# 어휘 관계 및 문맥 정보 기반의 도메인 감성사전 자동 구축 방안 연구

A Study on the Method for Automatically Constructing a Domain Specific Sentiment Lexicon Based Lexical Relation and Contextual Information


저자
(Authors)

박상민, 온병원
Sangmin Park, Byung-Won On



출처
(Source)

정보과학회논문지 47(10), 2020.10, 926-941 (16 pages)
Journal of KIISE 47(10), 2020.10, 926-941 (16 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10475009

APA Style

박상민, 온병원 (2020). 어휘 관계 및 문맥 정보 기반의 도메인 감성사전 자동 구축 방안 연구. 정보과학회논문지, 47(10), 926-941.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 47, No. 10, pp. 926-941, 2020. 10
https://doi.org/10.5626/JOK.2020.47.10.926

# 어휘 관계 및 문맥 정보 기반의 도메인 감성사전 자동 구축 방안 연구

## (A Study on the Method for Automatically Constructing a Domain Specific Sentiment Lexicon Based Lexical Relation and Contextual Information)

박 상 민 †　　온 병 원 ††
(Sangmin Park)　　(Byung-Won On)

**요 약** 감성사전은 감성 어휘들에 대한 집합으로 각 어휘들에 대한 감성의 극성이 부여되어 있으며, 감성 분석(Sentiment Analysis)을 위한 기초 자료로 활용된다. 하지만 이와 같은 감성 어휘들은 특정 도메인에 따라 극성이 역전되거나 유실될 수도 있으며 의존적인 감성 어휘가 존재할 수 있다. 예를 들면, 일반적으로 '잘 잤다'라는 단어는 긍정의 극성을 보이지만, 영화 도메인에서는 그 의미가 부정으로 쓰인다. 그렇기 때문에 감성사전은 분석하고자 하는 도메인의 특징이 반영되어 있어야 하며 도메인에 따라 알맞은 감성사전이 구축되고 활용되어야 한다. 이와 같은 문제를 해결하기 위해 현재 도메인 감성사전을 자동으로 구축하는 다양한 연구들이 나왔지만, 인간의 개입, 문맥적 요소 미고려, 지역적인 정보 반영 등이라는 문제점을 지니고 있다. 본 연구에서는 이와 같은 문제를 해결하기 위해 한국어 범용 감성사전인 'KNU 한국어 감성사전'과 글로브 벡터 그리고 접속사 관계를 활용하여, 특정 도메인의 전역적인 감성 정보와 문맥적 특징을 충분히 반영한 도메인 감성사전 구축 방안을 제안한다.

**키워드:** 감성사전, 도메인 감성사전, 반지도 학습, 감성 어휘, 언어 표상

***Abstract*** A sentiment lexicon is a set of sentiment words, each of which has its sentiment polarity, and is used as a basic method for sentiment analysis. However, the meaning of some words can be different or even the original meaning can disappear across domains. As such, many sentiment words are likely to depend on a specific domain. For example, the verb phrase "slept well" usually has a negative meaning, while it has a positive meaning in movie domains. Thus, given a particular domain such as hotel, the sentiment lexicon should be constructed so that many of the domain-dependent words reflect the meaning of the domain. Using the domain-dependent sentiment lexicon will render more accurate results than using existing sentiment lexicons that do not consider domain-dependent words in the sentiment analysis. To build the domain-dependent sentiment lexicons, various studies have been presented, but there are many limitations including the human intervention and the use of local information rather than contextual information. In this paper, we propose a novel method of automatically

· 본 연구는 2019년도 정부(과학기술정보통신부)의 한국연구재단의 개인기초연구사업(No. NRF-2019R1F1A1060752)의 연구비 지원으로 수행하였습니다.

† 비 회 원 : 군산대학교 소프트웨어융합공학과 학생
park1200656@gmail.com
†† 종신회원 : 군산대학교 소프트웨어융합공학과 교수
(Kunsan Nat'l Univ.)
bwon@kunsan.ac.kr
(Corresponding author임)

논문접수 : 2020년 3월 20일
(Received 20 March 2020)
논문수정 : 2020년 8월 12일
(Revised 12 August 2020)
심사완료 : 2020년 8월 13일
(Accepted 13 August 2020)


정보과학회논문지 제47권 제10호(2020. 10)

constructing a domain-dependent sentiment lexicon based on the global and contextual information and an existing sentiment lexicon (i.e., KNU sentiment lexicon, Glove vector, Conjunction relation).

**Keywords:** domain specific sentiment lexicon, semi-supervised learning, sentiment words, KNU sentiment lexicon, lexical relation, contextual information

## 1. 서 론

가트너는 비즈니스 인텔리전스(Business Intelligence) 책임자들이 점차 확대되어가는 정보 자산을 수용하고 기업은 조직의 내부와 외부 등에서 생성되는 다양한 데이터들을 분석 및 통찰력을 발견해 이를 다양한 의사결정에 활용하고자 한다고 언급했다. 특히 빅데이터에 대한 시장의 관심이 크며 새로운 데이터로부터 통찰력을 추출, 활용 그리고 행동하는 것의 가치가 증대되었다고 언급했다[1]. 특히, 인터넷과 SNS 등의 확산으로 비정형 텍스트의 폭발적인 증가와 이의 잠재적인 가치가 높아짐에 따라 비정형 텍스트 데이터를 분석하여 의미 있는 정보를 발견하는 텍스트 마이닝(Text Mining)의 중요성이 대두되고 있으며 그 중에서도 여론 조사, 시장 조사, 마케팅 등 다양한 분야에서 활용되는 감성 분석(Sentiment Analysis)의 중요성이 대두되고 있다[2]. 감성 분석은 대표적으로 사전 기반의 감성 분석과 기계학습 기반의 감성분석 그리고 딥러닝 기반의 감성 분석이 있지만, 가장 기초적이고 직관적이며 학습 데이터를 따로 필요로 하지 않는 방법은 사전 기반의 감성 분석이다. 사전 기반의 감성 분석은 분석 대상 텍스트 코퍼스로부터 긍정, 부정, 중립에 대한 감성 추출하고 이를 통해 어휘 수준의 감성 분석을 수행하는 것이며, 이와 같은 감성 어휘 목록을 감성사전(Sentiment Lexicon)이라고 한다. 감성사전을 구성하는 감성 어휘들은 특정 도메인에 따라 가지는 극성이 역전될 수도 있고 보존될 수도 있다. 또한 일반적으로 감성을 지니는 어휘가 특정 도메인에서는 감성이 없는 경우도 존재한다. 예를 들면, ‘슬프다’라는 감성 어휘는 일반적으로 부정의 감성을 지니고 있지만 영화 도메인에서의 ‘슬프다’라는 감성 어휘는 부정의 감성보다는 긍정의 감성으로 사용된다. 이와 같이 사전 기반의 감성 분석을 수행함에 있어 감성사전은 분석하고자 하는 도메인의 특성을 반영해야 한다. 도메인의 특성이 반영되지 않은 감성사전 해당 도메인 코퍼스의 분석에 있어 정확한 분석의 수행이 어렵기 때문이다. 그렇기 때문에 대부분의 감성사전은 감성 분석을 하고자 하는 도메인 코퍼스를 대상으로 연구자들이 직접 구축하여 활용한다. 하지만 이와 같이 연구자들이 직접 도메인 감성사전을 구축하는 경우 시간이 많이 소요될 뿐만 아니라 기준이 되는 감성 어휘 없이 감성사전을 구축할 경우 다양한 감성 어휘가 포함될 수 없다. 또한 도메인 지식이 부족할 경우 관련된 어휘를 찾기 어렵고 연구자의 주관이 들어가게 될 경우 주관적인 감성 어휘들이 추출될 수 있다는 문제점이 존재한다.

본 연구에서는 이와 같은 기존의 도메인 감성사전 구축과 관련된 문제를 해결하고 도메인 감성사전을 자동으로 구축하는 방법을 제안하며 제안 방안은 다음과 같다. 첫 번째, 수집한 도메인 코퍼스를 통해 도메인 감성사전을 구축해야하기 때문에 게시자의 감성이 많이 반영되어 있을 것으로 예상되는 칼럼, 후기 글, 댓글 등을 대상으로 수집한다. 두 번째, 수집된 도메인 코퍼스에서 감성을 지니는 형용사 어휘, 명사 어휘를 대상으로 후보 감성 어휘들을 추출하고 이들이 ’KNU 한국어 감성사전‘에 수록된 어휘이면 시드 감성 어휘로 그렇지 않으면 후보 감성 어휘로 분류한다. 세 번째, 도메인 코퍼스를 활용하여 글로브(GloVe) 벡터[3]를 학습하고 추출된 전체 감성 어휘에 대해 자카드 유사도(Jaccard Similarity)를 활용하여 어휘 간 유사도 행렬을 구축한다. 네 번째, 접속사를 도메인 코퍼스에서 추출하여 어휘 간 접속 관계 행렬을 구축한다. 다섯 번째, 어휘 간 유사도 행렬과 어휘 간 접속 관계 행렬 그리고 시드 감성 어휘를 활용하여 후보 감성 어휘에 대한 극성 레이블 전파를 수행, 이를 통해 후보 감성 어휘에 대한 극성을 자동으로 부여한다. 마지막으로 도메인 코퍼스의 의존 구문 분석(Dependency Parsing)을 활용하여 감성 어휘를 포함하는 어구를 추출하고 어구를 분석해 n-gram 형태의 감성 어휘를 추출한다. 추출된 n-gram 형태의 감성 어휘는 사전 기반의 감성 분석에 의해 극성이 자동으로 부여된다. 본 연구에서 제안한 도메인 감성사전 자동 구축 방안은 자동차 리뷰 사이트인 오토뷰[4]의 후기 게시판과 네이버 영화 리뷰 코퍼스를 수집하여 진행되었으며 자동차 도메인의 후보 감성 어휘에 대한 극성의 정확도는 86.6%로 대부분의 극성이 일치하는 것을 알 수 있었다. 영화 리뷰 도메인은 71%의 정확도를 보였는데 이는 띄어쓰기가 잘 수행되어있지 않은 리뷰가 많았으며 영화 리뷰에서도 다양한 주제(액션, 로맨스, 코미디 등)의 영화 장르가 섞여 있어 자동차 도메인보다는 낮은 정확도를 보였다. 또한 1-gram으로 나타낼 수 없는 감성 어휘들을 보완하기 위해 n-gram 형태의 감성 어휘들을 추출하였다는 점에서 차별성을 둘 수 있다.

## 2. 배경 지식 및 관련 연구

### 2.1 배경 지식

PMI(Point-wise Mutual Information)는 두 개체간

연관도를 측정하는 방법으로 두 개체가 독립적이면 0에 가까워지며 의존적이면 1에 가까워진다[5]. 예를 들면, 두 단어 $i$, $j$가 있고 $p(i)$와 $p(j)$는 $i$, $j$가 문서에 각각 등장할 확률이고, $p(i,j)$는 문서에서 $i$, $j$가 동시에 등장할 확률이라고 하였을 때, $PMI$는 식 (1)과 같이 계산된다.

$$PMI_{i,j} = \log_2 \frac{P(i,j)}{P(i)P(j)} \quad (1)$$

자카드 유사도(Jaccard Similarity)는 두 집합 사이의 유사도를 측정하는 방법으로 두 집합이 유사하면 1에 가까워지며 유사하지 않으면 0에 가까워진다[6]. 예를 들어 두 집합 A={a, b, c}와 B={b, c, d}가 있다고 가정할 경우 자카드 유사도는 두 집합의 전체 요소 중 공통적으로 가지고 있는 요소의 비율을 나타낸 값으로 $b,c/a,b,c,d = 0.5$ 의 유사도를 지니게 된다.

### 2.2 한국어 범용 감성사전

'KNU 한국어 감성사전'은 특정 도메인에 대한 감성사전을 자동으로 생성하기 위해 구축되었으며 그 특성상 특정 도메인에 영향을 받지 않는 인간의 보편적인 기본 감정을 표현하는 감성 어휘들로 구성되어 있다[7]. 이와 같은 감성 어휘들은 대표적으로 '감동받다', '사랑하다', '미워하다' 등이 있다. 'KNU 한국어 감성사전'은 표준 국어 대사전[7]에 수록된 단어들을 설명하는 문장인 뜻풀이를 수집하여 감성사전 구축을 위한 데이터로 활용한다. 수집된 뜻풀이는 딥러닝 기법 중 하나인 Bi-LSTM(Bi-directional Long Short-Term Memory) 모델에 의해 극성이 긍정, 부정, 중립으로 분류한다. 분류가 수행되면 긍정에 대한 극성을 나타내는 뜻풀이들에서는 긍정 감성 어휘들을, 부정에 대한 극성을 나타내는 뜻풀이들에서는 부정 감성 어휘들을 추출한다. 뜻풀이 외에도 기존에 구축된 감성사전 데이터에서 도메인에 독립적인 감성 어휘들을 추출하여 감성사전을 확장하였으며 텍스트 데이터에서 주로 사용되는 신조어나 이모티콘에 대한 감성 어휘 또한 추출하여 감성사전에 수록하였다. 구축된 'KNU 한국어 감성사전'은 도메인에 독립적인 14,843개의 n-gram 형태의 감성 어휘들로 구성되어 있으며 이는 기존에 구축된 범용 한국어 감성사전과의 차별성으로 둘 수 있다. 이와 같은 'KNU 한국어 감성사전'은 도메인 감성사전을 구축하는데 있어 기초자료로 활용된다.

범용적인 도메인의 감성 분석을 위해 구축되어진 한국어 범용 감성사전이 존재한다. 대표적으로 DecoSelex라는 한국어 기반 감성사전은 오피니언 마이닝(Opinion Mining)을 위해 구축되었으며, SentiWordNet을 통해 감성 어휘를 추출하고 한국어 Deco 사전을 활용하여 확장하는 방식으로 구축되었다[8]. 하지만 현재 구축된 DecoSelex 감성사전은 제공되고 있지 않다.

오픈 한글은 단어에 대한 원형, 품사 그리고 감성에 대한 정보를 제공해주는 오픈 서비스이다[9]. 감성사전에 수록된 단어는 집단지성을 활용하여 긍정, 부정, 중립에 대해 사용자들이 투표하고 각 극성에 대한 값이 누적 산출됨에 따라 신뢰도가 높아지도록 설계되었다. 하지만 오픈 한글은 현재 오픈 서비스의 문제로 인해 서비스가 잠정 중단 되었다.

서울대에서 KOSAC 말뭉치를 활용하여 한국어 감정 어휘 목록(감성사전)을 구축하였다. 수록된 감정 어휘 목록은 한국어 감성 분석 연구에 활용되어질 수 있도록 형태소 단위의 감성 특성을 제공한 어휘 목록이다[10]. 하지만 이 어휘 목록은 도메인에 대한 고려 없이 감성을 부여하였으며, 형태소 단위로 제공되기 때문에 쉽게 활용할 수 없다는 문제가 존재한다.

K-LIWC는 글의 언어학, 심리학적 특징을 분석하기 위해 구축된 한국어 글 분석 프로그램이다. 이는 기존에 구축된 어휘들을 기반으로 한국어 글 분석을 통해 언어적 특징의 분석이 가능하다[11]. 하지만 현재 K-LIWC는 서비스가 제공되고 있지 않아 활용하기 어렵다.

SentiWordNet은 워드넷(WordNet)의 Synset이라는 유의어 집단의 어휘들을 유의어, 반의어 관계를 통해 확장하고 분류기로 학습하여 긍정, 부정, 객관성에 대한 극정의 정도를 부여한 감성사전이다[12]. 하지만 이는 단순히 어휘의 관계 확장과 분류기 학습을 통해 극성의 정도가 부여되었기 때문에 일부 어휘에 대한 극성 정도가 바르지 않다. 예로, '연뇌막'이라는 어휘는 긍정적인 감성을 가지고, '비난하다'라는 어휘는 객관적인 감성을 지닌다. 또한 부정 감성 어휘 계산에 큰 영향을 미치는 요소가 충분히 고려되지 않아 부정확한 감성 정도를 제공한다. 이와 같은 SentiWordNet의 감성 어휘를 한국어로 번역하여 사용할 경우 다음과 같은 문제점이 발생한다. 첫 번째, 한국어에도 단어인 것이 영어에서는 어구로 존재한다. 대표적으로 '신물나다'라는 단어는 'sick of'라는 어구로 존재한다. 두 번째, 두 어휘의 감성 정도의 값이 일치하지 않는 경우가 있다. 대표적으로 한국어에도 '노발대발하다'라는 어휘의 극성 정도는 7.7점이지만 동일한 의미인 'infuriate'라는 단어는 2.5점이다. 마지막으로 한국어에서 서로 다른 형태의 어휘들이 영어에서는 하나의 어휘로 대역된다. 예를 들면, '역정나다', '성질나다', '화나다', '노하다', '분하다', '성나다', '약오르다', '골나다'라는 어휘는 모두 'angry'라는 하나의 단어로 번역된다.

이와 같은 기존의 범용 감성사전의 문제점은 수록된 어휘의 개수는 많으나 각 어휘들이 도메인 의존 여부와 상관없이 감성이 측정되어 구축되어있다는 문제점을 지닌

다. 그렇기 때문에 특정 도메인을 분석하기 위해 이와 같은 범용 감성사전을 활용할 경우 부정확한 결과가 도출될 수 있다는 문제점이 존재한다. 또한 대부분의 범용 감성사전은 1-gram 형태의 감성 어휘들을 지니고 있기 때문에 어휘의 형태가 풍부하지 않다는 문제점을 지니고 있다.

### 2.3 도메인 감성사전 구축 방안

도메인에 대한 감성사전을 자동으로 구축하기 위해 의존 구문 분석 어휘 간 동시 출현 확률 및 접속 관계를 활용하여 감성 어휘를 자동으로 식별한 연구가 수행되었다. 본 연구는 어휘 간 동시 출현 확률을 PMI (Pointwise Mutual Information) 기법을 통해 계산하였으며 접속사의 순접 역접을 정의하고 이를 활용하여 극성 전파를 통해 도메인 감성 어휘를 자동으로 추출한다[13]. 하지만 본 연구는 활용한 기초 자료가 도메인에 대한 감성 어휘의 특성을 고려하지 않은 감성사전들을 기초 자료로 활용하지 않았으며, 이와 같은 문제로 인해 극성이 잘못 고려된 감성 어휘가 감성 어휘들의 극성 자동 부여에 있어 큰 문제점을 발생시킬 수 있다. 또한 PMI는 단순하게 두 단어의 지역적인 출현빈도만 지니고 있기 때문에 접속 관계를 활용한다 해도 전역적인 정보가 충분히 반영될 수 없을 수 있다는 문제점을 지닌다. 특정 감정에 대한 감정 사전을 구축하기 위해 Emotion-aware LDA를 제안한 연구가 수행되었다. 이 모델은 도메인에 독립적인 시드 어휘들을 활용하여 도메인 감정 어휘를 식별하는데 활용되었다. 하지만 본 연구는 어휘들의 문맥 관계를 고려하지 않아 부정확한 감정 어휘가 추출될 수 있다는 단점이 존재한다[14]. 도메인별 감성사전을 자동으로 구축하기 위해 후보 감성 어휘 간의 시맨틱 유사성을 파악하고 단어 그래프를 구성하여 그래프 기반의 반지도 학습 기반의 레이블 전파를 수행하여 후보 감성 어휘에 대한 극성을 자동으로 전파하는 연구가 수행되었다. 하지만 본 연구 또한 접속 관계와 같은 중요 문맥 정보 충분히 반영되지 않았다는 점에서 감성 어휘의 극성 역전 관계를 포착할 수 없다는 문제점이 존재한다[15]. 특정 분야의 감성사전을 자동으로 구축하기 위해 그래프 기반 준지도 학습 방법을 제안하고 어휘 간 동시 출현 확률을 계산하여 감성사전을 자동으로 구축한 연구가 수행되었다[16]. 어휘 간 동시 출현 확률을 계산하기 위해 PMI 기법을 활용하여 어휘 간 공기 정보를 토대로 어휘들의 연관성을 계산하여 감성사전을 구축하였다. 하지만 본 연구는 접속 관계를 고려하지 않고 어휘 간 연관성만 고려했다는 점에서 유사도가 높지만 접속 관계에 따라 감성이 역전되는 문제는 해결하지 못했다는 단점이 있다. 분석하고자 하는 특정 도메인 코퍼스에 최적화된 도메인 감성사전을 구축하기 위해 일부 감성 키워드를 선정하고 Word2Vec을 활용하여 후보 키워드를 추출, 추출된 후보 키워드를 활용하여 긍정, 부정 감성 어휘를 추출한 연구가 수행되었다[17]. 특정 도메인을 잘 대표하는 도메인 감성사전을 구축하기 위해 회귀 분석 기법 중 하나인 엘라스틱넷(ElasticNet)을 활용하여 각 단어의 회귀 계수를 구하고 이를 통해 감성사전을 자동으로 구축한 연구가 수행되었다[18]. 언어의 기본 단위인 단어에 대해 감정 정보를 부여하고 감정 정보가 부여된 단어를 통해 한국어 감정 어휘 사전을 구축하는 연구가 수행되었다. 이 연구에서는 기초 감정 어휘에 대해 사용자 설문 조사를 수행 하여 감정의 정도 값을 산출하고, 나머지 감정 어휘의 감정 정도 값을 자동으로 부여하기 위해 사전의 표제어 설명부(Gloss)를 이용하였다[19]. 그 동안 겪어온 개인의 경험에 근거하여 개인의 생애에 대한 주관적인 평가는 주관적 웰빙 상태라고 한다. 이러한 주관적 웰빙 상태의 측정을 위해 SentiWordNet을 대역하여 기본 감정 어휘 사전을 구축하고 온라인 뉴스 기사의 댓글을 수집을 통해 댓글의 긍정, 부정 파악에 도움이 되는 감정 어휘를 추출하였다. 추출된 감정 어휘를 활용하여 상황적 감정어 목록을 구축하였으며 이를 활용하여 주관적 웰빙 상태를 측정하는 연구를 수행하였다[20]. 오픈 한글 서비스의 범용 감성사전을 활용하여 도메인에 독립적인 어휘들을 제외하고 특정 도메인에 대한 어휘들의 중요도를 빈도수를 통해 측정하여 특정 도메인 감성 어휘 목록을 구축하고 도메인 감성지수 산출을 통해 특정 도메인에 알맞은 감성사전을 구축하는 연구가 수행되었다. 하지만 이 연구에서 기초 자료로 활용되었던 범용 한국어 감성사전은 오픈 한글의 서비스 중단으로 활용할 수 없다. 게임 도메인에 대한 감성사전을 구축하기 위해 게임 도메인 코퍼스를 수집하고 이를 통해 감성사전을 구축한 연구가 수행되었다[21]. 하지만 이 방법은 사람이 직접 감성을 측정하고 이들의 가중 평균을 통해 점수를 산출했다는 점에서 시간이 오래 걸리고 도메인의 크기가 커지면 구축이 어렵다면 문제점을 지니고 있다. 또한 비교 평가 부분에 있어 SentiWordNet을 활용하였는데 이는 한국어로 대역할 경우 대역된 한국어와 영어의 의미 정확도가 떨어지는 문제점이 있어 비교 평가로 부적절하다. 감성 분석 성능의 향상을 위해 분석하고자 하는 도메인 특성에 알맞은 도메인 감성사전 구축을 위한 연구가 수행되었다. 본 연구에서는 도메인을 영화로 지정하였으며, 영화 장르에 따라 감성 어휘가 상이한 영화 리뷰 데이터만을 활용하여 형용사를 추출하고 PMI를 활용하여 장르별 감성사전을 구축하였다[22]. 하지만 단순히 형용사에 대한 어휘만을 통해 감성사전을 구축했다는 점에서 감성 분석의 성능 향상을 위해 충분한 감성 어휘가 수록되었다고는 볼 수 없다.

## 3. 제안 방안

본 연구에서는 'KNU 한국어 감성사전'을 기초자료로 활용하여 도메인 감성사전을 자동으로 구축하는 방안에 대하여 제안한다. 그림 1은 도메인 감성사전 자동 구축을 위한 프로세스 이다. 첫 번째, 구축하고자 하는 도메인에 대한 코퍼스를 수집한다. 해당 코퍼스를 수집하는데 있어 주의해야할 점은 본 연구의 목표가 도메인에 독립적인 감성 어휘를 식별하고 극성을 부여하는 것이기 때문에 감성에 대한 내용이 들어갔을 확률이 높은 칼럼이나 사용자 후기 글들을 수집하는 것이 중요하다. 두 번째, 도메인 코퍼스의 형태소 분석을 통해 감성을 표현하는 품사를 지니는 어휘들을 추출하고 후보 감성 어휘들 중 'KNU 한국어 감성사전'에 수록되어 있는 어휘를 시드 어휘로 구성한다. 세 번째, 두 번째 단계에 의해 추출된 모든 어휘 간 유사도를 글로브 벡터(GloVe)를 활용하여 계산하고 이를 어휘 간 유사도 행렬로 구축한다. 이와 같은 유사도 행렬은 이후에 시드 감성 어휘와 후보 감성 어휘들 간의 관계를 파악하는데 활용된다. 네 번째, 어휘 간 접속 관계를 식별하고 어휘 간 접속관계 행렬을 구축, 감성의 보존 또는 역전을 위한 단서로 활용한다. 다섯 번째, 어휘 간 유사도 행렬을 활용하여 어휘 간 접속관계 행렬이 전역적인 정보를 가질 수 있도록 유사도 및 접속 관계 정보를 전파한다. 여섯 번째, 전역적인 정보를 지니는 어휘 간 접속 관계 행렬과 시드 데이터를 활용하여 후보 어휘의 극성을 부여한다. 마지막으로, n-gram 형태의 감성 어휘를 추출하기 위해 문장에 대한 의존 구문 분석과 어구 분석을 수행하고 이를 통해 n-gram 형태의 감성 어휘를 식별 및 극성을 부여한다.

### 3.1 도메인 감성사전 자동 구축 방안

감성 어휘들은 특정 도메인에 따라 감성이 존재하지 않을 수도 있으며, 감성이 역전되거나 일반 도메인에서 없던 감성이 부여될 수도 있다. 표 1은 각 특정 도메인에 대해 일반적인 감성과 도메인의 감성에 대해 비교한 예시이다. 이처럼 특정 도메인에 대해 감성사전을 활용하기 위해서는 도메인의 특성을 반영한 감성사전을 사용하는 것이 중요하다. 예를 들어, 일반적으로 사용되는 감성사전의 감성 어휘를 특정 도메인에 적용한다면 부정확한 결과가 도출될 수 있기 때문이다. 이와 같은 문제로 도메인 분석을 위해 기존에 연구자들이 직접 도메인 감성사전을 구축하고 이를 활용하는 사례가 있다. 하지만 이는 도메인 감성사전 구축 간에 많은 시간과 비용이 들며, 사람이 수작업으로 식별하는 과정이기 때문에 도메인의 감성 어휘들을 충분히 식별하는 것이 어려울 뿐만 아니라, 개인의 주관적인 견해에 의해 감성의

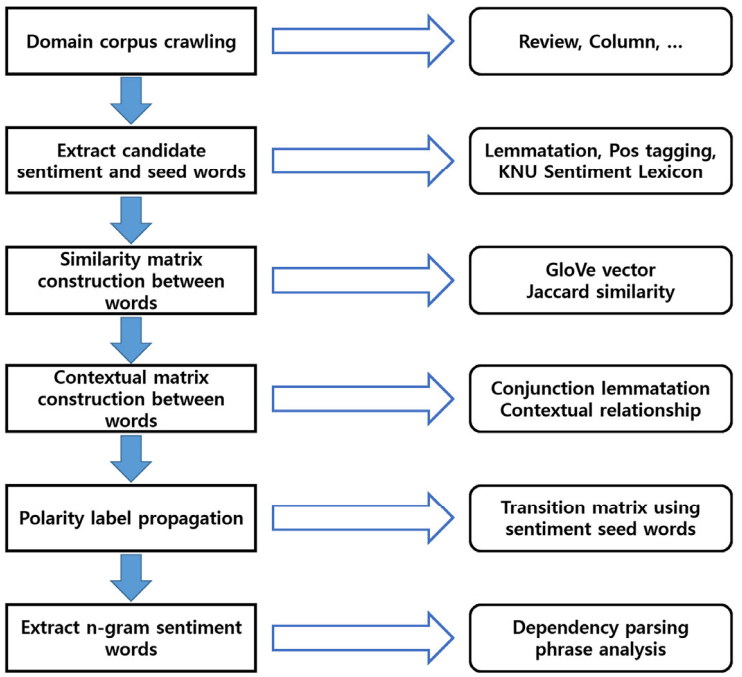


그림 1 도메인 감성사전 자동 구축을 위한 프로세스

Fig. 1 Process of the automatic domain specific sentiment lexicon construction

표 1 도메인 특정 감성 어휘
Table 1 Domain specific sentiment words

| Domain | Sentiment words | General polarity | Domain polarity |
|---|---|---|---|
| Car | Silence | Neutrality | Positive |
| | Hot | Neutrality | Negative |
| Movie | Sad | Negative | Positive |
| | Slept well | Positive | Negative |
| Clothes | Too big | Neutrality | Negative |
| | Sturdy | Neutrality | Positive |

극성이 부여되기 때문에 부정확한 극성이 부여될 수 있다는 문제점이 존재한다. 본 연구에서는 이와 같은 문제를 해결하기 위해 특정 도메인에 의존적으로 사용되는 감성 어휘를 자동으로 식별하고 감성에 대한 극성을 자동으로 부여하는 도메인 감성사전 자동 구축 방안을 제안한다. 그림 1은 제안 방안의 프로세스를 나타낸다. 제안 방안의 아이디어는 후보 감성 어휘들을 특정 품사를 통해 추출하고 후보 어휘 간의 유사, 문맥 관계를 파악한다. 그 후에, 'KNU 한국어 감성사전'을 활용하여 후보 감성 어휘에 대한 극성을 반지도 학습(Semi-supervised Learning)을 통해 부여한다. 또한 이와 같은 감성 어휘들을 활용하여 n-gram 형태의 감성 어휘를 식별, 감성을 자동으로 부여한다.

### 3.2 도메인 코퍼스 수집

감성사전은 감성 어휘들과 그 감성 어휘들에 대한 극성이 부여되어 있는 단어들의 모음집이다. 보통 감성사전을 구축하기 위해서는 사전을 구축하기 위한 도메인을 선정하고 해당 도메인 코퍼스를 수집한 후 수집된 도메인에서 식별되는 감성 어휘들을 추출하여 극성을 부여한다. 그렇기 때문에 많고 다양한 감성 어휘를 식별하기 위해서는 감성이 많이 들어가 있는 칼럼, 블로그 그리고 후기 댓글 등과 같이 게시자의 감성이 많이 내제되어 있을 것 같은 도메인 코퍼스를 수집하는 것이 중요하다. 본 연구에서는 이와 같이 감성 어휘가 많이 포함되어 있을 것으로 예상되는 자동차 후기 게시글과 영화 리뷰를 대상으로 연구를 수행한다.

### 3.3 전처리 및 접속사 원형 복원

본 연구에서는 도메인 코퍼스를 구성하는 어휘들의 일관성 유지를 위해 전처리를 수행한다. 예를 들면, '사랑해'와 '사랑함'과 같은 두 어휘는 사람이 보기에 같은 의미로 인식하지만 컴퓨터는 다른 의미로 인식하게 된다. 이처럼 두 어휘를 각각 '사랑'이라는 하나의 공통된 형태로 원형 복원을 수행하여 컴퓨터가 이와 같은 어휘들을 같은 어휘로 인식할 수 있게 한다. 또한 한국어의 특징 중 하나는 '그리고', '하지만' 등의 접속사가 문장과 결합하게 되면 형태가 변환되는 특징을 지니고 있다. 예

표 2 어휘에 대한 다양한 형태의 예
Table 2 Various types of words

| Lemma | Various word types |
|---|---|
| Love | Loved, Lovely, ... |
| Hate | Hated, dislike, ... |
| Delicious | Tasty, Appetizing, ... |

표 3 접속사에 대한 다양한 형태의 예
Table 3 Various types of conjunctions

| Lemma | Various conjunction types |
|---|---|
| 그리고 | ~이고, ~고 |
| 하지만 | ~이지만, ~지만, ~만 |
| 또는 | ~거나 |

를 들면, '그리고'와 같은 경우는 '나는 밥을 먹었고, 돈을 지불했다.'에서 '고'가 그 의미를 표현하고 있으며, '하지만'과 같은 경우는 '맛 집이라고 해서 갔지만 맛이 없었다.'에서 '지만'이 그 의미를 표현하고 있다. 이 또한 컴퓨터가 인식하는데 있어 다른 의미로 인식하는 문제점이 존재하기 때문에 본 연구에서는 도메인 코퍼스에서 접속사와 관련된 품사를 지니는 모든 형태소를 추출하고, 이들을 '그리고', '하지만' 등과 같은 하나의 접속어휘로 표현한다. 표 2와 표 3은 어휘 원형 복원과 접속사 원형 복원에 대한 예이다.

### 3.4 후보 감성 어휘 및 시드 감성 어휘 추출

후보 감성 어휘 추출 단계는 도메인 코퍼스의 어휘 중 감성을 가지고 있을 것으로 예상되는 어휘들을 추출하는 과정이다. 이와 같은 후보 감성 어휘들은 최종적으로 극성이 부여되어 도메인 감성사전을 구성하는 감성 어휘들이 되며, 어휘 간 유사도 행렬 구축, 접속 관계 전파 등에 활용된다. 후보 감성 어휘들을 추출하기 위해 우선 단어의 품사를 통해 감성을 지니는 후보 감성 어휘들을 추출한다. 주로 감성을 지니는 단어들은 형용사, 명사의 품사를 지닌다. 예를 들면, 형용사의 '사랑하는', '좋아하는', '훌륭한' 등이 있으며 명사의 예로는 '좋음', '멋짐', '날렵함' 등이 있다. 감성을 가지고 있을 것으로 판단되는 후보 감성 어휘들을 추출한 후 본 연구에서는 극성 레이블 전파를 위한 정보로 사용될 시드 감성 어휘를 추출한다. 시드 감성 어휘란 극성이 도메인에 독립적인 어휘들을 의미한다. 이와 같은 시드 감성 어휘를 추출하기 위해 본 연구에서는 'KNU 한국어 감성사전'을 활용한다. 'KNU 한국어 감성사전'은 인간의 일반적인 감성을 나타내는 보편적인(도메인에 의존적인) 어휘로 구성되어 있으며 총 14,843개의 n-gram 어휘로 구성되어 있다. 시드 감성 어휘를 추출하기 위해 'KNU 한국어 감성사전'에 수록되어 있는 어휘들을 후보 감성 어휘와 비교하여 후보 감성 어휘가 'KNU 한국어 감성

Algorithm 1: Extract seed sentiment words

Input: $CS, KS, s_c, k_s$

Output: $SS$

$CS$: Candidate sentiment words set; $s_c$: Candidate sentiment word;

$KS$: KNU sentiment lexicon words set; $k_s$: KNU sentiment lexicon word;

$SS$: Seed sentiment words set;

for $s_c$ in $CS$:

  if $s_c$ in $KS$:

    $SS$.add($s_c$) $CS$.delete($s_c$)

  else:

    continue

사전'에 수록되어 있는 경우 해당 어휘를 시드 감성 어휘로 추출한다. 이와 같은 방법을 통해 본 단계에서는 후보 감성 어휘 집합과 시드 감성 어휘 집합을 구축한다. 알고리즘 1은 후보 감성 어휘에서 시드 감성 어휘를 추출하는 방안에 대해 나타낸다.

### 3.5 후보 감성 어휘 간 유사도 행렬 구축

후보 감성 어휘 간 유사도 행렬의 구축은 어휘 간의 극성을 판단하는데 유용한 정보로 쓰인다. 예를 들면, 하나의 문서 또는 문장에서 '멋지다'와 '엣지있다'라는 단어의 동시 출현 빈도가 높으면 해당 두 어휘는 높은 연관성을 가지고 있다고 판단할 수 있다. 또한 '멋지다'라는 어휘가 시드 감성 어휘에 속해있고 그것의 극성이 긍정이라면 우리는 '엣지있다'라는 어휘 또한 긍정과 높은 연관성을 가지는 단어라고 예측할 수 있다. 반대로 '멋지다'와 '졸리다'라는 단어의 동시 출현 빈도가 낮으면 해당 두 어휘는 낮은 연관성을 가지고 있다고 판단할 수 있으며 '졸리다'라는 단어는 긍정과 낮은 연관성을 가지는 단어라고 예측할 수 있다. 본 연구에서는 이와 같은 어휘 간 유사도 행렬 구축을 위해 수집한 도메인 코퍼스를 글로브를 통해 학습한다. 글로브는 단어 동시 등장 정보를 보존하려는 임베딩 방법론으로 워드 투 벡터(Word2Vec)가 임베딩 된 두 단어 벡터의 내적이 코사인 유사도라면, 글로브는 동시 출현 확률을 나타낸다. 즉, 글로브는 통계 정보를 활용하여 워드 임베딩 모델을 구축하는 것을 목적으로 한다. 글로브는 통계 정보 활용을 위해 윈도우 사이즈 기반 동시 등장 행렬(Window based Co-occurrence)과 동시 등장 행렬을 기반으로

표 4 어휘 간 동시 등장 행렬

Table 4 Co-occur matrix between words

| Count | Good | Edge | Delicious | Hungry |
|---|---|---|---|---|
| Good | 0 | 3 | 0 | 0 |
| Edge | 3 | 0 | 0 | 0 |
| Delicious | 0 | 0 | 0 | 5 |
| Hungry | 0 | 0 | 5 | 0 |

학습을 수행한다. 동시 등장 행렬이란 단어 간 특정 윈도우 사이즈 내에서 두 단어가 동시 등장한 빈도수를 기록한 행렬이다. 표 4는 동시 등장 행렬의 예시이다. 동시 등장 확률(Co-occurrence Probability)은 특정 단어 i가 발생했을 경우 단어 k가 발생한 빈도수를 통해 계산한 조건부 확률 이다. 에서 타겟 단어를 i, 조건 단어를 k라고 가정할 경우, 동시 등장 행렬에서 I는 행의 모든 값의 합을 분모로, i행 k열의 값을 분자로 한 결과이다. 표 5는 동시 등장 확률에 대한 예시이다. 표 5에서 얼음이라는 단어가 주어졌을 경우 고체가 등장할 확률은 증기가 주어졌을 경우보다 동시 등장 확률이 높다. k가 물이거나 패션인 경우는 1에 유사한 값이 나온다. 이처럼 동시 등장 확률에서 관련성이 높거나 거의 없는 경우에는 1과 유사한 값으로 추출된다. 글로브는 이처럼 타겟 단어 k가 주어졌을 경우, 임베딩 된 두 단어 벡터의 내적이 두 단어의 동시 등장 확률 간 비율이 되도록 임베딩 하는 것을 목표로 한다. 표 5에서 보여주는 것과 같이 '고체'가 타겟 단어일 경우 '얼음'과 '증기'의 벡터 사이의 내적이 8.9가 되도록 하는 것이다. 본 연구에서는 이와 같이 동시 등장 정보를 나타내는 워드 임베딩

표 5 어휘 간 동시 등장 확률

Table 5 Simultaneous entrance probability between words

| Ratio | k=Soild | k=Gas | k=Water | k=Fashion |
|---|---|---|---|---|
| $P(k\|Ice)$ | 0.00019 | 0.000066 | 0.003 | 0.000017 |
| $P(k\|Steam)$ | 0.000022 | 0.00078 | 0.0022 | 0.000018 |
| $P(k\|Ice)/P(k\|Steam)$ | 8.9 | 0.085 | 1.36 | 0.96 |

기법인 글로브를 학습한 후 앞에서 추출한 후보 감성 어휘 집합과 시드 감성 어휘 집합들의 모든 어휘들의 벡터 값을 토대로 자카드 유사도(Jaccard similarity)를 수행하여 어휘 간 유사도 행렬을 구축한다. 자카드 유사도란 0~1 사이의 값을 가지며, 두 집합 사이의 유사도를 측정하는 데이터 마이닝 기법이다. 자카드 유사도는 두 집합 간의 유사도를 측정하는 기법으로 수식의 분자에는 두 집합의 교집합, 분모에는 두 집합의 합집합이 값으로 들어간다. 예를 들면, A={'a','b','c'}이고 B={'b','c','d'}인 경우 자카드 유사도는 $b,c/a,b,c,d=0.5$이다. 하지만 본 연구에서는 글로브 벡터를 자카드 유사도 측정을 위한 요소로 활용하기 때문에 분자를 각 벡터 요소들의 최소값, 분모를 각 벡터 요소들의 최대값으로 하여 자카드 유사도를 연산한다. 식 (2)는 두 집합 간 자카드 유사도 측정 방법을 나타내며 식 (3)은 벡터를 활용한 자카드 유사도 측정 방법을 나타낸다. 식 (2)와 식 (3)에서 $w,v$는 각각 두 벡터의 집합을 의미하며 $w_i,v_i$는 각 벡터 집합의 요소를 나타낸다.

$$sim_{jaccard}(w,v)=\frac{|w\cap v|}{|w\cup v|}=\frac{|w\cap v|}{|w|+|v|-|w\cap v|} \quad (2)$$

$$sim_{jaccard}(w,v)\approx\frac{\sum_{i=1}^{d}\min(w_i,v_i)}{\sum_{i=1}^{d}\max(w_i,v_i)} \quad (3)$$

글로브 벡터에 의해 유사한 단어들은 특정 단어와의 벡터 사이의 내적이 유사해야하기 때문에 유사한 단어들은 벡터의 좌표 평면상 유사한 위치에 분포할 것이며, 유사하지 않은 단어들은 벡터의 좌표 평면상 유사하지 않은 위치에 분포되어 있다. 그렇기 때문에 두 단어가 유사하게 등장했다면 1에 가까운 값이, 그렇지 않으면 0에 가깝지 않은 값이 추출된다. 이와 같은 자카드 유사도에 따라 어휘 간 유사도 행렬을 구축한다. 표 6은 어휘 간 유사도 행렬에 대한 예이다. 표 6에서 '엣지있다', '날렵하다'는 높은 유사도를 보이기 때문에 해당 두 어휘는 동시 출현 확률이 높은 단어로 볼 수 있다. 반대로 '엣지있다', '맛있다'는 낮은 유사도를 보이기 때문에 해당 두 어휘는 동시 출현 확률이 낮은 단어로 볼 수 있다. 이와 같이 구축된 유사도 행렬은 이후에 후보 감성 어휘의 극성을 자동으로 부여하기 위한 중요한 정보로 활용된다.

### 3.6 후보 감성 어휘 간 접속 관계 행렬 구축

접속사는 그 종류에 따라 문맥의 극성을 유지할 수도 있고 역전시킬 수도 있다. 표 7은 접속사에 따른 문맥 극성 변화의 예시이다. 표 7의 '하지만'과 '그렇지만'이라는 접속사는 두 문맥 간의 극성이 반대되는 관계에 주로 쓰이는 접속사이며, 각 어구에는 반대되는 극성을 가지는 감성 어휘가 사용된다(역접). 반대로 '그리고'와 '또는'이라는 접속사는 두 문맥 간의 극성이 동일한 관계에 주로 쓰이는 접속사이며, 각 어구에는 동일한 극성을 가지는 감성 어휘가 사용된다(순접). 이와 같이 감성 어휘 간의 관계를 파악하는데 접속사 정보는 유용한 정보로 쓰일 수 있다. 본 연구에서는 어휘 간 접속 관계 행렬 구축을 위해 두 감성 어휘 간 접속사가 순접으로 연결되어 있는 경우 어휘 간 접속 관계 행렬에 1의 값을, 두 감성 어휘 간 접속사가 역접인 경우 어휘 간 접속 관계 행렬에 -1의 값을 부여하고 만약 두 감성 어휘가 접속사로 연결되어 있지 않은 경우에는 0의 값을 부여한다. 표 8은 어휘 간 접속 관계 행렬에 대한 예시이다. 표 8에서 '엣지있다'와 '날렵하다'는 감성을 유지하는 순접 관계로 연결되어 있는 것을 볼 수 있다. 그렇기 때문에 두 단어는 같은 극성을 가질 확률이 높은 것으로 알 수 있다. 반대로 '깡통'과 '엣지있다'는 감성을 역전하는 역접 관계로 연결되어 있는 것을 볼 수 있다. 그렇기 때문에 두 단어는 다른 극성

표 6 어휘 간 유사도 행렬

Table 6 Similarity matrix between words

| | 엣지있다 | 날렵하다 | 깡통 | 맛있다 | 비싸다 |
|---|---|---|---|---|---|
| 엣지있다 | 1 | 0.88 | 0.2 | 0.1 | 0.53 |
| 날렵하다 | 0.88 | 1 | 0.12 | 0.08 | 0.5 |
| 깡통 | 0.2 | 0.12 | 1 | 0.24 | 0.33 |
| 맛있다 | 0.1 | 0.08 | 0.24 | 1 | 0.75 |
| 비싸다 | 0.53 | 0.5 | 0.33 | 0.75 | 1 |

표 7 접속사에 따른 문맥 극성 변화

Table 7 Contextual polarity change by conjunction

| Conjunction | Content | Label |
|---|---|---|
| 그리고 | 동네식당의 스테이크는 맛있고(긍정) 저렴하다(긍정). | Direct(1) |
| 하지만 | 아웃백의 스테이크는 맛있지만(긍정) 비싸다(부정). | Reverse(-1) |
| 또는 | 이 제품이 인기있는(긍정) 이유는 디자인 또는 저렴한가격(긍정) 때문이지 않을까? | Direct(1) |
| 그렇지만 | 넌 나의 충성스러운(긍정) 부하였어. 그렇지만 넌 나에게 모욕감(부정)을 줬어. | Reverse(-1) |

표 8 어휘 간 접속 관계 행렬
Table 8 Contextual matrix between words

| | Edge | Sharp | Can | Defective | Repair |
|---|---|---|---|---|---|
| Edge | 0 | 1 | -1 | 0 | 0 |
| Sharp | 1 | 0 | -1 | 0 | 0 |
| Can | -1 | -1 | 0 | 1 | 1 |
| Defective | 0 | 0 | 1 | 0 | 1 |
| Repair | 0 | 0 | 1 | 1 | 0 |

을 가질 확률이 높은 것으로 알 수 있다. 이와 같은 어휘 간 접속 관계 행렬은 두 단어의 관계에 대한 정보를 가지고 있으며 감성 어휘 간의 관계를 파악하고 극성 정보를 부여하는데 중요한 정보로 활용된다.

### 3.7 후보 감성 어휘 간 유사도 및 접속 관계 전파

글로브 벡터를 통해 구축된 어휘 간 유사도 행렬은 전역적인 정보를 가지고 있으나 접속 관계 정보가 들어있지 않은 정보를 가지고 있다. 그렇기 때문에 반대되는 감성을 가진 어휘 관계라도 동시 출현 확률이 높으면 1에 가까운 유사도를 지니고 있다. 또한 접속사 관계에 의해 구축된 어휘 간 접속 관계 행렬은 두 단어 간의 접속 관계에 의해서만 구축되었기 때문에 지역적인 정보만 지니고 있다. 그렇기 때문에 두 어휘 간의 지역적인 접속 관계만을 표현할 수 있다는 제약이 있다. 본 연구에서는 어휘 간 유사도 행렬이 지니는 동시 출현 확률에만 의존하는 한계점과 어휘 간 접속 관계 행렬이 지니는 지역적인 정보에 대한 한계점을 상호 보완하기 위해 어휘 간 접속 관계 행렬에 어휘 간 유사도 행렬을 업데이트함으로써 전역적인 접속 관계와 동시 출현 확률에 대한 정보를 보존할 수 있게 한다. 식 (4)와 5는 어휘 간 유사도 행렬을 업데이트하기 위한 수식이다.

$$E_v^{(t+1)} = \alpha S E^{(t)} + (1-\alpha) E^{(0)} \tag{4}$$

$$E_h^{(t+1)} = \alpha E_h^{(t)} S + (1-\alpha) E_v^{*} \tag{5}$$

식 (3)의 $E^{(0)}$는 초기의 어휘 간 접속 관계 행렬이며, $S$는 어휘 간 유사도 가중치 행렬이다. $S$는 식 (6)과 (7)에 의해 구축된다. 식 (6)의 $A$는 어휘 간 유사도 행렬이며 이를 통해 $A$의 가중치 행렬 $W$를 구축한다. 구축한 $W$를 i번째 행의 모든 값들의 합이 (i,i)의 입력으로 하는 대각 행렬인 $Z$를 구축하고 식 (7)에 의해 $S$를 생성한다. 식 (7)을 통해 생성된 $S$는 $W$를 대칭적으로 만들어주는 역할을 한다.

$$W_{ij} = \frac{A(x_i, x_j)}{\sqrt{A(x_i,*)}\sqrt{A(x_j,*)}} \\ , j \neq i \text{and} (x_i, x_j) \geq 0 (otherwise: 0) \tag{6}$$

$$S = Z^{-1/2} W Z^{-1/2} \tag{7}$$

$\alpha$는 현재까지 업데이트 된 접속 관계 행렬에 대한 가중치이며 $E_v^{(t+1)}$는 구하고자 하는 t+1번째 수직으로 업데이트 된 접속 관계 행렬이다. 이와 같은 접속 관계 행렬의 수직 전파 업데이트는 값이 수렴할 때 까지 업데이트를 수행한다. 식 (4)의 $E_v^{*}$는 식 (3)에 의해 최종적으로 수직 전파가 수행된 어휘 간 접속 관계 행렬이며, $E_h^{(t+1)}$는 구하고자 하는 t+1번째 수평으로 업데이트 된 접속 관계 행렬이다. 이와 같은 수직 전파가 완료된 접속 관계 행렬의 수평 전파 업데이트는 값이 수렴할 때 까지 업데이트를 수행한다. 결과적으로 식 (3)과 (4)를 통해 지역적인 정보만을 가지고 있던 어휘 간 접속 관계 행렬은 전역적인 정보를 포함하는 접속 관계 행렬로 업데이트 된다.

### 3.8 반지도 학습을 통한 후보 감성 어휘 극성 레이블 전파

글로후보 감성 어휘의 극성 레이블을 전파하여 후보 감성 어휘에 극성을 자동으로 부여하기 위해 수직, 수평 업데이트가 완료된 어휘 간 접속 관계 행렬과 시드 감성 어휘를 활용한다. $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$을 시드 감성 어휘의 집합이라 가정하고 $Y = \{y_{pos}, y_{neg}\}$이다($y_{pos}$는 긍정으로 1로 표기, $y_{neg}$는 부정으로 -1로 표기). $U = \{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$는 감성을 모르는 감성 후보 어휘라 가정한다. 그러므로 $X = l + u$이고 $X$는 시드 감성 어휘와 감성 후보 어휘의 전체 집합이다. $Y_L = \{y_1, y_2, \ldots, y_l\}$은 시드 감성 어휘에 대한 극성이며, $Y_U = \{y_{l+1}, y_{l+2}, \cdots, y_{l+u}\}$는 감성 후보 어휘에 대한 극성으로 0으로 초기화된 값을 넣는다(감성 후보 어휘에 대한 극성은 모르기 때문). 따라서 감성 후보 어휘의 극성을 탐지하기 위해서는 $X$와 $Y_L$에 기반 하여 우리는 감성 후보 어휘의 극성인 $Y_U$를 찾는다. 어휘 간 유사도 행렬 $A$는 $E^{*}$와 함께 레이블 전파에 활용되며, $A$는 식 (8)과 같이 재조정된다.

$$\widetilde{A_{ij}} = \begin{cases} 1-(1-E_{ij}^{*})(1-A_{ij}), E_{ij}^{*} \geq 0 \\ (1+E_{ij}^{*})A_{ij}, E_{ij}^{*} < 0 \end{cases} \tag{8}$$

표 9 식별된 감성 어휘에 따른 탐지된 구문과 추출된 n-gram
Table 9 Detected phrase and Extract n-gram by identified sentiment words

| Identified sentiment words | 아름다운 |
|---|---|
| Detected phrase | 아름다운 디자인을 가지고 있는 자동차이다. |
| Extracted n-gram | 아름다운 디자인을, 아름다운 디자인을 가지고 있는 |

Algorithm 2: n-gram polarity compute

Input: $S, s_w, s_p, N, n_w$

Output: $n_p$

$S$: Set of sentiment words; $s_w$: Sentiment words; $s_p$: Polarity of sentiment words;

$N$: Set of n-gram sentiment words; $n_w$: n-gram sentiment words; $n_p$: Polarity of n-gram sentiment words;

for $n_w$ in $N$:

  for $s_w$ in $S$:

    if $n_{w_i} = s_{w_i}$:

      $n_p = n_p + s_p$

  if $n_p > 0$:

    $n_p == 1$

  else if $n_p < 0$:

    $n_p == -1$

  else if $n_p == 0$:

    pass

$\tilde{A}$은 $A_{ij}$ 사이의 접속사가 순접($E^*_{ij} > 0$)인 관계이면 해당하는 값을 증가시키고, $A_{ij}$ 사이의 접속사가 역접($E^*_{ij} < 0$)인 관계이면 해당하는 값을 감소시킨다. 구축된 $\tilde{A}$를 통해 전이 확률 행렬 $T_{|X|\times|X|}$를 식 (9)에 의해 구축한다. 전이 확률(Transition Probability)이란 하나의 상태가 다른 상태로 전이될 확률을 의미한다. 즉, 식 (9)는 어휘 $x_i$와 어휘 $x_j$ 간의 관계를 일반화한 확률을 의미한다.

$$T_{ij} = p(i \rightarrow j) = \frac{\tilde{A}(x_i, x_j)}{\sum_{k=1}^{|X|} \tilde{A}(x_i, x_j)} \quad (9)$$

$f^t = \{f^t_1, f^t_2, \cdots, f^t_{|X|}\}$를 t번째 반복에 의해 계산된 $X$의 극성의 값이라고 가정하고, $f^0$를 초기 극성 값이라고 가정하자. $f^0$에는 시드 감성 어휘들의 값이 -1 또는 1로 구성되어 있을 것이며, 감성 후보 어휘들의 값은 0으로 구성되어 있다. 감성 후보 어휘들의 감성 레이블 전파는 식 (9)와 (10)에 의해 수행되며, 식 (9)와 (10)의 값이 수렴할 때까지 반복된다. 식 (9)에서는 $X$에 속한 감성 어휘를 그들의 이웃에 따라 레이블 전파를 수행한다. $\lambda$는 레이블 전파를 위한 가중치이며 0에서 1사이의 값을 지닌다. 식 (9)가 한번 수행되면 이어서 식 (10)을 통해 시드 감성 어휘의 초기 극성 값이 유지된다. $f^{t+1}_l$의 값이 수렴되면, 감성 후보 어휘에 대한 값이 음수인 경우는 극성을 부정으로, 값이 양수인 경우는 극성을 긍정으로 그리고 값이 0인 경우는 극성을 중립으로 취한다. 그리고 극성이 부여된 후보 감성 어휘는 도메인 감성 어휘라 명명한다.

## 4. 실험 및 평가

### 4.1 도메인 감성 어휘 추출 결과

본 연구에서는 자동차 도메인과 영화 리뷰 도메인의 데이터를 활용하여 도메인 감성 어휘 추출을 하였다. 자동차 리뷰에서는 118,302개의 어절을 활용하여 총 790개의 감성 어휘를 추출했고, 이중 442개의 도메인 감성 어휘를 식별, 극성을 자동으로 부여하였다. 또한 영화 리뷰 도메인에서는 300,000개 어절의 영화 리뷰 데이터를 활용하여 총 1878개의 감성어휘를 추출했고, 이중 1,099개의 도메인 감성 어휘를 식별, 극성을 자동으로 부여하였다. 이와 같은 통계를 통해 특정 도메인에서는 특정 감성 어휘가 보편적인 감성 어휘보다 많이 활용되는 것을 확인할 수 있었고, 이를 통해 특정 도메인에 특화된 감성 어휘를 식별하는 것이 도메인 텍스트를 분석하는데 있어 중요한 요소 중 하나인 것을 확인할 수 있었다. 그림 2는 n-gram을 탐지하기 위한 구문을 추출하는 예시로 감성 어휘가 포함된 구문을 의존 구문 분석을 통해 추출하고 표 9와 같이 식별한다. 표 10은 자동차 도메인과 영화 도메인에 대해 자동으로 추출된 1-gram, n-gram의 감성 어휘를 나타낸다. 1-gram 긍정 어휘 중 '빨르다'는 '빠르다'의 오타로 예상된다. 이것을


그림 2 의존 구문 분석 결과

Fig. 2 Result of the dependency parsing

표 10 형태에 따른 도메인 감성 어휘
Table 10 Domain sentiment words by types

| Domain | Pol | 1-gram | n-gram |
|---|---|---|---|
| Car | Pos | 날렵하다 | 고성능 타이어 |
| | | 빨르다 | 최상급 풀 옵션 사양 |
| | | 엣지있다 | 고급 세단과 유사한 수치 |
| | Neg | 미비하다 | 브랜드 격차의 한계 |
| | | 부진하다 | 브랜드 이미지의 한계 |
| | | 불과하다 | 앞서 제기된 아쉬움 |
| Movie | Pos | 끝내주다 | 고전 명작 |
| | | 절묘하다 | 인생 최고의 영화 중 |
| | | 몰두하다 | 멋진 동시 같은 영화 |
| | Neg | 추잡 | 답 없는 영화임 보면 후회함 |
| | | 처절하다 | 결말 왕 실망임 |
| | | 혼미하다 | 그냥 킬 링 타임용 |

표 11 추출된 도메인 감성 어휘의 통계
Table 11 Statistics of extracted domain sentiment words

| Domain | 1-gram | n-gram | Total |
|---|---|---|---|
| Car | 341 | 101 | 442 |
| Movie | 1,197 | 681 | 1,878 |

감안하면 사용자가 사용한 감성 어휘들 중 사용 빈도가 높고 오타인 감성 어휘 또한 제안 방안에 의해 자동으로 식별되는 것을 알 수 있다. 이처럼 오타뿐만이 아니라 게시자들에 의해 높은 빈도로 사용될 다양한 신조어 및 은어들 또한 식별 가능할 것으로 예상된다. 또한 제안 방안에 의해 자동으로 추출된 n-gram 감성 어휘들이 잘 식별되었다는 것을 알 수 있었다. 표 11은 도메인 코퍼스에서 추출된 각 형태에 대한 감성 어휘의 통계이다. 자동차 도메인의 1-gram은 총 341개의 감성 어휘들이 추출되었으며 n-gram 형태는 101개로 구성되어 있으며 영화 리뷰 도메인은 1-gram에서는 1,197개, n-gram은 681개로 구성되어 있다. 감성 어휘들의 통계치를 통해 n-gram 형태의 감성 어휘 또한 많이 식별 되는 것을 알 수 있었으며, 이처럼 1-gram 이외에도 n-gram 감성 어휘를 식별하는 것이 중요하다는 것을 알 수 있었다.

### 4.2 도메인 감성 어휘 정확도

표 12는 본 연구의 제안 방안에 의해 자동으로 구축된 도메인 감성사전의 정확도와 도메인 감성 어휘의 정확도를 나타낸다. 오토뷰 자동차 도메인과 네이버 감성 무비 코퍼스의 도메인을 활용하여 실험을 진행하였으며, 정확도 측정은 각 감성 어휘에 대해 연구자들이 직접 보고 극성을 부여한 후 자동으로 부여된 극성과 연구자에 의해 부여된 극성의 일치도를 통해 측정되었다. 자동차 도메인 감성사전의 정확도는 92.3%이며, 자동으로 극성이 부여된 도메인 감성 어휘의 정확도는 86.6%로 대부분의 극성이 일치하는 것을 확인할 수 있었다. 영화 도메인 감성사전의 정확도는 80.3%이며, 자동으로 극성이 부여된 도메인 감성 어휘의 정확도는 71% 자동차 도메인보다 낮은 정확도를 보였다. 이를 분석 해본 결과, 영화 리뷰 도메인에 활용된 코퍼스는 띄어쓰기 및 오탈자가 많아 형태소 분석과 원형복원이 원활하게 이루어지지 않았으며 리뷰 형태의 데이터다 보니 접속 관계를 충분히 포착하기 힘들었다. 또한 반어법의 사용으로 인해 어휘 간 관계를 잘 포착하지 못하는 문제점도 있었으며 마지막으로 특정 시드 어휘가 영화라는 도메인에서 극성이 역전되는 경우가 있어 감성사전 구축에 있어 약간의 제약이 있었다. 하지만 구축된 감성사전이 80.3%이상의 정확도를 보이는 것을 통해 정제되지 않은 데이터라도 어느 정도 활용 가능하다는 것을 알 수 있었다. 이를 통해 제안 방안을 통해 자동으로 구축된 도메인 감성사전의 실제 활용 가능성을 확인할 수 있었다. 표 13은 이와 같은 영화 리뷰 도메인이 도메인 감성사전 생성 간에 가지는 제약에 대한 예시이다. 표 14는 제안 방안과 기존 연구를 각각 활용하여 자동차 도메인에 대한 도메인 감성사전을 자동으로 생성하고 성능 비교 평가를 한 표이다. 기존 연구는 PMI 기법을 활용하여 어휘들의 연관 유사도를 고려하여 특정 도메인의 감성 어휘를 추출하였다. 감성 어휘의 정확도는 제안방안이 86.6%로 기존 연구보다 약 18.6%이상 높은 정확도를 보였으며 추출된 도메인 감성 어휘는 341개로 368개 많

표 12 도메인 간 감성 어휘 및 감성사전의 정확도
Table 12 Accuracy of the domain sentiment lexicon

| | Domain | Domain sentiment word | Domain sentiment lexicon |
|---|---|---|---|
| Acc | Car | 86.6% | 92.3% |
| | Movie | 71% | 80.3% |

표 13 도메인 감성사전 생성 간에 가지는 제약에 대한 예시
Table 13 Example of the constraint among construct domain specific sentiment lexicon

| Content | Cause |
|---|---|
| 여운이남는영화...너무슬프다 ...단순한에로영화랑은 차원이다른영화 | Seed word error |
| 대단하다1시간40분이지루하지않았다와 진짜 ㅋㅋ | Pos tagging error |
| 으어어어ㅇ어ㅈ어ㅈㄷㄹ어어엉무섭다 무서워 1점!! | Irony |
| 1편은 차암 재밌었는데... | Context constraint |

표 14 제안 방안과 기존 연구의 비교 평가
Table 14 Comparison of the proposed method and baseline

| | Baseline | Proposed method |
|---|---|---|
| Word accuracy | 68% | 86.6% |
| The number of sentiment words (1-gram) | 74 | 341 |
| The number of sentiment words (n-gram) | X | 101 |
| The number of total sentiment words | 74 | 442 |

았다. 기존연구와 다르게 n-gram 정보 또한 추출하여 다양한 감성 어휘의 형태를 수집하였다. 이와 같은 결과를 통해 문맥 접속 관계와 전역적인 어휘들의 연관 유사도를 고려한 것이 도메인 감성 어휘들을 추출하는데 결정적인 역할을 한다는 것을 알 수 있었으며, 제안 방안의 효용성을 입증하였다. 표 15는 제안 방안에 도메인의 특성이 고려되어 자동으로 구축된 도메인 감성사전과 'KNU 한국어 감성사전'을 활용해 사전 기반의 감성 분석을 수행하고 이를 비교 평가한 결과이다. 실험 결과에 따르면 도메인의 특성을 고려하여 도메인 감성 어휘를 수록한 도메인 감성사전은 도메인에 독립적인 감성 어휘들이 수록된 'KNU 한국어 감성사전'보다 새로운 감성 문장을 식별하거나 'KNU 한국어 감성사전'이 잘못 식별한 감성 문장을 올바르게 식별하는 것을 알 수 있었다. 이와 같은 실험 결과에 따라 특정 도메인을 분석하는데 있어 도메인의 특성이 충분히 고려된 도메인 감성 어휘를 활용한 도메인 감성사전을 구축하여 활용하는 것이 효과적이라는 것을 알 수 있다.

### 4.3 도메인 감성 어휘 추출 결과 분석

본 연구에서는 도메인 감성사전과 한국어 범용 감성사전을 자동차 도메인에 적용하여 감성 문장의 식별 및 상위 10개의 감성 어휘 비교 평가를 수행하였다. 그림 3은 자동차 도메인에서 감성 문장 분류 실험 결과이다. 자동차 도메인에서 감성 분석에 있어 도메인 감성사전을 활용한 경우 한국어 범용 감성사전을 활용한 것보다

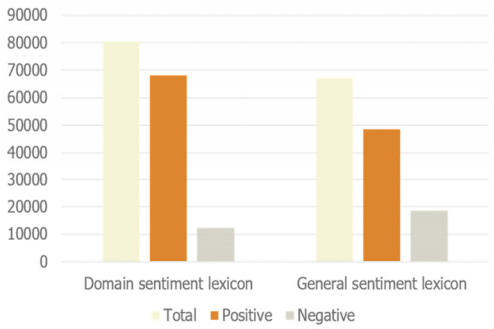


그림 3 자동차 도메인 감성사전과 한국어 범용 감성사전을 통한 감성분장 식별 결과
Fig. 3 Results of the sentiment sentence identified using the domain sentiment lexicon and KNU sentiment lexicon in the car domain

표 15 일반 감성사전과 도메인 감성사전의 감성 분석 비교 평가
Table 15 Comparison of the sentiment analysis using the general and domain sentiment lexicon

| Domain | Sentence | Domain sentiment lexicon | KNU sentiment lexicon |
|---|---|---|---|
| Car | 넓고 전면 유리는 배려심이 깊으며 슬라이딩 방식의 햇빛 가리개를 준비 해 두었다. | Positive | Negative |
| | 기간 해외 판매는 33만 2339대로 전년 동기 대비 5.7% 줄었다. | Negative | Neutrality |
| | 쌍용차에 빼앗긴 내수 3위의 자리를 탈환하기 위해 한국 GM은 이미 높은 할인율을 제공하고 있다. | Positive | Neutrality |
| | 차선 이탈 방지 시스템이 차선을 유지할 수 있게 돕는다. | Positive | Neutrality |
| | 지난해와 같은 달보다 17.5% 줄어든 수치다. | Negative | Neutrality |
| | 승객을 위한 공기청정기와 실내 등, 컵홀더까지 제대로 갖추었다. | Positive | Neutrality |
| Movie | 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다 | Positive | Neutrality |
| | 절대 평범한 영화가 아닌 수작이라는걸 말씀드립니다 | Positive | Neutrality |
| | 졸쓰레기 진부하고말도안됌ㅋㅋ 아..시간아까워 | Negative | Neutrality |
| | 아직도 이 드라마는 내인생의 최고! | Positive | Neutrality |
| | 이 영화가 왜 이렇게 저평가 받는지 모르겠다 | Positive | Neutrality |
| | 이틀만에 다 봤어요 재밌어요 근데 차 안에 물건 넣어 조작하려고 하면 차 안이 열려있다던지 집 안이 활짝 열려서 아무나 들어간다던가 문자를 조작하려고하면 비번이 안 걸려있고 ㅋㅋㅋ 그런 건 억지스러웠는데 그래도 내용 자체는 좋았어요 | Positive | Neutrality |

표 16 감성사전 별 상위 10개의 감성 어휘 식별 결과
Table 16 Result of the identification of the top-10 sentiment words using each of the sentiment lexicons

| Rank | General sentiment lexicon | Domain sentiment lexicon |
|---|---|---|
| 1 | 빠르다 | 아쉬움 |
| 2 | 만족하다 | 빠르다 |
| 3 | 부드럽다 | 만족하다 |
| 4 | 고급 | 부드럽다 |
| 5 | 부족하다 | 소음 |
| 6 | 충분하다 | 고급 |
| 7 | 아쉽다 | 부족하다 |
| 8 | 어렵다 | 충분하다 |
| 9 | 뛰어나다 | 아쉽다 |
| 10 | 안전하다 | 개선하다 |

표 17 상위 10개의 n-gram 감성 어휘 식별 결과
Table 17 Result of the identification top-10 n-gram sentiment words

| Rank | Domain sentiment words (n-gram) |
|---|---|
| 1 | 대한 아쉬움 |
| 2 | 고급 차 |
| 3 | 최고 수준 |
| 4 | 고급 세단 |
| 5 | 고급 유 |
| 6 | 개선됐다는 것 |
| 7 | 최상급 트림 |
| 8 | 고급 브랜드 |
| 9 | 고급 가죽 |
| 10 | 문제 발생 |

10,000개 이상의 문장을 더 많이 식별하였다. 또한 표 16은 출현 빈도에 의한 상위 10개의 감성 어휘의 비교 평가이다. 도메인 감성사전을 활용한 결과 범용 감성사전에서는 출현되지 않은 '아쉬움', '소음', '개선하다'와 같은 도메인에 의존적으로 볼 수 있는 어휘들이 새로 식별 된 것을 알 수 있다. 특히 표 17이 나타내는 것과 같이 제안 방안에 의해 추출된 n-gram 감성 어휘들은 도메인에 매우 의존적인 것을 알 수 있었다. 또한 표 17에서 '개선됐다는 것'의 오타인 '개선됬다는 것'이라는 어휘가 상위 6번째의 빈도로 출현되었다는 것을 통해 오탈자이지만 사람들이 자주 사용하는 어휘 또한 식별 가능하다는 것을 확인하였다. 이처럼 도메인 감성사전을 활용하면 범용 감성사전에서는 식별할 수 없는 도메인에 의존적인 부분들을 추가로 식별할 수 있다는 것을 실험을 통해 증명하였다. 표 18은 영화 리뷰 도메인에서의 감성분석 결과이다. 도메인 감성사전의 사전기반 감성분석 결과가 범용 감성사전 보다 약 6% 이상 높은

표 18 감성사전 별 감성분석 정확도 및 식별 감성 문장 통계
Table 18 Statistics of the accuracy and identified sentiment sentence by each sentiment lexicons

| | Accuracy | The number of identified sentiment sentence |
|---|---|---|
| General sentiment lexicon | 46.9% | 103,427 |
| Domain sentiment lexicon | 52.8% | 115,606 |

성능을 보였으며, 식별된 감성문장의 개수가 12,179개 더 많았다는 것을 보아 도메인 감성사전이 범용 감성사전보다 높은 성능과 더 많은 감성 문장을 식별할 수 있다는 것을 알았으며, 이를 통해 도메인 감성사전의 효용성을 입증하였다.

### 4.4 도메인 감성사전 통계 평가

표 19와 20은 제안 방안을 통해 생성된 도메인 감성사전과 기존에 존재하는 한국어 범용 감성사전의 t-검정 수행 결과이다. 성인 10명을 대상으로 설문조사를 수행하였으며 값은 리커트 척도에 의해 산출되었다. 1에 가까운 값은 도메인을 대표하는 단어가 적다는 의미이며 5에 가까운 값은 도메인을 대표하는 단어가 풍부한 의미이다. 귀무가설은 "도메인 감성사전과 한국어 범용 감성사전은 도메인 감성 어휘들을 비슷하게 가지고 있다"이며, 대립가설은 "도메인 감성사전은 한국어 범용 감성사전 보다 도메인 감성 어휘들을 풍부하게 가지고 있다"이다. 유의 수준은 0.05에서 수행되었으며, 단측 검증을 수행하였다. 수행결과 두 도메인의 p-값이 모두 유의수준보다 낮은 결과가 산출된 것을 보아 귀무가설이 기각되

표 19 자동차 도메인에서의 도메인 감성사전과 한국어 범용 감성사전 간의 t-검정
Table 19 The t-test between the domain and general sentiment lexicon in the car domain

| Car domain | Domain sentiment lexicon | General Sentiment lexicon |
|---|---|---|
| Mean | 4.5 | 3.2 |
| Variance | 0.5 | 0.4 |
| Observed value | 10 | 10 |
| Pearson C.C | 0.745355992 | - |
| Mean difference | 0 | - |
| degrees of freedom | 9 | - |
| t-statistics | 8.510497719 | - |
| P(T<=t) one-tailed test | 6.72948E-06 | - |
| t-reject value one-tailed test | 1.833112933 | - |
| P(T<=t) two-tail test | 1.3459E-05 | - |
| t-reject value two-tail test | 2.262157163 | - |

표 20 영화 리뷰 도메인에서의 도메인 감성사전과 한국어 범용 감성사전 간의 t-검정

Table 20 The t-test between the domain and general sentiment lexicon in the movie domain

| Movie domain | Domain sentiment lexicon | General Sentiment lexicon |
|---|---|---|
| Mean | 3.9 | 2.2 |
| Variance | 0.766666667 | 0.177778 |
| Observed value | 10 | 10 |
| Pearson C.C | 0.361157559 | - |
| Mean difference | 0 | - |
| degrees of freedom | 9 | - |
| t-statistics | 6.529880877 | - |
| P(T<=t) one-tailed test | 5.38251E-05 | - |
| t-reject value one-tailed test | 1.833112933 | - |
| P(T<=t) two-tail test | 0.00010765 | - |
| t-reject value two-tail test | 2.262157163 | - |

었다는 것을 알 수 있었다(대립가설 채택). 이와 같은 결과는 제안 방안을 통해 생성된 도메인 감성사전은 해당 도메인에 있어 한국어 범용 감성사전보다 그 도메인을 잘 대표하는 감성 어휘들이 풍부하다는 것을 알 수 있었다.

## 5. 결론 및 향후 연구

기존에 구축되어 공개된 감성사전은 도메인에 대한 특정 어휘들의 극성이 고려되지 않았으며, 특정 도메인에서만 쓰이는 감성 어휘는 식별되지 않은 채 구축되어 있다. 이와 같은 기존의 감성사전으로 특정 도메인에 대한 감성 분석을 수행하게 된다면 해당 도메인의 감성 식별이 어렵거나 부정확한 분석 결과가 도출될 가능성이 있다. 특히, 특정 도메인을 조사하는 시장 조사나 여론 조사 그리고 마케팅 등의 분야에서의 부정확한 분석 결과는 치명적인 문제점을 야기 시킬 수 있다. 그렇기 때문에 특정 분야에 대해 감성사전 기반의 정확한 감성 분석을 수행하기 위해서는 도메인에 알맞은 감성사전을 구축하고 활용하는 것이 유용하다. 본 논문에서는 기존의 감성사전의 문제점을 해결하기 위해 특정 도메인에 대한 감성사전을 자동으로 구축하는 알고리즘을 제안한다. 도메인 감성사전을 자동으로 구축하는 방안은 크게 2가지로 단계로 나뉜다. 첫 번째 단계는 도메인 감성사전을 자동 생성하기 위한 기초 자료로 활용하기 위한 한국어 범용 감성사전을 구축하는 것이다. 본 연구에서는 국립국어원 표준 국어 대사전의 뜻풀이를 활용하여 딥러닝 기법 중 하나인 Bi-LSTM 모델 통해 감성을 분류하였다. 감성이 분류된 뜻풀이는 해당 뜻풀이가 나타내는 감성에 따라 인간의 보편적인 감정을 나타내는 도메인에 독립적인 감성 어휘 추출된다. 이와 같은 감성 어휘들은 총 14,843개이며 1-gram 이외에도 n-gram의 다양한 형태로 추출되었다. 그리고 해당 한국어 감성사전을 'KNU 한국어 감성사전'이라 명명하였다. 두 번째 단계는 구축된 'KNU 한국어 감성사전'을 활용하여 도메인 감성사전을 자동으로 생성하는 단계이다. 도메인 감성사전을 자동으로 생성하기 위해 'KNU 한국어 감성사전'을 활용하였으며, 분석하고자 하는 도메인을 수집하였다. 수집된 도메인을 통해 감성을 갖는 품사에 대한 후보 감성 어취들을 추출하였고, 어휘 간 유사도 관계와 접속 관계를 통해 자동으로 극성을 부여하였다. 또한 의존 구문 분석을 통해 어절 단위(n-gram)의 감성 어휘를 추출하고 이에 대한 극성을 자동으로 부여하였다. 본 연구의 제안 방안은 특정 도메인에서 독립적으로 사용되는 감성 어휘들을 포함하여 감성사전을 구축하기 때문에 특정 다양한 도메인 분석가들이 기존의 감성사전에 특정 도메인의 감성 어휘들을 추가할 필요 없이 바로 사용할 수 있다. 또한 신조어나 새로 생긴 은어들을 알고리즘에 의해 빠르게 식별할 수 있다는 장점이 있다. 예를 들면, 새로운 신조어나 은어들이 생기면 분석가들은 해당 어휘들을 일일이 찾고 토론을 거쳐 극성을 부여해야 하지만 본 제안 방안은 어휘들의 동시 출현 확률과 접속 관계에 의해 자동으로 빠르게 식별 가능하며 극성을 자동으로 부여해주기 때문에 빠르고 효율적이다. 본 논문에서 제안한 공헌은 다음과 같다. 첫 번째, 표준 국어 대사전에 수록된 뜻풀이들을 활용하여 도메인에 독립적인 감성 어휘들을 추출하고 이를 통해 한국어 범용 감성사전을 구축한 최초 연구이다. 두 번째, 구축된 'KNU 한국어 감성사전'은 누구든 쉽게 접근하고 제한 없이 이용할 수 있도록 공개되어 있으며 특정 도메인 감성사전을 빠르게 구축하기 위한 유용한 기초 자료로 활용될 수 있다. 세 번째, 'KNU 한국어 감성사전'은 도메인에 독립적인 감성 어휘로 구성되어 있기 때문에 도메인 감성사전 구축을 위한 기초 자료 이외에도 기본적인 감성분석에 활용될 수 있다. 네 번째, 어휘들의 동시 출현 빈도에 기반 한 어휘 간 유사도 정보, 어휘 간 접속 관계 정보를 활용하여 도메인 감성사전을 자동 구축한 연구이다. 다섯 번째, 1-gram 감성 어휘뿐만 아니라 의존 구문 분석과 어구 분석을 통해 n-gram 감성 어휘를 자동으로 추출, 극성을 자동으로 부여하였다. 여섯 번째, 제안 방안은 특정 도메인에 알맞은 감성사전을 자동으로 구축해주기 때문에 분석하고자 하는 코퍼스의 도메인에 제약받지 않고 감성사전 기반의 감성 분석 수행이나 대량의 학습데이터 구축 그리고 딥러닝 입력의 자질로 활용될 수 있다.

향후 연구로는 일부 도메인을 선정하여 도메인 감성

사전 자동 구축을 수행해본 후 분석을 수행할 예정이다. 또한 기존 시드 감성 어휘의 극성 변화를 탐지할 수 있는 알고리즘을 추가하여 시드 어휘의 극성이 역전될 수 있도록 할 예정이다. 딥러닝의 어텐션 메커니즘(Attention Mechanism)을 활용하여 문장에서 주의 되는 부분과 도메인 감성 어휘 간의 관계를 이용하여 감성 어휘를 딥러닝 기법에 의해 자동으로 식별하는 알고리즘을 수행할 예정이다. 또한 4.2장에서 언급한 감성사전 구축간에 생기는 제약을 보다 효과적으로 처리하기 위해, 띄어쓰기 모듈, 오탈자 정제 모듈을 적용할 예정이며, 시드 어휘의 극성을 고정시키지 않고 필요에 따라 시드 어휘의 극성 또한 역전시킬 예정이다. 또한 반어법 문맥을 보다 정교하기 탐지하기 위해 규칙을 더 추가할 예정이며, 앞서 언급한 극성이 역전된 시드 어휘를 활용할 예정이다. 마지막으로 대용량 코퍼스에 대한 감성사전을 빠르게 구축하기 위해 맵리듀스(MapReduce) 기법을 적용할 예정이다.

## References


[1] CIO, Expansion of BI and Analytical Role, 2013, Available at http://www.ciokorea.com/t/544/9118/15551 (Accessed 2018)

[2] Wikipedia, Sentiment Analysis, Available at https://en.wikipedia.org/wiki/Sentiment_analysis (Accessed 2018)

[3] Pennington, J., R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Represent," *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[4] AutoView, Available www.autoview.co.kr (accessed 2019)

[5] Wikipedia, Pointwise Mutual Information, Available https://en.wikipedia.org/wiki/Pointwise_mutual_information (accessed 2019)

[6] Wikipedia, Jaccard Similarity, Available https://en.wikipedia.org/wiki/Jaccard_index (accessed 2019)

[7] Park, S. M, C. W. Na, M. S. Choi, D. H. Lee and B. W. On, "KNU Korean Sentiment Lexicon : Bi-LSTM based Method for Building a Korean Sentiment Lexicon," *Journal of Intelligence and Information Systems*, Vol. 24, No. 4, pp. 219-240, 2018.

[8] Shin, D. H., D. Cho, and J. S. Nam, "Building the Korean Sentiment Lexicon DecoSelex for Sentiment Analysis," *Journal of Korealex*, No. 28, pp. 75-111, 2016.

[9] OpenHagul, Sentiment Lexicon, Available at openhangul.com/restrict (Accessed 2018)

[10] 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, pp. 49-67, 2015.

[11] 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, pp. 93-121, 2005.

[12] Baccianella, S., A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," *Proc. of the International Conference on Language Resources and Evaluation*, LREC, pp. 2200-2204, 2010.

[13] Sheng, H., N. Zhendong and S. Chongyang, Automatic construction of domain-specific sentiment lexicon based on constrained label propagation, *Knowledge-Based Systems*, Vol. 56, pp. 191-200, 2014.

[14] Yang, M., D. Zhu and K. P. Chow, "A Topic Model for Building Fine-grained Domain-specific Emotion Lexicon," *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 421-426, 2014.

[15] Tai, Y. J. and H. Y. Kao, "Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation," *Conference: Proceedings of International Conference on Information Integration and Web-based Applications & Services*, 2013.

[16] Kim, J. H., Y, J. Oh and S. H. Chae, "The Construction of a Domain-Specific Sentiment Dictionary Using Graph-based Semi-supervised Learning Method," *Science of Emotion & Sensibility*, Vol. 18, No. 1, pp. 97-104, 2015.

[17] ang, H. S., K. Y. Jeong, and E. Y. Jang, "Efficient method to generate sentiment vocabulary for specific topic based on Word2Vec," *Proc. of Korean Institute of Information Scientists and Engineers*, pp. 652-654, 2017.

[18] Kim, S. B., S. J. Kwon, and J. T. Kim, "Building Sentiment Dictionary and Polarity Classification of Blog Review By Using Elastic Net," *Proc. of Korean Institute of Information Scientists and Engineers*, pp. 639-641, 2015.

[19] Choi, S. J. and O. B. Kwon, "The Study of Developing Korean SentiWordNet for Big Data Analytics - Focusing on Anger Emotion -," *The Jounal of Society for e-Business Studies*, Vol. 19, No. 4, pp. 1-19, 2014.

[20] 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, pp. 83-105, 2016.

[21] Jung, W. Y., B. C. Bae, S. H. Cho and S. J. Kang, "Construction and Evaluation of a Sentiment Dictionary Using a Web Corpus Collected from Game Domain," *Science of Emotion & Sensibility*, Vol. 18, No. 5, pp. 97-104, 2018.

[22] Lee, S. H., J. Choi, and J. W. Kim, "Sentiment Analysis on Movie Review Through Building Modified Sentiment Dictionary by Movie Genre,"

*Journal of Intelligence and Information Systems*, Vol. 22, No. 2, pp. 97-113, 2016.

박 상 민

2020년 군산대학교 소프트웨어융합공학과의 소프트웨어융합공학 석사. 현재 솔트룩스 AI Labs 재직 중. 관심분야는 자연어 처리, 텍스트 마이닝, 인공지능, 감성사전, 감성분석, 감정분석

온 병 원

2007년 미국 펜실베이니아주립대학교의 컴퓨터공학과 박사. 캐나다 브리티시컬럼비아 대학교 박사후연구원. 2010년 미국 일리노이대학교 ADSC센터 선임연구원 서울대학교 차세대융합기술연구원 연구교수. 현재 군산대학교 소프트웨어융합공학과 부교수. 관심분야는 데이터 마이닝, 정보검색, 빅데이터, 인공지능