라는 유의어 집단에 포함되어 있는 단어들의 긍정, 부정, 객관성에 대한 감성 정도를 부여한 감성사전이다. SenticNet은 단어들에 대한 일반적인 감각 개념의 감성을 추론한 감성사전이다. KNU 한국어 감성사전의 감성어휘를 확장하기 위해 SentiWordNet

과 SenticNet의 상위 k개의 긍·부정 값을 가지는 감성 어휘의 synset을 번역하여 추가한다. k는 synset에서 번역되어 추출된 감성 어휘들이 기 추출된 감성 어휘들과 중복될 때까지 반복 수행한다.

3.2.2 감정동사, 감성단어사전0603

김은영은 감정동사 전체 어휘를 대상으로 각 감정동사의 특징을 정리하고 이에 대한 연구를 하였으며 474개의 감정동사를 제시하였다. 감성단어사전0603은 ‘한국어 감정표현단어의 추출과 범주화(Son et al., 2012)’의 프로젝트 수행을 위해 BK21 플러스에 업로드 된 감성 단어 목록이다. 총 428개의 단어로 구성되어 있으며 감성 범주, 빈도, 감성 정도로 구성되어 있다. KNU 한국어 감성사전의 감성 어휘 확장을 위해 김은영이 제시한 474개의 감정 동사 목록과 감성단어사전 0603에 구축되어 있는 감성 어휘 중 도메인에 독립적인 어휘를 택하여 KNU 한국어 감성사전에 추가한다.

3.3 감성 어휘 추출 방안 : 신조어, 이모티콘

감성 분석에 사용되는 데이터들은 주로 텍스트 데이터이다. 텍스트 데이터는 표준어 이외에도 신조어, 이모티콘 등으로 구성되어 있다. 하지만 현재 구축되어 있는 대부분의 감성사전은 신조어와 이모티콘에 대한 정보가 없거나 부족하기 때문에 신조어나 이모티콘에 대한 감성을 탐지할 수 없다. 이 논문에서 제안하는 KNU 한국어 감성사전은 이와 같은 문제를 해결하기 위해 위키 백과에 등재된 신조어와 이모티콘을 수집하여 추가한다.

3.4 감성 어휘 추출 방안 : 수작업

표준국어대사전, SentiWordNet, SenticNet, 감정동사, 감성단어사전0603을 바탕으로 구축한 KNU 한국어 감성사전은 온라인 문서에 주로 쓰이는 감성 어휘가 들어있지 않을 수 있다. 이 논문에서는 네이버 중에서도 사람들이 주로 후기나 의견을 게시하는 여행 블로그, 자동차 블로그, 패션·뷰티 블로그, 뉴스 칼럼 등의 도메인에서 사용되었던 감성 어휘 중 도메인에 독립적인 감성 어휘를 수작업으로 수집하여 감성 사전을 확장한다. 또한 감성 어휘간의 유의어, 반의어를 통해 사전을 확장한다.

4. 실험 환경 및 실험 결과

4.1 실험 환경

이 논문의 실험을 위해 표준국어대사전의 형용사, 부사, 동사, 명사에 대한 모든 뜻을 수집하였다. 아래의 표 5는 학습 데이터와 테스트 데이터에 대한 통계를 나타낸다.

〈Table 5〉 The number of training data and test data

	train data set	test data set
the number of gloss	5,544	304,450

각 뜻풀이에 대해 한국어 정보처리를 위한 파이썬 패키지인 'KoNLPy(Park et al., 2014)'에 내장되어 있는 Twitter 형태소 분석기(Lee et al., 2010)를 사용하여 형태소 분석을 수행하고 데이터베이스에 저장하였다. 실험에 사용된 컴퓨터

는 Intel(R) Core(TM) i7-4790으로 CPU 성능은 3.60Hz이고 GPU는 GeForce GTX TITAN X를 사용하였다.

4.1.1 뜻풀이 감성 분류 모델

이 논문에서는 뜻풀이 분류를 위해 Bi-LSTM 모델을 사용하였다. Bi-LSTM 모델은 각 단어의 벡터를 입력으로 받기 때문에 이 연구에서는 FastText(Facebook, 2018)를 통하여 각 단어를 벡터 값으로 표상하였다. 표 6은 사용된 FastText의 파라미터(Parameter)이다.

〈Table 6〉 Used parameter values of FastText

	model	window size	dim
value	Skip-gram	2	50

표 7은 Bi-LSTM 모델 구축에 사용된 파라미터이다. 입력 차원(Input Dimension)은 1450 (29*50)이며, 최적화(Optimization)는 Adam 기법을 사용하였다. 또한 드롭아웃(Drop Out)기법을 적용하여 과적합(Over Fitting)을 방지하였다.

〈Table 7〉 Used parameter values of Bi-LSTM

	epoch	batch size	learning rate	drop out
value	100	50	Adam(0.0001)	0.26

4.2 실험 결과

4.2.1 뜻풀이 감성 분류

이 논문에서는 Bi-LSTM을 사용하여 뜻풀이 감성 분류 모델을 구축하고 구축된 모델을 통해

뜻풀이의 감성을 분류하였다. 뜻풀이의 감성 분류 결과는 표 8과 같다.

〈Table 8〉 Results of sentiment classification of train data set and test data set

	pos	neg	obj
train data	1,844	1,900	1,800
test data	55,489	44,524	204,437

뜻풀이의 정확도, 정밀도, 재현율, F1-score는 다음과 같다.

〈Table 9〉 Results of sentiment classification's accuracy, precision, recall and F1-score

	accuracy	precision	recall	F1-score
train data	0.9689			
test data (pos)	0.903	0.903	0.9281	0.9154
test data (neg)	0.886	0.886	0.9336	0.9092
test data (all)	0.8945	0.8945	0.93085	0.9123

표 9의 테스트 데이터에 대한 정확도, 정밀도, 재현율 그리고 F1-score는 분류된 테스트 데이터 100개를 무작위로 추출하여 평가자의 확인으로 측정 되었다. 테스트 데이터의 감성 분류 결과는 정확도와 정밀도에서 0.8945, 재현율에서는 0.93085 그리고 F1-Score에서는 0.9123이 측정되었다. 이를 통해 구축한 뜻풀이 감성 분류 모델이 감성 분류에 있어 유의미한 성능을 보이는 것을 알 수 있다.

4.2.2 감성 어휘 추출

이 논문에서 제안하는 뜻풀이를 통한 감성 어휘 추출은 두 가지 방법으로 나뉜다. 첫째, 뜻풀이 감성 분류 모델을 위하여 구축한 학습 데이터에서 감성 어휘를 추출하는 것이다. 둘째, 뜻풀이 감성 분류 모델에 의해 분류된 뜻풀이에서 감성 어휘를 추출하는 방법이다. 긍정으로 분류된 뜻풀이는 55,489개이고 부정으로 분류된 뜻풀이는 44,524개로 모든 뜻풀이에서 감성 어휘를 추출하는 것은 어렵다. 이와 같은 문제를 해결하기 위해 소프트맥스에 의해 산출된 각 뜻풀이의 확률 값을 기준으로 하여 상위 500개의 뜻풀이마다 감성 어휘를 추출하고 기 추출된 감성 어휘와 중복을 검사한다. 테스트 데이터에서 추출된 감성 어휘가 기 추출된 감성 어휘와 모두 중복될 경우 감성 어휘 추출을 종료하고 그렇지 않은 경우 추가로 500개의 뜻풀이에서 감성 어휘를 추출, 중복 검사를 수행한다. 표 10은 뜻풀이를 통해 추출된 감성 어휘의 통계를 나타낸다.

〈Table 10〉 Results of the number of extracting sentiment words using gloss

	very neg	neg	pos	very pos
senti word	4,201	4,105	1,807	2,268

표 11은 다양한 외부 소스를 통해 추출한 감성 어휘의 통계이다. SentiWordNet과 SenticNet에서 총 557개의 감성 어휘를 추출하였다. SentiWordNet과 SenticNet은 각각 긍·부정어 상위 500개, 100개의 synset을 통해 감성 어휘 확장이 수행되었다.

감정동사, 감성단어사전0603을 통해 총 338개의 감성 어휘를 추가하였으며 특정 어휘에 대한 다양한 형태를 추가하였다. 예를 들면, ‘외롭다’

라는 단어에 대하여 ‘외로워지다’, ‘외로워하다’ 등의 형태를 추가하였다. 또한 최근 온라인에서 많이 사용되는 신조어, 이모티콘을 KNU 한국어 감성사전에 추가하여 온라인에서 사용되는 신조어와 이모티콘에 대한 감성을 탐지할 수 있도록 하였다. 마지막으로, 이 논문에서는 수작업을 통해 웹에서 사용되는 감성 어휘를 수작업으로 추출하였으며 유의어, 반어의 등의 관계를 통해 확장하였다.

〈Table 11〉 Results of the number of extracting sentiment words using external source

	SWN SenticNet	senti.verb senti wordDic	neologism emoticon	manual
senti word	557	338	228	1,339

4.2.3 KNU 한국어 감성사전 통계

표 12는 KNU 한국어 감성사전의 통계치이다. 통계치를 보면 KNU 한국어 감성사전을 구성하는 감성 어휘는 긍정보다 부정에 더 많이 편향되어 있다는 것을 알 수 있다. 이를 통해 부정적인 감성 어휘가 긍정적인 감성 어휘에 비해 다양한 표현을 가지고 있다는 것을 알 수 있었다.

〈Table 12〉 The number of KNU Sentiment Lexicon's sentiment words

	very neg	neg	obj	pos	very pos
senti word	4,797	5,029	154	2,266	2,597

이 논문에서는 2-gram, 어구, 문형 정보를 통해 1-gram으로 표현할 수 없는 감성 어휘에 대하여 추출하였다. 예를 들면, ‘맛있지’라는 어휘의

감성은 긍정으로 볼 수 있다. 하지만 ‘맛있지’라는 단어의 앞이나 뒤에 특정 단어가 결합됨에 따라 다른 감성을 나타낼 수 있다. 예를 들면, ‘맛있지 않다’는 ‘맛있지’와 ‘않다’가 결합하여 부정의 감성을 나타낸다. 표 13은 긍·부정 어휘의 형태별 통계를 나타낸다. 이를 통해 KNU 한국어 감성사전에는 2-gram, 어구, 문형과 같은 다양한 감성 어휘가 있다는 것을 알 수 있다.

〈Table 13〉 The number of sentiment word's type

	1-gram	2-gram	phrase	pattern	neologism, emoticon
senti word	6,223	7,861	278	253	228

표 14는 KNU 한국어 감성사전의 감성 어휘에 대한 예이다.

〈Table 14〉 Example of the sentiment words in KNU Sentiment Lexicon

		senti word	
pos	1-gram	귀엽다 (2)	
		리즈시절 (1)	
	2-gram	바르며 상냥하게 (2)	
		씩씩하고 힘차다 (2)	
	phrase	아는 것이 깊고 (1)	
		아름다운 말과 글 (2)	
pattern	마음에 여유가 있다 (2)		
	더할 나위 없이 좋음 (2)		
neg	1-gram	가난 (-2)	
		탄식하다 (-2)	
	2-gram	바람을 피우며 (-1)	
		탈이 생기다 (-1)	
	phrase	하는 동 마는 동 (-1)	
		사리에 어두운 사람 (-2)	
	pattern	아무 값어치나 의의가 없다 (-1)	
		힘에 부치는 데가 있다 (-2)	

4.2.4 감성사전 별 정성·정량 평가

이 절에서는 기존에 구축되어 있는 한국어 기반의 감성사전과 이 논문에서 제안한 KNU 한국어 감성사전을 비교 평가 하였다. 비교 대상으로 기존에 널리 알려진 서울대학교에서 구축한 한국어 감정 어휘 목록, 오픈 한글의 감성어 사전을 선택하였다.

서울대의 한국어 감정 어휘 목록은 총 16,362가지의 1-gram, 2-gram, 3-gram의 감정 표현을 가지고 있다. 5가지 표현 유형을 정의하였으며 감정 어휘의 감정 정도는 출현 빈도와 비율로써 계산되었다(Shin et al., 2016). 2011년부터 2013년까지 구축한 한국어감정분석코퍼스(KOSAC)를 활용하였으며, 기계적인 알고리즘을 통해 구축된 사전으로 특정 도메인에 영향을 받는 어휘에 대해서는 충분히 고려되지 못하였다.

오픈 한글의 감성어 사전은 국립국어원의 표준국어대사전에 등재되어 있는 단어 중 명사, 형용사, 동사, 부사를 우선순위로 하여 랜덤하게 사용자들이 감성을 부여하고 이를 통해 감성을 분류하도록 하였다. 투표를 통해 나온 감성을 비율로써 계산하여 긍·부정 확률을 퍼센트로 계산하였다(An et al., 2015). 하지만 이는 국립국어원의 사전에 등재되어 있는 단어에 대해서만 고려되었으며 특정 도메인에 영향을 받는 어휘에 대해서는 충분히 고려되지 못하였다. 감성어의 개수는 총 33,987개로 구성되어 있다.

이 논문에서 제안한 KNU 한국어 감성사전은 국립국어원의 표준국어대사전을 기반으로 등재되어있는 뜻을 풀이를 긍·부정으로 분류하여 도메인에 영향을 받지 않는 감성 어휘를 1-gram, 2-gram, 어구, 문형 형태로 추출하였다. 또한 온라인 텍스트에 자주 등장하는 신조어, 이모티콘

에 대한 감성 어휘를 추가하였다.

아래의 표 15와 16은 각각의 감성사전에 대한 정성적, 정량적 평가이다.

〈Table 15〉 Comparison of korean sentiment lexicon between seoul univ, open hangul and KNU Sentiment Lexicon about qualitative indicator

	source	method	consider domain	service
seoul univ	kosac	counting	no	yes
open hangul	std korean	voting	no	no
KNU	std korean	extracting voting	yes	yes

〈Table 16〉 Comparison of korean sentiment lexicon between seoul univ, open hangul and KNU Sentiment Lexicon about quantitative indicator

	1-gram	2-gram	3-gram	n-gram
seoul univ	3,476	6,579	6,307	none
open hangul	33,987			
our lexicon	6,451	8,135	226	31

표 15와 16을 보면 이 논문에서 제안한 KNU 한국어 감성사전은 도메인에 영향을 받지 않는 감성 어휘만을 고려하여 구축하였기 때문에 기존의 감성사전 보다 적은 어휘의 개수를 보인다. 하지만 기존의 감성사전 보다 다양한 형태의 어휘 정보를 가지고 있으며 3명의 투표자가 직접 감성 어휘를 추출하고 감성 정도를 측정했다는 점에서 보다 정확한 감성 정도를 가지고 있을 것

으로 예상된다. 또한 도메인에 영향을 받지 않는 어휘로 구성되어있다는 점에서 다양한 도메인에서 보다 정확한 감성사전으로 활용될 수 있을 것이다.

4.2.5 KNU 한국어 감성사전 데모 페이지

그림 8은 KNU 한국어 감성사전 데모 페이지이다. KNU 한국어 감성사전 데모 페이지에서는 감성사전에 기록되어 있는 감성 어휘에 대한 정보를 확인할 수 있다. 그림 9와 그림 10은 KNU 한국어 감성사전 데모 페이지에 대한 사용 예이다.

KNU 한국어 감성사전

단어: 입력

어근:

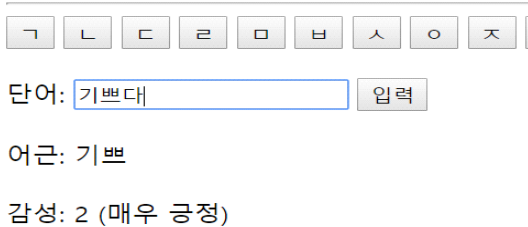
감성:

〈Figure 8〉 Demo page of KNU Sentiment Lexicon

그림 9-1과 9-2는 입력된 단어에 대한 정보를 보여주는 예이다. 입력된 단어가 감성사전에 등재되어 있는 경우 해당 단어에 대한 어근과 감성에 대한 정보를 보여준다. 입력된 단어가 감성사전에 등재되어 있지 않은 경우 사전에 없는 단어라는 메시지를 출력한다.

그림 10은 자음별 감성 어휘 목록에 대하여 나타낸다. 빨간 박스 안의 자음 버튼을 누르면 데모 페이지에서는 해당 자음으로 시작하는 감성 어휘, 어근, 감성 정도를 보여준다. KNU 한국어 감성사전의 데모페이지는 'dilab.kunsan.ac.kr/knu/knu.html'에서 사용할 수 있다.

KNU 한국어 감성사전

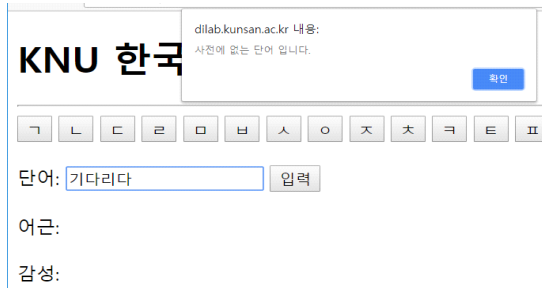


단어: 기쁘대 입력

어근: 기쁘

감성: 2 (매우 긍정)

〈Figure 9-1〉 Results of input word which in KNU Sentiment Lexicon



dilab.kunsan.ac.kr 내용: 사전에 없는 단어입니다. 확인

KNU 한국

단어: 기다리다 입력

어근:

감성:

〈Figure 9-2〉 Results of input word which not in KNU Sentiment Lexicon

KNU 한국어 감성사전



단어: 가난
어근: 가난
감성: -2

단어: 가난했미
어근: 가난했미
감성: -2

단어: 가난살미
어근: 가난
감성: -2

단어: 가난살미하다
어근: 가난
감성: -2

〈Figure 10〉 List of sentiment words according to consonants

5. 결론 및 향후 연구

감성사전은 긍정, 부정의 감성을 지니는 어휘에 대한 사전이다. 감성 어휘는 특정 도메인에 따라 감성의 종류나 정도가 달라질 수 있다. 이와 같은 이유로 정확한 감성 분석을 수행하기 위해서는 해당 도메인에 알맞은 감성사전을 구축해야한다. 특정 도메인에 대한 감성사전을 빠르고 효율적으로 구축하기 위해서는 연구자의 주관보다는 구축되어 있는 범용 한국어 감성사전을 활용하는 것이 유용하다.

이 논문에서는 특정 도메인에 대한 감성사전을 빠르게 구축하기 위한 기초 자료로 활용될 수 있는 범용 한국어 감성사전을 구축하고 이를 ‘KNU 한국어 감성사전’이라고 명명하였다.

KNU 한국어 감성사전은 특정 도메인에서 사용되는 감성 어휘 보다는 인간의 보편적인 기본 감성 표현을 나타내는 14,843개의 감성 어휘로 구성되어 있다. KNU 한국어 감성사전을 구축하기 위해 표준국어대사전의 뜻을 딴어 기법을 사용하여 감성을 분류하였고 분류 간에 89.45%의 정확도를 보였다. 감성이 긍정으로 분류된 뜻풀이에서는 긍정에 관련된 감성 어휘를, 부정으로 분류된 뜻풀이에서는 부정에 관련된 감성 어휘를 추출하였다. 또한 다양한 외부 소스를 활용하여 감성 어휘와 신조어, 이모티콘을 KNU 한국어 감성사전에 추가하였다.

KNU 한국어 감성사전은 도메인에 독립적인 감성 어휘로 구성되어있기 때문에 다양한 분야의 실무자들이 감성사전 기반의 감성 분석에 있어 제약없이 활용할 수 있다. 또한 특정 분야에 정밀한 감성사전을 구축하기 위해 기초 자료로써 활용될 수 있다. 예를 들면, KNU 한국어 감성사전의 감성 어휘는 도메인에 의존적인 단어

를 확장하기 위한 시드 감성 어휘로 활용될 수 있다.

이 논문에서 구축한 KNU 한국어 감성사전의 공헌은 다음과 같다. 첫째, 표준국어대사전에 수록되어있는 뜻을 활용하여 도메인에 독립적인 기본 감성 어휘를 추출하고 이를 통해 감성사전을 구축한 최초의 연구이다. 둘째, KNU 한국어 감성사전은 실제 활용 가능한 범용 한국어 감성사전으로써 특정 도메인에 알맞은 감성사전을 효율적으로 구축하기 위한 기초 자료로 활용될 수 있다. 셋째, KNU 한국어 감성사전은 도메인에 독립적인 감성 어휘로만 구성되어 있다는 점에서 어떠한 도메인에도 제약받지 않고 감성사전 기반의 감성 분석 수행이나 대량의 학습 데이터 구축 그리고 딥러닝 입력의 자질로써 활용될 수 있다.

향후 연구로는 KNU 한국어 감성사전을 활용하여 특정 도메인의 감성사전을 자동으로 구축하는 알고리즘을 개발할 것이다. 뜻풀이 감성 분류 모델의 입력으로 구문 분석과 의미 분석 결과를 자질로 추가하여 문장 분류를 성능을 증가시킬 것이다. 또한 KNU 한국어 감성사전을 활용하여 대량의 학습 데이터 세트를 자동으로 구축하여 기계 학습에 적용해볼 것이며 마지막으로 세종 말뭉치와 같은 다양한 코퍼스를 활용하여 감성 어휘를 지속적으로 추가할 예정이다.

참고문헌(References)

- An, J. K. and H. W. Kim, "Building a Korean Sentiment Lexicon Using Collective Intelligence," *Journal of Intelligence and Information Systems*, Vol. 21, No. 2(2015), 49~67
- Baccianella, S., A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", *Proceedings of the International Conference on Language Resources and Evaluation*, LREC(2010), 2200-2204
- Cambria, E., S. Poria, D. Hazarika, and K. Kwok, "SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings", *AAAI*(2018)
- Choi, S. J. and O. B. Kwon, "The Study of Developing Korean SentiWordNet for Big Data Analytics - Focusing on Anger Emotion -," *The Journal of Society for e-Business Studies*, Vol. 19, No. 4(2014), 1~19
- Choi, S. J., Y. E. Song, and O. B. Kwon, "Analyzing Contextual Polarity of Unstructured Data for Measuring Subjective Well-Being," *Journal of Intelligence and Information Systems*, Vol. 22, No. 1(2016), 83~105
- Christopher, O., Understanding LSTM Networks, 2015, Available at <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> (Accessed 2018)
- CIO, Expansion of BI and Analytical Role, 2013, Available at <http://www.ciokorea.com/t/544/9118/15551> (Accessed 2018)
- Facebook, FastText, Available at <https://fasttext.cc/> (Accessed 2018)
- Fellbaum, C., WordNet, Available at <http://wordnet.princeton.edu/> (Accessed 2018)
- Jang, H. S., K. Y. Jeong, and E. Y. Jang, "Efficient method to generate sentiment vocabulary for specific topic based on

- Word2Vec," *Proceedings of Korean Institute of Information Scientists and Engineers* (2017), 652~654
- Kim, D. H., T. M. Cho, and J. H. Lee, "A Domain Adaptive Sentiment Dictionary Construction Method for Domain Sentiment Analysis," *Proceedings of the Korean Society of Computer Information Conference*, Vol. 23, No. 1(2015), 15~18
- Kim, E. Y., "A Study on the Korean Emotion Verbs", *A Doctoral Dissertation at JeonNam National University*(2004)
- Kim, S. B., S. J. Kwon, and J. T. Kim, "Building Sentiment Dictionary and Polarity Classification of Blog Review By Using Elastic Net," *Proceedings of Korean Institute of Information Scientists and Engineers* (2015), 639~641
- Kim, S. I., sentiment lexicon0603, 2015, Available at http://datascience.khu.ac.kr/board/bbs/board.php?bo_table=05_01&wr_id=91 (Accessed 2018)
- KOSAC, Korean Sentiment Lexicon, Available at word.snu.ac.kr/kosac/lexicon.php (Accessed 2018)
- Lee, C. H., J. M. Sim, and A. S. Yoon, "The Review about the Development of Korean Linguistic Inquiry and Word Count", *KOREAN JOURNAL OF COGNITIVE SCIENCE*, Vol. 16, No. 2(2005), 93~121
- Lee, D. J., J. H. Yeon, I. B. Hwang, and S. G. Lee, "KKMA : A Tool for Utilizing Sejong Corpus based on Relational Database," *Journal of KIISE : Computing Practices and Letters*, Vol. 16, No. 11(2010), 1046~1050
- Lee, S. H., J. Choi, and J. W. Kim, "Sentiment Analysis on Moive Review Through Building Modified Sentiment Dictionary by Moive Genre", *Journal of Intelligence and Information Systems*, Vol. 22, No. 2(2016), 97~113
- National Institute of the Korean Language(NIKL), Standard Korean Dictionary, Available at stdweb2.korean.go.kr (Accessed 2018)
- OpenHagul, Sentiment Lexicon, Available at openhagul.com/restrict (Accessed 2018)
- Park, C. Y., stdkor, Available at <https://github.com/forkonlp/stdkor> (Accessed 2018)
- Park E. J. and S. J. Jo, "KoNLPy: Korean natural language processing in Python", *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology*(2014), 133~136
- Park, S. M. and B. W. On, "Latent-Based Product Reputation Mining", *Journal of Intelligence and Information Systems*, Vol. 23, No. 2(2017), 39~70
- Shin, D. H., D. Cho, and J. S. Nam, "Building the Korean Sentiment Lexicon DecoSelex for Sentiment Analysis," *Journal of Korealex*, No. 28(2016), 75~111
- Shin, H. P., M. H. Kim, and S. Z. Park, "Modality-based Sentiment Analysis through the Utilization of the Korean Sentiment Analysis Corpus," *EONEOHAG : Journal of the Linguistic Society of Korea*, No. 74 (2016), 93~114
- Son, S. J., M. S. Park, J. E., Park and J. H. Son, "Korean Emotion Vocabulary: Extraction and Categorization of Feeling Words", *Science of Emotion & Sensibility*, vol. 15(2012), 105-120
- Teng, Z., D. T. Vo, and Y. Zhang, Teng, Zhiyang,

- Duy-Tin Vo and Yue Zhang. “Context-Sensitive Lexicon Features for Neural Sentiment Analysis.” *EMNLP*(2016)
- Wikipedia, Deep Learning, Available at https://en.wikipedia.org/wiki/Deep_learning (Accessed 2018a)
- Wikipedia, Emoticon, Available at <https://ko.wikipedia.org/wiki/%EC%9D%B4%EB%AA%A8%ED%8B%B0%EC%BD%98> (Accessed 2018b)
- Wikipedia, Neologism, Available at https://ko.wikipedia.org/wiki/%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD%EC%9D%98_%EC%9D%B8%ED%84%B0%EB%84%B7_%EC%8B%A0%EC%A1%B0%EC%96%B4_%EB%AA%A9%EB%A1%9D (Accessed 2018c)
- Wikipedia, Sentiment Analysis, Available at https://en.wikipedia.org/wiki/Sentiment_analysis (Accessed 2018d)
- Wikipedia, Text Mining, Available at https://en.wikipedia.org/wiki/Text_mining (Accessed 2018e)

Abstract

KNU Korean Sentiment Lexicon: Bi-LSTM-based Method for Building a Korean Sentiment Lexicon

Sang-Min Park* · Chul-Won Na** · Min-Seong Choi** · Da-Hee Lee** · Byung-Won On***

Sentiment analysis, which is one of the text mining techniques, is a method for extracting subjective content embedded in text documents. Recently, the sentiment analysis methods have been widely used in many fields. As good examples, data-driven surveys are based on analyzing the subjectivity of text data posted by users and market researches are conducted by analyzing users' review posts to quantify users' reputation on a target product. The basic method of sentiment analysis is to use sentiment dictionary (or lexicon), a list of sentiment vocabularies with positive, neutral, or negative semantics. In general, the meaning of many sentiment words is likely to be different across domains. For example, a sentiment word, 'sad' indicates negative meaning in many fields but a movie.

In order to perform accurate sentiment analysis, we need to build the sentiment dictionary for a given domain. However, such a method of building the sentiment lexicon is time-consuming and various sentiment vocabularies are not included without the use of general-purpose sentiment lexicon. In order to address this problem, several studies have been carried out to construct the sentiment lexicon suitable for a specific domain based on 'OPEN HANGUL' and 'SentiWordNet', which are general-purpose sentiment lexicons. However, OPEN HANGUL is no longer being serviced and SentiWordNet does not work well because of language difference in the process of converting Korean word into English word. There are restrictions on the use of such general-purpose sentiment lexicons as seed data for building the sentiment lexicon for a specific domain.

In this article, we construct 'KNU Korean Sentiment Lexicon (KNU-KSL)', a new general-purpose Korean sentiment dictionary that is more advanced than existing general-purpose lexicons. The proposed

* Department of Software Convergence Engineering, Kunsan National University

** Department of Software Convergence Engineering, Kunsan National University

*** Corresponding Author: Byung-Won On

Department of Software Convergence Engineering, Kunsan National University

558 Daehak-ro, Gunsan-si, Jeollabuk-do 54150, Korea,

Tel: +82-63-469-8913, Fax: +82-63-469-7423, E-mail: bwon@kunsan.ac.kr

dictionary, which is a list of domain-independent sentiment words such as ‘thank you’, ‘worthy’, and ‘impressed’, is built to quickly construct the sentiment dictionary for a target domain. Especially, it constructs sentiment vocabularies by analyzing the glosses contained in Standard Korean Language Dictionary (SKLD) by the following procedures:

First, we propose a sentiment classification model based on Bidirectional Long Short-Term Memory (Bi-LSTM).

Second, the proposed deep learning model automatically classifies each of glosses to either positive or negative meaning.

Third, positive words and phrases are extracted from the glosses classified as positive meaning, while negative words and phrases are extracted from the glosses classified as negative meaning.

Our experimental results show that the average accuracy of the proposed sentiment classification model is up to 89.45%. In addition, the sentiment dictionary is more extended using various external sources including SentiWordNet, SenticNet, Emotional Verbs, and Sentiment Lexicon 0603. Furthermore, we add sentiment information about frequently used coined words and emoticons that are used mainly on the Web. The KNU-KSL contains a total of 14,843 sentiment vocabularies, each of which is one of 1-grams, 2-grams, phrases, and sentence patterns. Unlike existing sentiment dictionaries, it is composed of words that are not affected by particular domains.

The recent trend on sentiment analysis is to use deep learning technique without sentiment dictionaries. The importance of developing sentiment dictionaries is declined gradually. However, one of recent studies shows that the words in the sentiment dictionary can be used as features of deep learning models, resulting in the sentiment analysis performed with higher accuracy (Teng, Z., 2016). This result indicates that the sentiment dictionary is used not only for sentiment analysis but also as features of deep learning models for improving accuracy. The proposed dictionary can be used as a basic data for constructing the sentiment lexicon of a particular domain and as features of deep learning models. It is also useful to automatically and quickly build large training sets for deep learning models.

Key Words : Sentiment Lexicon, Sentiment Analysis, Deep Learning, Text Mining, Bi-LSTM

Received : September 17, 2018 Revised : December 18, 2018 Accepted : December 24, 2018

Publication Type : Regular Paper Corresponding Author : Byung-Won On

저 자 소개



박상민

군산대학교 소프트웨어융합공학과 대학원 석사 과정 2학기 재학 중이며 연구 분야는 한국어 감성사전 구축, 빅데이터 기반의 여론조사 소프트웨어 개발이며, 관심 분야는 데이터 마이닝과 인공지능이다.



나철원

군산대학교 소프트웨어융합공학과 학부 3학년 재학 중이며 대학원에 진학할 예정이다. 연구 분야는 감성 문서 자동 크롤링이며, 관심 분야는 크롤링과 인공지능이다.



최민성

군산대학교 소프트웨어융합공학과 학부 3학년 재학 중이며 연구 분야는 딥러닝 학습 데이터 자동 구축이며, 관심 분야는 분산 시스템과 인공지능이다.



이다희

군산대학교 소프트웨어융합공학과 학부 2학년 재학 중이며 연구 분야는 광고성 댓글 판별이며, 관심 분야는 데이터 마이닝과 인공지능이다.



온병원

2007년, 미국 펜실베이니아 주립대학교의 컴퓨터공학과에서 박사학위를 취득한 후, 캐나다 브리티시 컬럼비아 대학교에서 박사 후 연구원으로 재직하였다. 2010년, 미국 일리노이 대학교의 차세대디지털과학센터에서 선임연구원으로 근무하였고, 서울대학교 차세대융합기술연구원에서 연구교수를 역임하였다. 현재는 군산대학교 소프트웨어융합공학과 교수로 재직 중이다. 주요 연구 분야로는 데이터 마이닝, 정보검색, 빅데이터, 인공지능 등이다.