The Journal of Studies in Language

The Journal of Studies in Language 36(4), 525-552 (2021)

DECO 언어자원 · 의존 파서와 LSTM에 기반한 하이브리드 자질기반 감성분석 플랫폼 연구

황창회* · 유광훈 · 남지순*****

한국외국어대학교

A Hybrid Feature-based Sentiment Analysis Platform Based on DECO-Dependency Parser and LSTM

Hwang, Chang-Hoe*, Yoo, Gwang-Hoon and Nam, Jee-Sun*****

Hankuk University of Foreign Studies

***First Author / **Second Author / ***Corresponding Author**

OPEN ACCESS

https://doi.org/10.18627/jslg.36.4.202102.525

pISSN : 1225-4770
eISSN : 2671-6151

Received: January 10, 2021

Revised: January 29, 2021

Accepted: February 08, 2021

This is an Open-Access article distributed under the terms of the Creative Commons Attribution NonCommercial License which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright©2021 the Modern Linguistic Society of Korea

본인이 투고한 논문은 다른 학술지에 게재된 적이 없으며 타인의 논문을 표절하지 않았음을 서약합니다. 추후 중복게재 혹은 표절된 것으로 밝혀질 시에는 논문게재 취소와 일정 기간 논문 제출의 제한 조치를 받게 됨을 인지하고 있습니다.

ABSTRACT

The Journal of Studies in Language 36.4, 525-552. This study introduces the implementation of the DecoFESA platform, which proposes a hybrid approach to FbSA (Feature-based Sentiment Analysis) using the DECO-LGGs (Local Grammar Graphs based on DECO dictionary) assigned with dependency parsing and LSTM (Long Short-Term Memory). DecoFESA is developed to leverage DECO-LGG linguistic resources and NLP_HUB dependency parser to process semantic information and analyze sentence structures for Target, Feature, and Sentiment detection at sentence level. It also uses of one of the efficient deep learning algorithms, LSTM, to classify sentiments of OOV (Out of Vocabulary) texts. For the performance evaluation, we scraped 1,500,000 cosmetics reviews of online shopping malls by a crawler. DecoFESA shows the overall robust performance: 0.95 and 0.80 f-measure for POSITIVE and NEGATIVE with 0.845 and 0.83 accuracies for Target and Feature extraction, which manifests the efficiency and practicality of the hybrid approach to FbSA utilizing extensive DECO-LGG linguistic resources collaborated with dependency parsing and LSTM. **(Hankuk University of Foreign Studies)**

Keywords: Feature-based Sentiment Analysis, DECO dictionary, Local Grammar Graph, Dependency Parser, Long Short-Term Memory

1. 서 론

본 연구는 도메인 유형별 감성 극성 정보를 사전정보와 의존 파서(parser)를 통해 처리하고 이를 기계학습 방법을 통해 보완하는 하이브리드 자질기반 감성분석

- 본 연구는 2020년도 한국외국어대학교 교내학술연구비 지원에 의해 수행되었음. 연구의 초기 버전은 HCLT-2020 학술대회에서 구두 발표된 바 있으며, 토론자 및 익명의 심사자 선생님들의 지적과 도움말에 감사의 말씀을 전한다.

(Feature-based Sentiment Analysis) 방법론을 제시하고, 이를 실제로 처리하기 위해 구현된 DecoFESA 플랫폼을 소개하는 것을 목표로 수행되었다.

감성분석(sentiment analysis)은 다양한 사용자 생성문(user-generated sentence)에 포함되어 있는 주관적 오피니언(opinion)을 수집 및 분석하기 위한 오피니언 마이닝(opinion mining) 분야로, 개별 문장의 전체 오피니언을 감성표현의 극성(polarity) 방향성에 따라 분류하는 방법이 주로 사용되어온 분야이다. 하지만 본 연구의 주요 목표에 해당하는 자질기반 감성분석은 단순 의미적 지향성(sentiment orientation)을 바탕으로 오피니언 전체를 분석하는 방법론에서 벗어나, 감성분석 핵심 어휘인 감성표현과 해당 감성표현이 평가하고자 하는 '평가대상'의 구체적 세부 정보에 대한 처리를 목표로 한다. 이때 분석의 대상이 되는 요소들은 일반적으로 Liu(2012; 2015)의 오피니언 5원소쌍으로 정의된다.

(1) 오피니언 5원소쌍(opinion quintuple)
{개체명, 자질명, 감성표현, 오피니언 평가자, 시간}

(1)에서 제시된 표현 중, 감성표현(sentiment) 부류와 감성분석의 평가대상(target)에 해당하는 개체명(named-entity) 및 자질명(feature)은 오피니언 구성의 핵심 요소로 분류된다. 따라서 코퍼스에서 나타나는 구체적인 평가대상인 개체명/자질명과 감성표현을 통합적으로 분석하는 자질기반 감성분석에는 문장 내 오피니언 핵심 원소들에 대한 의미 정보 및 구문적 관계에 대한 처리가 요구된다. 이러한 방법론은 평가문 내에서 감성 극성이 표시되는 모든 자질명에 대한 극성 분석을 수행하는 자질-어휘 감성분석이 아니라, 한 문장 내의 감성표현에 해당하는 주요 서술어를 감지하고, 평가대상으로서의 자질명을 범주화하여 감성분석 결과를 도출하는 자질-카테고리 감성분석의 관점(Xue and Li, 2018)과 상응된다.

그런데 사용자생성문 텍스트에 나타나는 감성분석 원소들은 가장 기초적인 형태적 실현인 단일어 외에도 다양한 다단어 표현(multi-word expression) 형태로 실현되는 경우가 빈번하다. 가령 특정 텍스트 유형에서는 긍정 극성의 차원에서 단일어휘 '좋다'와 같은 표현보다 '마음에 든다'와 같은 긍정 극성의 다단어 표현의 출현이 훨씬 빈번한데, 다단어 표현의 경우 일반적인 사전 표제어로 등재하거나 의미적 특징에 의한 자동 조합 또는 유추하기가 어려운 어휘적 특이성(lexical idiosyncrasy)을 보이기 때문에 이들에 대한 별도의 언어자원이 제공되지 않으면 감성분석의 성능을 저해하는 중요한 장애물이 된다(Piao et al., 2003).

사용자생성문 텍스트에서 나타나는 개체명과 자질명 또한 다음과 같이 도메인별로 다단어 표현의 양상으로 나타나기 때문에 이에 대한 별도의 처리가 요구된다. 다음을 보자.

(2) ㄱ. 아이폰 Xs는 화면이 딱 적당해요.(+)
ㄴ. 저는 뿌링클 시즈닝이 입에 안맞더라구요.(-)
ㄷ. 발색이 금방 날아가는 보브 실키핏 립스틱.(-)

위 예시의 문장들은 각각 스마트폰/배달음식/화장품과 관련된 사용자의 의견을 담고 있으며, 서로 다른 도메인에서 실현되는 만큼 자질기반 감성분석을 수행하기 위한 어휘 및 어구의 양상이 다르게 실현된다. 이를 자세히 살펴보면, 스마트폰 리뷰인 (2ㄱ)에서는 스마트폰 제품명에 해당하는 '아이폰 Xs'가 분석 대상인 개체명으로 나타나고, (2ㄴ)은 '뿌링클'이, (2ㄷ)에서는 '보브 실키핏 립스틱'이 2개 이상의 어휘가 결합한 다단어 표현으로 나타나고 있다. 대상의 특성에 해당

하는 자질명 또한 (2ㄱ)에서는 '화면'이, (2ㄴ)과 (2ㄷ)에서는 '시즈닝'과 '발색'이 각각 자질명으로 나타나고 있다. 이러한 개체명/자질명 실현 양상이 의미하는 것은 해당 표현들이 코퍼스의 도메인에 따라 완전히 다른 양상을 띨 수 있는 것이다. 특히 개체명의 구 단위 실현 양상의 경우, (2ㄷ)의 개체명은 어절 구분자가 생략된 하나의 토큰이나, 구성성분이 누락되거나 변형되어 가령 '실키핏 립스틱'과 같이 축약된 형태로 실현되는 것이 가능하다.

이러한 문제점의 해결책을 제시하기 위해 현재까지 많은 논의가 지속되어 왔지만, 주로 규칙적으로 등장하는 패턴을 인식 및 추론하는 데에 적합한 순환 신경망(Recurrent Neural Network: RNN) 계열 알고리즘에 조건부 무작위 장(Conditional Random Fields: CRF)과 같은 여타 알고리즘이 목적에 따라 혼합되는 형태로 활용되는 연구들이 주를 이루었다. 하지만 이러한 연구들의 성과는 학습 데이터에 대한 과도한 의존성 문제가 해결되지 않는 한, 지속 가능한 자질기반 감성분석 접근법으로 기대하는 데에는 한계가 있다.

본 연구는 개체명 결합성이 가장 복잡하게 나타나는 코퍼스 도메인 중 하나인 화장품 도메인을 대상으로 하였다. 화장품 도메인은 온라인 상품 구매가 활발히 이루어지는 상품 분야 중 하나로 텍스트 수집이 용이하다는 장점이 있으며, 동일한 맥락에서 화장품 구매자의 평가문에 대한 감성분석 시스템이 구축될 필요가 존재한다. 본고에서는 화장품 도메인의 감성분석을 위해 개체명 및 자질명 구성과 관련된 결합 정보를 분석하여 얻어진 패턴을 방향성 비순환 그래프(directed acyclic graph)의 형태로 기술하는 부분 문법 그래프(Local Grammar Graph: LGG) 프레임을 사용하였다. 감성표현의 경우는 범용 극성 표현뿐만 아니라 도메인에 특화된 극성 표현 유형을 의미처리하기 위한 패턴 문법을 결합 제약을 고려하여 구성하여, 화장품 코퍼스 내 의미 정보를 효과적으로 처리하기 위한 언어자원 DECO-LGG를 구축하였다. DECO-LGG가 적용되어 정보가 처리된 문장들은 의존 파싱(parsing) 과정 및 의미 정보 통합 과정을 거쳐 자질기반 감성분석의 출력값에 해당하는 {개체명, 자질명, 감성 극성}의 형태로 저장되며, 이는 추후 시각화 과정을 통해 간략화된다. 이때 해당 과정에서 미등록단어(Out Of Vocabulary: OOV)에 의해 출력값을 갖지 못한 문장들의 경우 장단기기억(Long Short-Term Memory: LSTM) 딥러닝 모델에 기반하여 극성 정보를 할당받으며, 이와 연결되는 평가대상 개체명과 자질명은 한국어의 '강력한 후행 언어(rigid head-final language)' 속성을 고려하여 후치된 요소부터 자동 선택되도록 설정하였다. 본 연구에서는 이상의 처리 과정을 단순 코드 형태가 아닌 스탠드얼론(stand-alone)으로 작동하는 'DecoFESA' 플랫폼으로 구현하여 사용자가 직접 자신이 구축한 감성분석 언어자원을 활용하여 자질기반 감성분석 결과를 저장하고 시각화 결과를 참고할 수 있도록 하였다.

도메인 코퍼스에 대한 자질기반 감성분석 관련 핵심 어휘들의 의미 정보를 처리하는 기능을 수행하는 DECO-LGG 감성분석 언어자원은, DECO (Dictionnaire Electronique du COreen) 한국어 전자사전(남지순, 2018)과 LGG 패턴문법 기술(Gross, 1997; 1999) 프레임을 통해 구현되고, Unitex (Paumier, 2003) 플랫폼을 통해 적용되어 특정 도메인 코퍼스 내부에서 실현되는 감성분석 원소들의 의미처리를 수행하는 데에 사용된다. 기계가독형 전자사전과 연동되어, 어절 시퀀스 정보를 활용하는 DECO-LGG 언어자원은 실제 언어 기술의 양상과 밀접하게 맞닿아 있음에 따라, 확률 기반 처리방식에 의존하는 기존 딥러닝 기반 감성분석 방법의 한계를 넘어 보다 실질적이고 정확한 의미 처리가 가능하다는 장점이 있다. 구문 패턴에 기반하여 구축된 언어자원 DECO-LGG에 의해 반자동으로 처리된 도메인 텍스트의 의미 정보는 문장 단위의 구문 분석을 통해 획득한 논항 정보와 통합적으로 분석되어 문장의 평가대상 개체명/자질명/감성표현을 페어링(pairing)하는 언어 정보 기반의 자질기반 감성분석을 수행한다. DecoFESA 플랫폼에서는 OOV 출현 때문에 감성분석 결과가 도출되지 못한 문장들을 LSTM 모델을 통해 보완하여, 전체 과정을 자질기반 감성분석을 위한 하이브리드 방법론으로 처리하게 된다.

이후 부트스트래핑(boot-strapping) 방식을 통해 반복 확장이 가능한 언어자원을 기반으로 코퍼스의 의미처리를 수행하고, 처리된 코퍼스를 실제 알고리즘을 통해 분석한 후, 분석 결과를 출력값으로 제시하는 일련의 과정은 이에 대한 시각화 요약 정보의 제공과 함께 DecoFESA 플랫폼에서 모두 진행된다. 이상의 처리 과정은 다음과 같이 도식화된다.

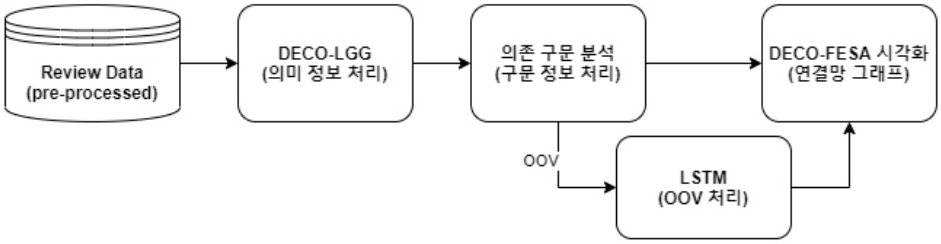

그림 1. DecoFESA 시스템 프로세스

다음 2장에서는 자질기반 감성분석을 수행하기 위해 기존에 진행된 오피니언 원소의 추출 및 처리 관련 선행연구를 검토하고, 실제 자질기반 감성분석을 수행한 연구와 한국어 SNS 텍스트의 자질기반 감성분석 연구를 소개한다. 3장에서는 화장품 코퍼스의 자질기반 감성분석을 위한 의미 정보 자원인 DECO-LGG의 구성 과정과 패턴 문법의 양상에 대해 살펴본다. 4장에서는 언어자원이 적용된 코퍼스의 의존 구문 분석에 관련된 논의를 통해 의미 정보가 문장 내에서 어떠한 방식으로 페어링 되는지에 대해 제시하고, 이어지는 5장에서는 4장의 처리를 통해 분석되지 못한 문장들의 극성을 분류하는 LSTM 분류기의 이론적 구성과 극성 예측(prediction)에 관련된 사항을 서술한다. 6장에서는 본 연구의 방법론을 검증하기 위해 200문장으로 구성된 평가용 정답지를 사용한 실험 결과를 제시한다. 마지막 장에서는 평가 결과에 대한 정리와 더불어, 해당 방법론을 통해 구성된 DecoFESA 플랫폼과 감성분석 결과의 동적 시각화 결과를 소개하고, 향후 연구 방향에 대해 논의한다.

2. 관련 연구

자질기반 감성분석을 수행하기 위한 국내외의 연구들은 일반적으로 문서 내에서 나타나는 자질의 추출 방법론 및 긍·부정 의미 분류를 중심으로 진행되었다. 이에 먼저 자질 추출과 관련된 내용을 살펴보면 이는 크게 빈도기반 방식, 구문 처리 방식, 기계학습 방식의 세 가지로 분류된다.

자질 처리의 빈도기반 방식의 경우, 분석 대상 문서 내에서 고빈도로 나타나는 일반 명사들을 통해 자질 추출 및 처리를 수행한다. 오피니언 문서에서 나타나는 특정한 단일어 명사 및 복합 명사 표현들의 실현빈도가 나머지 어휘의 실현빈도보다 훨씬 더 높으며, 이러한 표현들이 자질 명사일 가능성이 큰 것으로 관찰되었기 때문이다. Hu와 Liu(2004)의 연구에서 기술된 자질 검출 방법은 먼저 문장을 형태소 단위로 분석한 후, 단일 명사와 복합 명사만을 추려내어 자질어휘의 후보군을 구성하였으며, 문장 내에서 공기하는 단어의 빈도가 전체 문장의 최소 1%에서 발생해야 한다는 역치를 상정함에 따라, 해당 역치를 넘는 명사의 조합 빈도를 유의미한 명사 조합으로 특정한다. 이 방식은 데이터 마이닝에서 고전적으로 사용하는 선험적(apriori) 알고리즘으로, 이때 자질 명사는 같은 문장에서 나타나면 순서나 시퀀스를 고려하지 않는다. 이외에도, 해당 연구에는 하나 이상의 감성 단어와 공기하는 최단 거리의 명사 또는 복합 명사를 추출하는 방법이 사

용되었으나, 문장 내 구문 의존성이 고려되지 않았으며, 빈도 및 단어 간 거리를 기반으로 처리하는 방식이므로, 언어학적 특징을 고려하지 못하는 단점이 존재하였다. 또 다른 빈도기반 연구로, Hai et al.(2011)은 어휘공기 행렬을 통해 명시적 자질명 및 이와 공기하는 감성 어휘의 연관 규칙을 활용하여 비명시적으로 나타나는 자질명을 추출하는 방법을 제안하기도 하였다. 이러한 빈도기반 방법의 주요 단점은 자연적으로 빈도가 높은 명사와 명사구가 자질명으로서 오분석될 수 있다는 점이다. 또한, 상대적으로 언급량이 적을 수 있는 구체적인 자질명 부류는 이와 같은 빈도기반 방법을 통해 처리하기 어렵다는 한계가 발생한다.

구문 기반 방법의 경우, 자질 표현 추출을 위해 명사의 빈도에 초점을 맞추는 대신 구문 관계를 활용하여 자질어휘를 추출한다. 이는 감성 어휘가 가지고 있는 의미적 특성에 따라 자질명이 구문적으로 공기할 가능성이 크기 때문이다. 구문 정보 기반 접근법의 장점은 저빈도라도 규칙에 입각하면 자질을 찾을 수 있다는 점이다. Zhao et al.(2010)에서는 주석 코퍼스를 기반으로 추출된 자질 처리 구문 패턴과 입력 구문의 유사성 점수를 계산하여 구문 기반 자질 추출의 범용성을 높이는 방법을 제안하였다. 하지만 구문 기반 방법은 범용성을 획득하기 위해 많은 구문 관계를 직접 기술해야 한다는 단점이 있어, 비교적 높은 정확도의 장점과 동시에 낮은 재현율의 문제점을 가진다. Mirtalaie et al.(2018)에서는 구문 기반 방법과 사전을 활용하여 자질 표현을 추출하였는데, 가령 “The button is inconvenient”와 같은 의존 구문 정보를 “Feature → Dep1 → VB - B VB ∈ {is, are} Dep1 ∈ {nsubj}”와 같이 일반화하고, 이를 특정 상품에 대한 상의어, 하의어 정보를 가진 Product tree와 같은 사전을 활용하여 노이즈를 제거하는 방식으로 자질 추출 과정을 진행하였다.

지도적/비지도적 기계학습 방법론 또한 자질기반 감성분석을 위한 자질 추출에 활용되는데, Samha et al.(2015)는 시퀀스 정보별 레이블(label)에 기반한 기계학습 방법인 조건부 무작위장(Conditional Random Field: CRF)을 활용하여 자질 표현을 추출하는 방법을 제안한 바 있다. CRF의 경우 토큰마다 라벨링을 하는 형태소 태깅과 같은 작업에 주로 사용되는 방법으로, 해당 연구에서는 자질을 추출하기 위해 전처리 및 품사 태깅이 완료된 텍스트에 Beginning-Middle-Ending의 형태로 주석하는 방식으로 학습 데이터를 구성하여 CRF 모델에 학습시켰다. Alghunaim et al.(2015)은 오피니언 리뷰에서 나타나는 감성 정보와 자질정보를 서포터 벡터 머신(Support Vector Machine: SVM)을 통해 학습시켜 자질 표현을 추출하는 벡터 기반 지도적 알고리즘 제안하였다. SVM은 지도적 기계학습 방법 중에서도 일반화 능력이 뛰어난 방법으로, 다양한 커널 함수를 사용하여 입력 공간에서 분류되기 힘든 비선형 문제를 고차원 공간으로 매핑(mapping)시켜 최대의 마진(margin)을 가진 비선형 데이터의 선형 판별을 수행한다. 비지도적 방법을 사용해 자질명을 추출한 연구에는 Huang et al.(2012)이 있다. 해당 연구에서는 레스토랑에 대한 사람들의 의견을 체계적으로 표현한 대규모 사용자 리뷰에서 기능-의견 쌍(pair)을 추출하기 위한 “RevMiner”를 제안하였다. 자동 부트스트래핑 방식이 적용된 기능-의견 쌍 추출기(attribute-value extractor)를 기반으로 유사도에 대한 스코어링과 극성계산을 수행한다. 자질 추출을 위해서는 이미 설정해 놓은 5개의 클러스터를 정의하고, 관련 자질어휘를 그룹화하여 인터페이스상에 시각화하는 방식을 활용하였다.

앞선 연구들과 같은 다양한 자질 추출 방법론에 기반하여 수행되는 자질기반 감성분석 연구에도 기계학습을 통한 방법론이 사용되는데, 먼저 지도적 학습과 관련된 연구로 He et al.(2018)이 있다. 해당 연구는 음식점 관련 사용자 리뷰를 학습 데이터로 사용하여, 문장 단위 구문 분석을 수행함에 따라 얻어진 의존 관계를 토대로 Attention 가중치를 할당하는 LSTM 모델을 제안하였다. 여기서는 동일 도메인에 대한 자질기반 감성분석에서 84.61%의 정확도를 보였다. 비지도적 기계학습에 가까운 자질기반 감성분석 연구로는 Garcia-Pablos et al.(2018)이 있는데, 여기서는 연속단어 임베딩(continuous word embedding)과 최대 엔트로피 분류기를 결합한 토픽 모델링 접근법을 통해 자질기반 감성분석을 수행

한다. 해당 연구에서 제시된 모델은 사용자가 입력한 최소한의 토픽 시드 어휘들을 기반으로 도메인 자질 분류/자질-감성 어휘 분리/극성 분류를 동시에 수행한다.

한국어에 대한 자질기반 감성분석의 연구 성과는 현재로는 찾아보기 쉽지 않다. Song et al.(2019)의 연구에서는 어텐션(attention) 기반 LSTM을 사용하여 주석된 한국어 화장품리뷰 코퍼스의 자질 및 극성 어휘를 형태소 레벨로 학습시켜 화장품 코퍼스에서 나타나는 자질-극성 어휘를 페어링하여 자질기반 감성분석을 수행하였다.

앞서 언급한 자질기반 감성분석 연구들은 공통적으로 오피니언 대상 어휘들을 고정해 놓거나, 평가대상을 자질명으로 설정하여 진행되었다. 이에 따라 텍스트에서 나타나는 대상 어휘를 감지해내지 못하며, 오로지 수기로 주석된 코퍼스 데이터 내에 한정된 표현들에 대한 분석만을 학습에 포함한다는 한계가 존재하였다. 이러한 한계를 극복하기 위하여 본 연구에서는 사전 및 구문 패턴을 기반으로 특정 도메인 코퍼스에서 나타나는 감성분석 관련 요소들을 감지하고 주석하는 반자동 시퀀스를 적용하여 텍스트가 평가하고자 하는 대상에 대한 추출이 가능하도록 구성하였다. 또한, 자질명의 경우, 범용 사전을 통해 분석된 결과를 토대로, 빈도에 따라 범주를 구성하여 자질명으로 분석될 수 있도록 도메인별 자질 사전을 구축하였다. 이를 바탕으로 텍스트에서 나타나는 감성분석 핵심성분들의 통사적 관계를 분석할 수 있도록 의존 파서에 의한 구문 분석을 시도하였다. 이때 이러한 처리를 진행하면서 나타나는 OOV에 대해서는 LSTM을 활용한 딥러닝 모델을 통해 극성을 최종 평가할 수 있도록 구성하였다.

3. 자질기반 감성분석과 DECO-LGG 언어자원

3.1 자질기반 감성분석에 필요한 언어자원 유형

자질기반 감성분석을 수행하기 위해선 문장 내에 존재하는 오피니언 구성의 핵심 요소를 처리하는 것이 중요하다. Liu(2015)의 오피니언 5원소쌍 중 평가대상 개체명/자질명/감성표현이 이에 해당한다. 다음을 보자.

(3) ㄱ. 시슬리 좋아요.
ㄴ. 세타필이 바세린보다는 보습력이 떨어져요.
ㄷ. 비디비치 썬크림 체험단 이벤트로 받은건데 끈적이지 않고 촉촉해요

(3)에 제시된 예문들은 실제 온라인 쇼핑몰에 등록되어 있는 화장품 리뷰들을 보인다. 여기서 개체명으로 분류되어야 하는 어휘로 (3ㄱ)에는 '시슬리', (3ㄴ)에는 '세타필'과 '바세린'이 나타나고 있으며 (3ㄷ)에는 복합어인 '비디비치 썬크림'으로 실현되었다. 이러한 개체명 어휘를 처리하기 위해서는 해당 어휘들이 개체명에 해당됨을 알 수 있는 기계가독형 사전에 기반한 단일어 처리 과정이 먼저 요구되며, (3ㄷ)에서와 같이 개체명의 실현이 다단어 표현으로 나타나는 경우, 해당 개체명 부류들을 인식하기 위한 패턴 문법이 요구된다. 자질명의 실현 유형은 몇 가지로 나누어지는데, (3ㄱ)과 같이 평가대상에 대한 자질이 나타나지 않은 경우와, (3ㄴ)의 '보습력'과 같이 명시적으로 드러난 경우, (3ㄷ)과 같이 명시적으로 존재하지는 않지만, '촉촉해요'와 같은 감성표현에 내재되어 있는 경우가 있다. 이에 따라 오피니언 구성 요소 중 자질명의 실현은 수의적일 수 있으며, 명시적/암시적 자질명을 처리하기 위해서는 이와 관련된 의미 정보의 처리가 사전에 기반하여 처리되어야 한다. 또한 감성표현은 (3ㄱ, ㄷ)의 '좋아요'와 '촉촉해요'와 같이 단일어휘 형태에서 감성 극성

이 나타나는 표현들이 있는 경우와, '보습력이 떨어져요'와 같이 부분적 시퀀스를 통해 극성이 드러나는 다단어 표현이 존재하기 때문에 이를 효과적으로 처리하기 위한 패턴 기술이 필요하다.

본 연구에서는 이러한 단일어 및 다단어 표현을 효과적으로 표상하고 이를 텍스트 분석에 적용할 수 있도록, DECO 전자사전에 기반하여 유한상태 트랜스듀서(Finite State Transducer: FST) 형식으로 구현되는 LGG 프레임을 사용하였다. LGG는 Unitex 플랫폼에서 방향성 그래프 형식으로 기술된 후 자동으로 FST로 컴파일되어 텍스트 분석에 적용 가능한 문법이 된다. 이는 순환 전이망(recursive transition network) 문법의 하나로, 한국어의 어휘, 통사, 의미 정보 및 극성 정보를 처리하는 기계가독형 활용형사전 DECO를 기반으로 작동한다. 구축된 LGG 패턴 문법이 FST 문법으로 컴파일되면, 코퍼스에 적용되어 문장의 의미처리를 수행한다.

하지만 이러한 의미처리에 앞서, 특정 상품 및 서비스에 대한 사용자 평가문은 소셜 미디어 텍스트와 같은 비정형 텍스트 형태로 실현되기 때문에 다음과 같은 문제점이 잠재되어 있다.

(4) 헤라쿠션 너무 좋은거같아요

위 예문의 텍스트는 실제 사용자 평가문에 대한 크롤링 단계에서 나타난 비정형문 중 하나이다. 평가대상에 해당하는 개체명과 이를 평가하기 위한 극성 어휘가 매우 단순한 형태로 기술되어 있으나, 의존 명사에 대한 띄어쓰기가 지켜지고 있지 않아 텍스트 분석 상에서 OOV로 분석될 확률이 매우 높다. 이처럼 온라인 코퍼스의 특성상 어문 규정을 준수하지 않아 OOV로 분석되는 어절들을 최소화 하기 위해 본 연구에서는 패턴 사전 구축을 위한 DECO 사전 적용의 마지막 단계에서 OOV로 나타나는 어절들을 전처리하기 위한 LGG 전처리 모듈(최성용 · 남지순, 2018)을 모든 텍스트의 처리 과정에서 적용하였다.

이상과 같이 전처리가 적용된 코퍼스 내 문장의 의미처리를 수행하기 위해 개체명/자질명/감성표현을 인식하기 위한 언어자원이 적용되면, 주어진 문장의 최종 감성, 또는 오피니언이 무엇인지를 결정하는 과정이 필요하다. 다음을 보자.

이번에 새로 나온 <ENTITY>노르딕 바디 모이스처</ENTITY>는 평상시에 <ENTITY>뉴트로지나</ENTITY>의 <FEATURE>보습감</FEATURE>은 <POSITIVE>좋아했지만</POSITIVE> 사용해본 결과 <FEATURE>용량</FEATURE>이 <NEGATIVE>너무 적었어요</NEGATIVE>.

위 문장에는 본 연구에서 사용한 언어자원이 감지할 수 있는 개체명/자질명/감성표현이 모두 인식되어 마크업 되어있다. 그런데 이처럼 여러 개의 개체명과 자질명, 감성표현이 실현되는 문장의 경우, 실제 문장이 표상하고자 하는 오피니언이 '[개체명]노르딕 바디 모이스처-[자질명]용량-[감성]NEGATIVE'임을 선택해주기 위해서는 문장 내에 존재하는 핵심 술어를 찾아내야 하며, 해당 술어에 의존 관계를 가지는 개체명/자질명이 오피니언의 대상에 해당된다는 것이 분석되어야 한다. 이러한 과정을 위해 본 연구에서는 문장의 구문 정보를 분석하기 위한 도구로써 의존 구문 분석기를 활용한다. 위의 과정에서 LGG 언어자원에 의해 처리된 정보들은 의존 파서의 분석 과정에 활용되어, 문장 안에서 나타나는 대상 개체명/자질명/감성표현을 연결해주는 정보로 활용된다. 본 연구에서 구축하는 언어자원은 코퍼스 내에서 나타나는 오피니언 구성 요소들의 처리 패턴에 대한 분석을 통해 부트스트랩 방식으로 확장될 수 있고, 단일어/다단어 표현에 대한 활용형을 경제적으로 처리하는 것이 가능하다는 점에서 차별성을 가진다.

3.2 자원 구축에 활용되는 DECO 사전의 기본 구성

특정 도메인의 사용자생성문에서 나타나는 개체명/자질명/감성표현을 처리하기 위한 코퍼스 주석 언어자원을 구축하기 위해 본 연구에서 사용한 DECO 한국어 전자사전에는 30만여 개의 표제어가 수록되어있다. 여기에는 기본적인 형태-통사 정보들과 더불어 다양한 의미-개체명-도메인-감성 정보 등이 할당되어 있다. 실제 DECO 사전에 수록된 표제어의 예를 보이면 다음과 같다.

(5) ㄱ. 갤럭시.NS01+ZNW+XXPR+XITP+…
ㄴ. 카메라.NS01+ZNF+XXPR+XQFT+XITP+…
ㄷ. 화질.NS03+ZNX+XQFT+XITP+…
ㄹ. 예쁘다.AS16+ZAZ+QXPO+…

DECO 사전의 표제어에는 품사 및 활용 클라스 <N/V/A/DS-->와 품사 세부범주 <ZN/ZV/ZA/ZD->와 관련된 내용이 먼저 할당되며, 세부적인 언어 정보들이 후속적으로 부착되는 형태로 구성된다. 이러한 부가 정보들 가운데 개체명과 관련된 정보는 (5ㄱ, ㄴ)의 <XXPR>과 같이 <XX-AA> 형식의 태그로 할당되며, 자질명과 관련된 정보는 (5ㄴ, ㄷ)의 '카메라/화질' 표제어에 부착된 태그와 같이 <XQFT> 형태로, 도메인 관련 정보는 <X-AAA> 형태로 부착되어 있다. 이러한 종류의 태그들은 DECO 사전의 개체명·도메인 관련 어휘 분류체계에 따라 개별 표제어에 부착된 의미 정보이다. 또한 DECO 사전에는 표제어의 감성 정보를 일련의 태그 정보로 제공하는데, 가령 (5ㄹ)의 '예쁘다'는 긍정의 극성값을 갖는 표제어로서 <QXPO> 태그 정보가 할당되어 있다. 이러한 태그들은 DECO 사전의 감성 어휘 분류체계에 따라 세분되어 있으며, <QX-AA> 형식으로 각 표제어에 부착된다. 감성 어휘 분류체계는 비단 긍·부정의 극성 정보 뿐만 아니라 표제어에 대한 부정적 표현 및 심리적 속성 정보를 제공해준다는 점에서 감성분석용 언어자원 구축 시 유용하게 활용된다.

3.3 평가대상 개체명과 자질명

본 연구에서는 특정 도메인을 설정하여 자질기반 감성분석 연구를 수행하기 위해 화장품 리뷰 텍스트를 그 대상 도메인으로 선정하였다. 여기서 평가대상 및 자질명을 주석 처리하기 위한 언어자원을 구축하기 위해, DECO 사전에서 사용하는 EntLEX/FeaLEX와 DomLEX 분류 표제어를 활용하였다. EntLEX/FeaLEX 분류체계는 11가지의 '개체명' 대분류와 1개의 '자질명' 분류 태그로 구성되는데, 각각의 태그들에 따른 세부 정보는 다음과 같다.

(6) ㄱ. 인물 — 인명(XXPE) | 직무/직업(XXHU) | 조직(XXOR)
ㄴ. 공간 — 자연공간(XXGE) | 인공공간(XXLO)
ㄷ. 시간 — 명시적 시간(XXTI) | 비명시적 시간(XXEV)
ㄹ. 사물 — 구체물 — 이동 불가(XXCO) | 이동 가능/상위(XXTH) | 이동 가능/하위(XXPR)
ㅁ. 사물 — 추상물(XXCR)
ㅂ. 자질 — 자질 명사(XQFT)

DECO 사전의 Ent/FeaLEX를 활용하여 사전 표제어 내 개체명을 확보하기 위해, 화장품 도메인 내에서 브랜드/조직과 관련된 표제어를 포함하는 <XXOR> 태그와, '크림/로션'과 같은 제품의 상위어들을 포함하는 <XXTH> 태그, 화장품 개별제품명을 포함하는 <XXPR> 태그가 사용되었다. 마찬가지로 자질명의 일괄 색인에는 FeaLEX에 할당된 자질명 의미 태그인 <XQFT>를 사용하여 색인 된 표제어들을 검토하였다. 하지만 개체명/자질명 태그 정보가 광범위한 표제어에 할당되어 있는 의미 범주이기 때문에, 해당 어휘들이 나타나는 코퍼스 도메인 정보를 통해 표제어 목록을 한정시켜줄 필요가 있다. 이를 위해 DECO 사전의 도메인 분류체계인 DomLEX를 활용하였다.

(7) ㄱ. 정치사회 — 정치(XPOL) | 사회(XSOC)
ㄴ. 경제쇼핑 — 경제(XECO) | 전자(XITP) | 의류(XCLO) | 화장품(XCOS)
ㄷ. 문화예술 — 문화(XCUL) | 영화(XMOV) | 스타(XACT) | 가수(XSIN) …
ㄹ. 서비스관광 — 생활(XLIF) | 세계(XWOR) | 관광(XTUR) | 맛집(XRES) …

(7)의 도메인 분류체계에 따르면, 화장품 도메인과 관련된 표제어에는 <XCOS> 태그가 할당된다. 따라서 화장품 도메인에 대한 개체명 및 자질명을 획득하기 위한 도메인 특정자(specifier)로 <XCOS> 태그를 활용하였으며, 이러한 과정을 통해 다음과 같은 표제어 리스트를 확보하였다. 이를 바탕으로 도메인 어휘 사전에 할당되어야 할 개체명/자질명 리스트를 획득하였다.

(8) ㄱ. <아벤느.NS01+ZNW+XXOR+XCOS+…>
ㄴ. <입생.NS02+ZNW+XXOR+XCOS+…>
ㄷ. <틴트.NS01+ZNW+XXTH+XCOS+…>
ㄹ. <크림.NS02+ZNW+XXTH+XCOS+…>
ㅁ. <플로르드망힐렉스토너.NS01+ZNW+XXPR+XCOS+…>
ㅂ. <캐비어수분강화집중케어실크마스크.NS01+ZNW+XXPR+XCOS+…>
ㅅ. <색.NS02+ZNW+XQFT+XCOS+…>
ㅇ. <색감.NS02+ZNW+XQFT+XCOS+…>
ㅎ. <성분.NS02+ZNW+XQFT+XCOS+…>

이를 (8)의 예시를 통해 살펴보면, <XXOR(+XCOS)> 태그가 할당되어 화장품 도메인의 브랜드/제조사에 해당하는 (8ㄱ, ㄴ)의 '아벤느/입생' 등과 같은 표제어나, (8ㄷ, ㄹ)의 '틴트/크림'과 같이 <XXTH> 태그가 부착된 화장품 제품 종류에 대한 표제어에는 대체로 단일어휘 형태의 표제어들이 나타나는 것을 알 수 있다. 하지만 화장품 도메인 명사 중 가장 많은 부분을 차지하는 화장품 개별제품명 태그 <XXPR>가 할당된 표제어는 (8ㅁ, ㅂ)의 '플로르드망힐렉스토너/캐비어수분강화집중케어실크마스크' 과 같이 다단어 명사구 구성이 띄어쓰기 없이 등재되어 있는 경우들이 많았으며, 이에 따라 해당 표제어들을 도메인 사전에 추가하는 경우 개체명 패턴 구성의 목적에 부합하는 형태로 등재해야 한다. 화장품 도메인의 자질 명사를 추출한 (8ㅅ-ㅎ)의 경우, 표제어가 단일어휘 형태로 나타나는 경우에도, '색/색깔' 등과 같이 하나의 의미 카테고리로 정규화(normalize)해 줄 수 있는 경우들을 고려하여 도메인 사전을 구성하는 것이 필요하다.

또한 도메인 사전 어휘들을 위 사전 태그에 따라 동일하게 구성하는 경우, DECO 사전 표제어와 중첩되는 경우 사전의 처리 효율성을 저하하는 부작용을 낳는다. 본 연구에서는 DECO 사전의 범용성을 훼손하지 않는 차원에서 도메인 사전을 최대한 활용하기 위해 도메인 사전 어휘 한정 태그 <DOM>을 할당하였다. <DOM> 태그는 DECO 사전의 태그 분류에 중첩되지 않는 태그로, DECO 사전의 범용 태그를 바탕으로 코퍼스 내 어휘를 색인 및 처리하고자 하는 경우 해당 태그를 통해 도메인 어휘들을 예외 시킬 수 있다.

위 과정을 통해 구성된 개체명/자질명 도메인 사전에 화장품 도메인의 개체명 인식률을 높이기 위해, 본 연구에서는 코퍼스 수집 시 출현한 686개의 개체명 메타정보를 분석하여 개체명 사전 내 어휘를 확장하였다.

3.3.1 개체명 처리를 위한 언어자원

화장품 코퍼스에서 나타나는 평가대상으로서의 개체명 어휘들을 처리하기 위해 구축된 사전은 예문 (9)와 같으며, 표제어에 수록되어있는 태그들을 통해 화장품 개체명임이 드러난다.

(9) ㄱ. <랑콤.NS01+ZNW+DOM+DomInfo=XCOS+DomEFT=XXOR+EntType=BRD>
ㄴ. <블로썸.NS02+ZNW+DOM+DomInfo=XCOS+DomEFT=XXPR+EntType=MOD>
ㄷ. <오일.NS03+ZNW+DOM+DomInfo=XCOS+DomEFT=XXTH+EntType=TYP>

도메인 사전의 모든 표제어에는 ‘NS’라는 품사 정보 및 조사 부착과 관련된 활용클라스 번호가 부착되어 있으며, 이에 도메인 사전에 기반한 표제어라는 것이 <DOM> 태그를 통해 공통적으로 명시된다. 이외에 화장품 도메인 사전의 개체명 표제어는 두 가지 도메인 의미 정보가 표제어의 특성에 따라 태그의 형태로 부착되어 있으며 이는 다음과 같이 설정되었다.

(10) ㄱ. 도메인 사전 어휘 태그 → <DOM>
ㄴ. 도메인 유형 태그(화장품) → <DomInfo=XCOS>
ㄷ. 도메인 개체명 어휘 분류 태그 → <DomEFT=XXOR/XXPR/XXTH>
ㄹ. 개체명 구성 타입(type) 태그 → <EntTyp=BRD/MOD/TYP>

사용자 생성문의 평가대상인 개체명을 인식하기 위해 구성된 도메인 사전 표제어 정보에는 도메인과 관련 분류 정보, 개체명 분류 정보를 비롯하여 화장품 코퍼스의 제품명 결합 구성에 관련된 세부 분류가 반영되어야 하는데, 표제어에 할당된 도메인 의미 정보는 <DOM> 태그를 제외한 모든 태그들이 기존의 <XXXX> 형태의 태그가 아닌 <XXXX=YYYY> 형태의 태그로 할당되어 있음을 알 수 있다. ‘XXXX’ 부분이 태그의 키(key), ‘YYYY’ 부분이 해당 키가 가지는 값(value)으로 구성되는 해당 태그는 도메인 사전 내에서 각각 도메인 유형(DomInfo), 도메인 개체명 어휘 분류(DomEFT)에 대한 정보를 키, 해당 키에 대한 값을 DECO 사전 태그로 할당받게 되어있으며, (9ㄱ)의 ‘랑콤’과 같은 표제어에는 <DomInfo=XCOS+ DomEFT=XXOR>이 할당됨에 따라 해당 표제어가 ‘화장품 도메인의 XXOR’ 개체명임이 표시되게 된다. 또한 개체명 구성어휘 표제어의 경우 <EntTyp=XXX>의 키-값 태그가 부착되어 있는데, 이는 표제어가 가지는 의미 결합성 정보가 할당되어 있는 태그로, (9ㄴ)의 표제어 ‘오일’의 경우 제품의 종류에 해당하는 표제어임에 따라 <EntType=TYP> 태그가 부착되어 있다. 해당 태그에 부여되어야 하는 값의 종류는 개체명 구성의 양상을 반영

하는데, 본 연구에서는 브랜드(BRD)/수식 어휘(MOD)/제품 종류(TYP)의 세 가지 개체명 타입 분류를 LGG 구문 패턴과 결합하여 화장품 코퍼스의 개체명 의미 정보를 처리하였다. 이때의 구문 패턴 또한 앞서 수집된 개체명 메타데이터의 분석을 통해 획득할 수 있으며, LGG를 통해 구체화된 개체명 구문 패턴의 모습은 다음 <그림 2>와 같다.

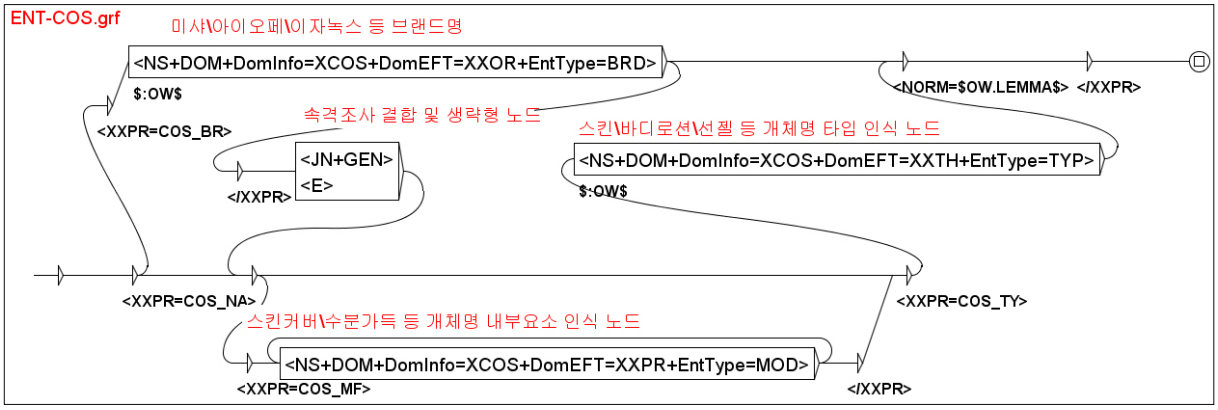

그림 2. 화장품 도메인 코퍼스의 개체명 인식 LGG

위 개체명 인식 LGG는 해당 도메인에서 나타나는 개체명의 결합적 특징을 구조화하여 기술하고 있다. 해당 패턴에는 개체명 구성어휘 각각이 실현 가능한 위치가 고려되어 있으며, 도메인 사전의 태그 형태로 반영된 정보들을 활용하여 패턴을 인식한다. 대부분의 노드(node)에는 명사 태그 <NS>와 도메인 태그 <DOM>이 '+' 연산자를 통해 연결되어 있음을 알 수 있는데, 이는 도메인 개체명 구성의 기본적인 단위를 전체 사전의 범위 내에서 한정 짓기 위해 기술된 것이다. 각각의 노드에는 해당 위치에 맞는 개체명 구성어휘 부류를 호출하기 위한 세부 정보가 할당되어 있으며, 다음과 같은 태그 해석을 지닌다.

(11) ㄱ. 브랜드명→ <NS+DOM+DomInfo=XCOS+DomEFT=XXOR+EntType=BRD>
의미: <명사+도메인 사전어휘+화장품 도메인+집단+브랜드>
ㄴ. 제품 수식어→ <NS+DOM+DomInfo=XCOS+DomEFT=XXPR+EntType=MOD>
의미: <명사+도메인 사전어휘+화장품 도메인+개체명+수식어>
ㄷ. 제품 타입→ <NS+DOM+DomInfo=XCOS+DomEFT=XXTH+EntType=TYP>
의미: <명사+도메인 사전어휘+화장품 도메인+개체명 상위어+제품종류>

전체 경로상에는 화장품의 브랜드 및 제품 타입(예: 파운데이션, 마스크팩)과 같이 단일 평가대상 어휘로 나타날 수 있는 개체명 유형의 출현을 필요조건으로 설정하여 과생성을 막고, 처리에 충돌을 일으킬 수 있는 부류들을 최대한 억제하는 한편, 조사 및 다른 범주의 어휘가 나타날 때까지 최장 결합의 형태를 제한하지 않는 방식으로 구성된다. LGG-오토마톤을 통해 인식된 단일어/다단어 형태의 개체명에는 XML 형식으로 구상된 <XXPR=COS_XX> 형태의 화장품 개체명 태그가 할당되며, 추후 자질기반 감성분석 과정에서 평가대상인 개체명 분석에 사용된다. 이에 따라 해당 그래프를 통해 실제 인식되는 표현들의 예를 제시하면 <그림 3>과 같다.

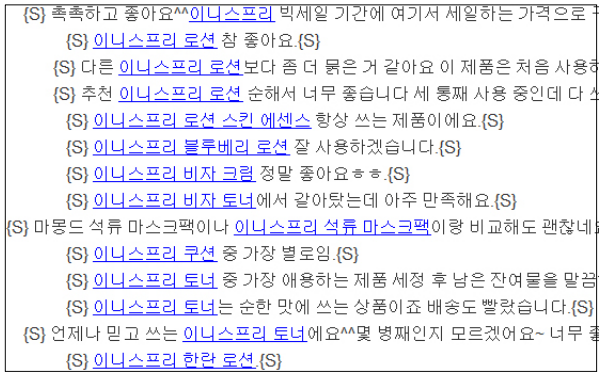

그림 3. LGG를 통해 인식된 화장품 도메인 코퍼스의 개체명 일부

3.3.2 자질명 처리를 위한 언어자원

자질명 또한 DECO 사전을 통해 확보된 도메인 어휘들을 기반으로 코퍼스에서 나타나는 일반 명사들의 빈도를 기술통계적으로 분석하여 그 외연을 확장하였다. 현재 도메인에서는 '가격/성분/향기/용량/보습력'과 같은 어휘들이 이러한 부류에 속한다. 다음을 보자.

(12) ㄱ. <값.NS02+ZNW+DOM+DomInfo=XCOS+DomEFT=XQFT+NORM=디자인>
ㄴ. <제형.NS02+ZNW+DOM+DomInfo=XCOS+DomEFT=XXPR+NORM=제형>
ㄷ. <고보습.NS03+ZNW+DOM+DomInfo=XCOS+DomEFT=XXTH+NORM=보습력+DomPola=QXPO>
ㄹ. <보습력.NS02+ZNW+DOM+DomInfo=XCOS+DomEFT=XXPR+NORM=보습력+DomPola=QXDE>

위 예시의 자질명을 처리하기 위해 구축된 도메인 사전 또한 앞서 개체명 표제어 예시에서와 같이, 도메인 사전 태그/도메인 분류/개체명-자질명 분류 태그에 대해서 같은 키-값 형태의 태그를 부여하고 있다. 하지만 해당 표제어가 자질명에 해당하므로 <DomEFT>의 태그의 값은 DECO 사전 자질명 태그인 <XQFT>로 할당되어 있다.

화장품 도메인의 자질명 처리를 위한 도메인 사전에서 추가된 의미 정보는 정규화 어휘 정보와 도메인 극성 정보이다. 먼저 정규화 어휘 정보는, 자질명으로 나타나는 다양한 어휘들이 특정한 종류의 카테고리화되는 것이 가능하기 때문에 설정된 것으로, '질감/입자/제제/제형/타입'과 같은 어휘들은 화장품 코퍼스 내에서 '제형'에 대한 자질명으로 범주화 하는 것이 가능하다. 이에 따라, 자질기반 감성분석을 위한 자질명의 의미처리를 위해선 각각의 범주에 속한 어휘들이 동일한 평가 자질을 의미하도록 지정해주어야 한다. 이에 따라 화장품 도메인 사전의 자질명을 약 20개의 어휘를 통해 범주화하였으며, 범주를 대표하는 어휘를 지정하여 <NORM=대표어휘>로 할당하여 추후 자질기반 감성분석 시스템 상에서 이를 동일한 자질명 카테고리로 처리할 수 있게 하였다.

새로 추가된 또 다른 의미 태그는 도메인 극성 정보로, 이는 도메인 코퍼스에서 해당 표제어가 가지는 일반적인 극성에 따라 부착되는 태그이다. 이러한 유형에는 '고보습/고광택' 등과 같이 자체적으로 긍정의 의미를 내포하고 있는 표제어에 부착되었으며, 이는 서술격 조사 '-이다'와 공기하였을 때 극성이 명확해지는 특징을 보인다. 이와 같이 하나의 확실한 극성을 택하는 경우와 반대로, '보습력/수분감'과 같이 존재 형용사 '있다/없다'와 공기함에 따라 극성이 달라지는

경우에는 상대극성(QXDE) 태그가 할당되었다. 부여된 도메인 극성 정보는 이후 감성표현 처리에 활용된다.

앞선 자질명들이 자질명을 명시적으로 드러난 문장을 처리하기 위한 경우였다면, 반대로 자질이 어휘를 통해 명시적으로 실현되지 않고 암시적(implicit)으로 함축되는 다음과 같은 유형이 존재한다.

(13) ㄱ. 그래도 생각보다는 번들거리지 않아요. [흡수력]
ㄴ. 반나절도 안돼서 다시 건조해지네 [보습력]
ㄷ. 다른 크림들보다 저렴하진 않음 [가격]

위의 예문에 나타난 극성 형용사들은 자질명과 공기하지 않았으나 암시적인 자질명을 함축하고 있다. 이러한 암시적 자질명을 인식하기 위하여 코퍼스에서 명시적 자질명과 공기하여 나타나는 극성 형용사들을 분석한 후, 이들을 다시 문서단어행렬(document-term matrix)을 통해 빈도 분석하여 유의미한 '명시적 자질명-암시적 자질명' 관계를 추출하였다. 이와 같은 자질명 추출을 위한 이중 증식(double propagation)을 진행하였으며, 이렇게 극성 어휘에 부여되는 암시적 자질명은 추후 자질기반 감성분석 시에 명시적 자질명이 실현되지 않은 경우 자질로 기능하게 하였다.

3.4 단일어/다단어 감성표현(Sentiment Expression)

도메인 코퍼스에서 나타나는 단일어와 다단어 감성표현 인식을 위해 DECO 사전의 감성 어휘 분류체계의 하나인 PolLEX 극성 어휘 분류체계를 활용하였다. 해당 분류에 속하는 표현들은 단일어휘 형태의 극성 표현들로, '예쁘다/천사'와 같은 긍정 표현들은 <QXPO> 태그가 할당되어 있으며, '못생기다/멍청이'와 같은 부정 표현들에는 <QXNG>가 할당된다. 이러한 분류체계만을 활용하여 13,000개의 표제어와 그 활용형에 대한 정보를 극성 처리할 수 있다. <표 1>은 현재 DECO 사전에 수록된 극성 어휘의 분포를 보인다.

표 1. DECO-PolaLEX 극성 어휘 분포

구분	명사	동사	형용사	부사	합계
강한긍정<QXSP>	174	312	326	550	1,362
보통긍정<QXPO>	1,133	1,882	1,263	1,709	5,987
강한부정<QXSN>	254	1,040	763	1,052	3,109
보통부정<QXNG>	2,887	3,531	1,649	2,097	2,541

이에 더하여, 특정 도메인에서만 의미를 갖는 도메인 의존적 감성표현을 기술하기 위해 무극성 어휘의 빈도정보 및 연어 관계 등을 추출 및 분석함에 따라 감성 극성과 관련한 어휘적 특이성을 보이는 구 단위 표현들을 수집하였으며 이를 세 가지 패턴으로 분류하였다.

(14) ㄱ. 체언 + 체언 구성 : 착붙 템(+), 가루 날림(-), 요플레 현상(-)
ㄴ. 체언 + 용언 구성: 다크닝이 사라지다(+), 피부톤을 잡다(+)
ㄷ. 용언 + 용언 구성: 바르기 쉽다(+), 떠 보이다(-)

위 표현들은 공통적으로 해당 어구를 구성하는 체언 및 서술어가 모두 극성이 존재하지 않았음에도 불구하고, 다단어 결합으로 극성이 발현되는 관용표현의 형식을 취하기 때문에 개별 표현에 맞는 극성을 패턴 문법을 통해 인식 및 처리해 줄 필요가 있다. 이에 따라 본 연구에서는 해당 유형을 처리하기 위한 LGG 패턴 문법을 구축하였다. 이와 반대로 코퍼스 도메인과 관계없이 지정된 극성으로 실현되는 범용 다단어 극성 표현들 또한 존재한다. 이는 다음과 같이 나타난다.

(15) ㄱ. 체언 + 체언 구성: 끝판 왕(+), 기분 다운(-)
ㄴ. 체언 + 용언 구성: 마음에 들다(+), 인기가 없다(-)
ㄷ. 용언 + 용언 구성: 끝내 주다(+), 물고 늘어지다(-)

이러한 표현들 또한 세 가지의 구문 패턴을 통해 유형화할 수 있으므로 코퍼스에서 나타나는 범용 다단어 극성 표현에 포함하여 기존에 공개되어 있는 관용어 사전 등을 일부 이용하여 그래프 문법으로 구성하였다.

단일어/다단어 극성 표현들은 문장 내에서 일련의 극성 전환어(Polarity Shifting Device: PSD)로 기능하는 어구와 공기하여 극성이 가감 또는 반전될 수 있다(남지순, 2012; Nam, 2014). 다음 예문을 보자.

(16) ㄱ. 듣던 것보다는 나쁘지 않아요/안 나빠요.
ㄴ. 터치감이 마음에 들진 않아요/안 들어요.
ㄷ. 사람들 말처럼 최고는 아니예요.
ㄹ. 반나절도 안됐는데 수분감이 사라졌어요.

(16ㄱ)의 문장은 일반적으로 나타나는 단일어휘 극성 표현에 '않다'가 결합하는 장형 부정문 형태와, 부정부사 '안'이 결합한 단형 부정문 형태를 보여준다. (16ㄴ) 또한 앞선 경우와 마찬가지로 부정용언 및 부사가 다단어 감성표현에 부착됨에 따라 나타나는 부정문에 해당된다. 주요 극성을 나타내는 어휘가 체언으로 나타나는 경우 (16ㄷ)과 같이 보어화 된 명사 뒤에 '아니다' 또는 '없다'가 PSD로 극성을 반전시킬 수 있으며, 이외에도 '사라지다' 등의 일부 서술어가 '어휘적 부정표현'(남지순, 2018)으로 PSD에 해당된다. PSD 표현들은 의미적 차원에서 감성 어휘의 방향성에 영향을 미치기 때문에 감성표현 처리 시 고려되어야 한다. 본고에서는 해당 공기 관계를 보이는 감성표현들을 처리하기 위한 처리 패턴을 문장성분 간의 결합 제약(composition constraints)을 고려하여 기술하였다. 이러한 PSD 처리를 위해 구성된 그래프 문법 중에서, 단형 부정소 결합 패턴을 감지하는 LGG 문법의 예를 들면 <그림 4>와 같다.

해당 그래프는 단형 부정소 '안'을 부정부사 형태의 PSD를 의미하는 <PSD=NADV>로써 태깅하고, 이와 공기하는 단일어휘 감성표현을 포착하여 결합 시퀀스 단위의 극성 정보를 처리하는 패턴 문법이다. 해당 그래프에서 '<AS/VS+QX-->' 형태의 노드들은 각각 극성이 존재하는 형용사/동사를 호출하기 위해 설정된 것으로, '안/못'의 경우 명사와의 결합이 부자연스럽고, 일반 부사와 공기하는 경우 후치 결합하기 때문에 처리 패턴에서 배제되었다. 또한 용언 앞에서 실현되는 단형 부정소는 '안'은 '-하다'류 용언과의 사용이 어색하기 때문에 결합 가능한 감성표현 부류에서 '-하다'류 용언을 제외하는 문법을 구성하였다(~HAP). 또 다른 단형 부정 어휘인 '못' 또한 해당 패턴에서 배제되었는데, 이는 사용자 후기 글에서 '못'이 부사로써 감성표현과 결합하는 경우가 거의 존재하지 않았기 때문이다. 그 결과 부정소

‘안’을 중심으로 단형 부정 패턴 처리 자원을 구성하였으며, 각각의 경로상에서 호출된 유극성 용언들은 단형 부정소와 구(phrase)를 이룸에 따라 극성이 다시 분류되는 방식이 사용되었다.

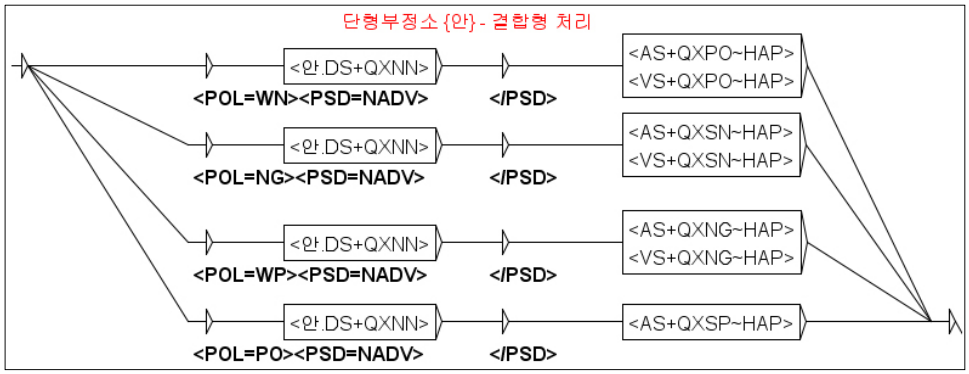

그림 4. 단형 부정소 결합 패턴 그래프 예

해당 과정을 통해 최종 표상되는 구 단위 극성 정보는 (17)에서와 같이 <POL=XX> 형태의 XML 태그를 통해 주석되며, 이는 자질기반 감성분석 시 극성 점수(score) 계산에 활용된다.[1)]

(17) ㄱ. 기분이 <POL=WN><PSD=NADV>안</PSD> 좋아요</POL>
ㄴ. 세일한다는데 <POL=NG>구미가 <PSD=NADV>안</PSD> 당기네</POL>

구문 패턴에 따라 PSD로 분류되어야 하는 두 번째 유형은 장형 부정 구문으로, 해당 구문을 처리하기 위한 LGG 그래프는 장형 부정 문형과 결합하는 어휘를 체언/용언으로 분리하여 각각의 어휘 범주와 결합 가능한 패턴을 별도로 기술하는 방식으로 구성되었다. 이러한 처리 방식은 아래 그림과 같이 구현되어 {VS/AS-QX--} 서브 그래프를 통해 호출된 동사 및 형용사의 극성 어휘들이 장형 부정 표현과 공기하는 경우에는 {PSD-Aux-PRED} 서브 그래프가 해당 문형을 처리하고, {NS-QX--} 서브 그래프에서 호출된 체언 극성 표현들의 장형 부정 어휘의 공기 관계가 포착되는 경우 {PSD-Aux-NOUN} 서브 그래프가 이를 처리하도록 한다.

<그림 5>의 우측에 삽입된 서브 그래프들은 순차적으로 용언/체언과 결합하는 장형 부정 문형을 처리하기 위해 구성된 그래프로, 용언 처리를 위한 경로에는 부정 의미가 내포된 보조 동사인 <않다>가 설정됨에 따라 선행하는 극성 용언과 ‘-지 않다’의 형태로 결합할 수 있도록 수 있도록 설정되어 있다. 이에 반해 체언 처리 경로는 <아니다>를 포함함에 따라 ‘-이/가 아니다’의 형태로 극성 명사와 결합할 수 있도록 설정되었다. 해당 그래프에는 각각의 선행 결합 품사 범주에 부합하는 {‘같다/생각하다’ 류 구문 구조 + 장형 부정} 구문을 처리하기 위한 경로가 서브 그래프의 형식으로 표상되어 있는데, 해당 서브 그래프들이 서술어 및 체언의 ‘-이다’ 결합형과 연결되는 {관형형 전성어미 + 의존 명사}을 처리할 수 있도록 고려되어 실질적으로 더 많은 장형 부정 문형을 처리할 수 있도록 구성하였다.

1) 이때 극성 점수는 {강한 긍정(SP):1.5, 일반 긍정(PO):1.0, 약한 긍정(WP):0.5} 형태로 계산되며, 부정 표현의 경우 이에 역수를 취한다. 어휘 의미 정보 분석을 위해 부착되었던 <PSD=XXXX> 태그는 본처리 과정에서 제거된다.

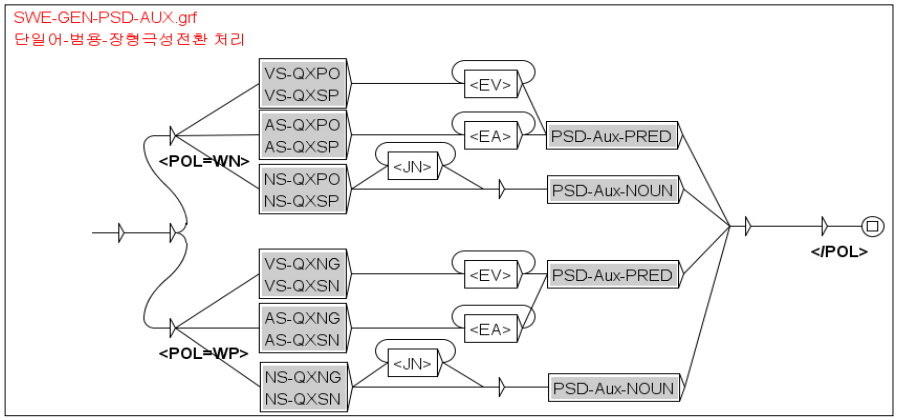

그림 5. 장형 부정 문형 처리를 위한 LGG 모듈의 전개도

추가적으로, 감성 술어의 수식어 부류에는 '문맥을 강화하는 부사(amplifier)' 유형과 '문맥을 약화하는 부사(down toner)' 유형이 존재한다. 다음 예시를 살펴보자.

(18) ㄱ. 이전보다 훨씬 좋아졌어요.
ㄴ. 이번에 나온 건 화장이 덜 밀리네요.

위 예시에서 나타난 정도 강화부사와 약화부사는 각각 (18ㄱ)에서 '훨씬', (18ㄴ)에서 '덜'로, 이들은 각각의 문장 안에서 긍정 표현 '좋아졌어요'와 부정 표현 '화장이 밀리네요'의 극성을 강화/약화한다. 이러한 표현들은 거의 모든 종류의 극성 표현과 공기할 수 있음에 따라 이를 처리하기 위한 경로 패턴 또한 고려되어야 하며, 이는 <그림 6>과 같은 방식을 통해 기술되었다.

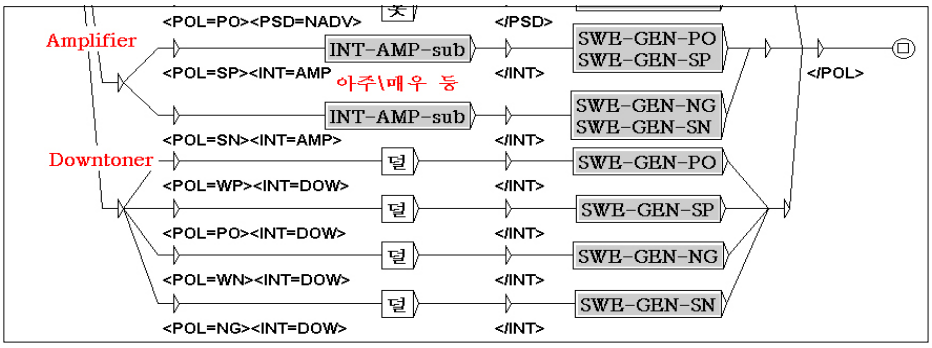

그림 6. 단형 부정소, 강화 & 약화부사어

4. DECO-LGG와 의존 파서 기반 감성분석

이상에서 소개한 DECO-LGG의 구문 패턴을 통해 처리된 코퍼스는 처리 범위 내에서 나타날 수 있는 모든 경우의 개체명/자질명/감성표현에 해당하는 패턴들을 인식하여 의미 태그 주석 처리를 수행한다. 이에 대한 예시로 다음을 보자.

(19) ㄱ. 이니스프리의 수분감은 세타필보다 좋아요.
ㄴ. 이상하던 디자인이 이제 예쁘게 보여요.

(20) ㄱ. <XXPR=COS_NA>이니스프리<NORM=이니스프리></XXPR>의 <XQFT>수분감<NORM=보습력></XQFT>은 <XXPR=COS_NA>세타필<NORM=세타필></XXPR>보다 <QXPO=GEN>좋아요</QXPO>.
ㄴ. <POL=NG>이상하던</POL> <XQFT>디자인</XQFT>이 이제 <POL=PO>예쁘게</POL> 보여요.

언어자원 DECO-LGG는 (19)와 같은 감성분석 대상 문장을 의미 주석하여 (20)과 같은 텍스트를 출력한다. 이때 오피니언 핵심 구성 요소에 대한 의미 정보가 주석 처리되는데, 언어자원이 코퍼스 전체를 색인하며 패턴 시퀀스에 의해 감지된 모든 표현들을 주석처리 하기 때문에 하나의 평가문 내에 복수의 대상 개체명/자질명/감성표현이 할당되는 문제가 발생한다. 따라서 주요 서술어와 실제 평가대상에 해당하는 개체명 및 자질명을 연결하기 위한 과정과, 상반된 극성 정보가 다수 공기하는 문장에 대한 극성 결정이 요구된다. 이는 병렬연결된 구문을 제외한 복문과 단문은 하나의 주요 서술어를 가지며, 문장의 오피니언을 개체명/자질명/극성 평가 1:1:1 관계로 분석하는 자질-카테고리 분석 모델에 기초한 것으로, 이에 따르면 (19ㄱ)의 경우 XXPR이 할당된 두 개의 개체명 '이니스프리', '세타필' 중, 주 서술어의 필수논항(주어)에 해당하는 '이니스프리'를 긍정 극성 정보를 내포하는 주 서술어와 페어링해야 하며, 부정 극성 어휘 '이상하다'와 긍정 극성 어휘 '예쁘다'가 공기하는 (19ㄴ) 문장의 경우 주절의 정보에 가중치를 두어 최종 극성은 '긍정'으로 분석되어야 한다. 이와 같은 자질 기반 감성분석을 수행하기 위해선 문장의 핵심 어휘에 해당하는 서술어의 위치를 찾아내야 하며, 해당 서술어와 연결되는 논항 요소들을 감지하기 위한 구문 분석이 요구된다. 이에 따라 본 장에서는 DECO-LGG를 통해 제공된 평가대상 개체명/자질명/극성 어휘 정보와 문장 내 구문의 의존 관계를 종합적으로 처리하여 감성분석을 진행하는 것을 소개한다.

의존 구문 분석을 위해 사용된 파서로 본 연구에서는 카이스트에서 개발한 'NLP_HUB' 의존 파서(함영균 외, 2012)를 사용하여, 문장 단위에서 나타나는 대상/자질명 페어링과 구문 깊이(depth)에 따른 극성 어휘 가중치 처리를 진행하였다. 의존 구문 분석은 기본적으로 핵(ROOT)이 되는 단어와 그 의존 요소 사이의 관계를 처리하는데, 특정 어순을 제한하지 않아서 격조사에 따라 어순이 자유로운 한국어의 구조에 용이하다. 'NLP_HUB' 의존 파서는 세종 의존 관계 태그셋(tagset)에 따라 구문을 처리하는데, 각 어절은 지배소와 해당 어절 사이의 의존 관계로 분석된다. 세종 의존 관계 태그셋은 문장의 구문을 분석했을 때, 각 어절이 지배소와 가질 수 있는 의존 관계의 유형을 정리한 체계로, <표 2>(Choi et al., 2012)에 정리된 구문 태그셋과 기능 태그셋을 조합하여 사용된다. 구문 태그의 경우 처리 대상 어절이 체언인지, 용언인지, 관형사인지, 부사인지 등을 구분하며, 기능 태그의 경우 지배소와 가지는 관계가 주격인지, 목적격인지, 관형격인지 등을 표현한다.

표 2. 세종 의존 관계 태그셋

구문 태그셋		기능 태그셋		
NP	체언	SBJ	주어	
VP	용언	OBJ	목적어	
VNP	긍정 지정사구(명사+이다)	CMP	보어	
AP	부사구	MOD	MOD	관형어
			NMOD	체언 수식어
			VMOD	용언 수식어
DP	관형사구	AJT / ADV	부사어	
IP	감탄사구	CNJ	접속어	

위 (19ㄱ, ㄴ)의 예문들에 대한 의존 구문 분석은 (21)과 같이 진행된다. 각 어절의 의존 관계는 구문 태그와 기능 태그를 '_'로 결합하여 표현되며, 해당 지배소의 정보가 대괄호 형태로 제공된다.

(21) ㄱ. 의존 구문 분석 결과(이니스프리의 수분감은 세타필보다 좋아요.):
- 이니스프리의 → NP_MOD [수분감은]
- 수분감은 → NP_SBJ [좋아요]
- 세타필보다 → NP_AJT [좋아요]
- 좋아요 → VP_ROOT

ㄴ. 의존 구문 분석 결과(이상하던 디자인이 이제 예쁘게 보여요.):
- 이상하던 → NP_MOD [디자인이]
- 디자인이 → NP_SBJ [보여요]
- 이제 → AP_ADV [보여요]
- 예쁘게 → VP_VMOD [보여요]
- 보여요 → VP_ROOT

종합적으로, (19ㄱ)의 예문이 DECO-LGG를 통해 (20ㄱ)과 같이 의미 정보를 할당 받고, 구문의 의존 관계 정보를 담은 (21ㄱ)으로 분석되어, 핵어(ROOT)의 직접지배를 받는 부가적 정보인 용언 수식어(VMOD)를 포함한 부사어(AJT) 외의 필수논항이 대상 개체명/자질명 페어링 및 극성 판별에 우선 처리될 수 있도록 진행된다. 이를 통해 DECO-LGG 의미 태깅과 구문 분석 결과가 (22)와 같이 최종 분석된다.

(22) ㄱ. DECO-FbSA 분석 결과(이니스프리의 수분감은 세타필보다 좋아요.):
- <XXPR>이니스프리</XXPR>의 → NP_MOD [수분감은]
- <XQFT>수분감</XQFT>은 → NP_SBJ [좋아요]
- ~~세타필보다 → NP_AJT [좋아요]~~
- <POL=PO>좋아요</POL> → VP_ROOT

ㄴ. DECO-FbSA 분석 결과(이상하던 디자인이 이제 예쁘게 보여요.):
- <POL=NG>이상하던</POL> → NP_MOD [디자인이] -Weight
- <XQFT>디자인</XQFT>이 → NP_SBJ [보여요]
- 이제 → AP_ADV [보여요]
- <POL=PO>예쁘게</POL> → VP_VMOD [보여요] +Weight
- 보여요 → VP_ROOT

4.1 주요 서술어의 의존 구문 분석을 통한 평가 개체명/자질명 페어링

한 문장에서 다수의 개체명/자질명이 공기할 때, 주요 서술어의 필수논항 및 깊이(depth) 정보를 처리하는 것은 매우

중요하다. 이는 단순히 부가어 태그를 부여받은 어절만 선별하는 방식이 아니라, 해당 어절이 지배 관계를 가지는 의존소까지 그룹화하여 처리해야 함을 뜻하며, 반대로 이는 필수논항과 의존 관계를 가지는 의존소 어절의 깊이별 의미 정보 처리가 필요함을 의미한다.

<그림 7>은 (22ㄱ)의 분석 결과를 깊이 정보와 함께 시각화한 것으로, 본 연구에서는 입력 문장에 DECO-LGG를 적용하여 오피니언 구성의 핵심 의미 정보인 개체명(XXPR), 자질(XQFT), 극성(QXPO) 정보를 처리하고, 의존 구문 분석을 통해 문장성분과 핵어와의 거리, 즉 핵어를 최상단으로 계산한 문장의 깊이를 처리함에 따라 서술어의 필수논항의 의미 정보만을 오피니언 구성 요소로써 선별하는 문장 단위의 개체명/자질 정보 페어링을 수행한다.

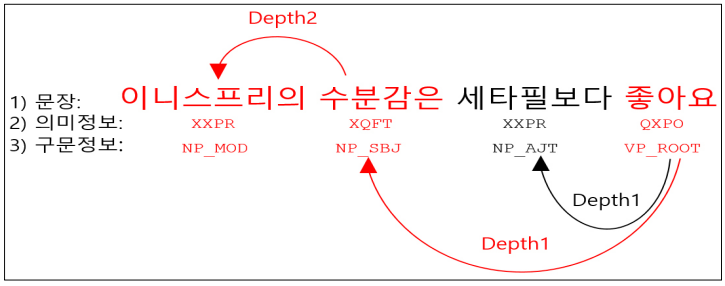

그림 7. 의존 구문 분석을 통한 대상/자질명 페어링

필수논항과 부가어의 구문 의존 관계를 구별하기 위해서, 핵어의 직접지배를 받는 구문의 의존 관계 분석에 깊이-우선 탐색(depth-first search) 알고리즘을 활용하였다. 이를 통해 깊이 정보를 정수형으로 활용할 수 있는데, (23ㄱ)과 같이 부가어가 없는 문장에서 필수논항 내 복수의 개체명/자질명이 공기할 때, 핵어와의 깊이가 가까운 개체명 어절이 최종 대상 개체명 및 자질명으로 가중치를 받게 된다.

(23) ㄱ. 에이블씨엔씨의 미샤가 가성비는 정말 좋아요.

ㄴ. <XXPR=COS_NA>에이블씨엔씨</XXPR>의 <XXPR=COS_NA>미샤</XXPR>가 <XQFT>가성비</XQFT>는 <POL=SP>정말 좋아요</POL>.

ㄷ. 의존 구문 분석 결과:

- 에이블씨엔씨의 → NP_MOD [미샤가]
- 미샤가 → NP_SBJ [좋아요]
- 가성비는 → NP_SBJ [좋아요]
- 정말 → AP_ADV [좋아요]
- 좋아요 → VP_ROOT

(23)의 예문을 가중치 처리함에 따라 도출된 결과인 (24)를 통해 알 수 있듯 '에이블씨엔씨'와 '미샤'가 둘 다 DECO-LGG를 통해 개체명 태그(XXPR)를 할당 받았음에도 불구하고, 핵어와 가까운 '미샤'가 실제 대상 어휘로 페어링되었음을 확인할 수 있으며, 자질명을 감성표현과 페어링하는 과정 또한 이와 동일한 처리 과정을 통해 수행됨을 알 수 있다.

(24) DECO-FbSA 분석 결과(미샤가 이자녹스보다 가성비가 훨씬 좋아요.):

- 에이블씨엔씨의 → NP_MOD [미샤가]
- <XXPR>미샤</XXPR>가 → NP_SBJ [좋아요] +Weight
- <XQFT>가성비</XQFT>가 → NP_SBJ [좋아요]
- <POL=SP>정말 → AP_ADV [좋아요]
- 좋아요</POL> → VP_ROOT

4.2 의존 구문 분석 기반 극성어 가중치 처리

어휘의 극성 정보에 기반한 감성분석의 경우, 단일 문장에 서로 다른 다수의 극성 정보가 충돌할 때 이를 효과적으로 처리할 필요가 있다. 의존 구문분석을 통한 깊이 정보는 이에 효과적으로 적용될 수 있다.

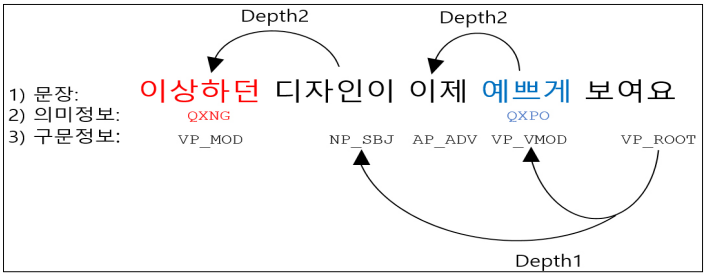

그림 8. 의존 구문 분석을 통한 극성 어휘 가중치 처리

<그림 8>은 서로 다른 극성 어휘가 공기하는 (19ㄴ)의 문장 처리 예시이다. 의존 관계를 나타내는 화살표는 지배소가 의존소에 가지는 방향성을 나타낸다. 해당 문장은 부정 어휘와 긍정 어휘가 동시에 존재하지만, 긍정 의미 정보(QXPO)를 할당받은 '예쁘게'의 깊이가 핵어에 가깝기 때문에, 문장의 주어(NP_SBJ)를 수식하기 위한 부정 극성(QXNG) 관용어 '이상하던' 보다 높은 가중치를 받도록 처리하여 해당 문장의 최종 극성을 긍정으로 분석한다. 이는 긍·부정 어휘를 처리하는 '어휘연산 감성분석(Lexical algorithmic Sentiment Analysis)' 방법(Thelwall et al., 2012)에 기반하여, 각 어휘의 극성 점수를 깊이로 나눔에 따라, 깊이가 핵어와 가까운 감성표현일수록 더 높은 점수를 받도록 가중 처리한다.

가중처리 방식과 더불어, 극성 분석에 부가어를 제외하는 방식 또한 활용되었는데, 이는 다수의 극성 어휘가 공기하는, (25)와 같은 복문 구조에 적용된다.

(25) ㄱ. 광고에선 좋은 말이 가득한데 실제 성분은 싸구려다.

ㄴ. 광고에선 <POL=PO>좋은</POL> 말이 가득한데 실제 <XQFT>성분</XQFT>은 <POL=NG>싸구려</POL>다.

ㄷ. 의존 구문 분석 결과:

- 광고에선 → NP_NMOD [말이]
- 좋은 → VP_MOD [말이]

- 말이 → NP_SBJ [가득한데]
- 가득한데 → VP_VMOD [싸구려이다]
- 실제성분은 → NP_SBJ [싸구려이다]
- 싸구려이다 → VNP_ROOT

(25ㄴ)처럼 나타나는 복수의 극성 정보는 아래의 (26)과 같이 처리되어, 문장의 최종적인 극성은 '부정'으로 결정된다.

(26) DECO-FbSA 분석 결과(광고에선 좋은 말이 가득한데 실제 성분은 싸구려다):
- ~~광고에선 → NP_NMOD [말이]~~
- ~~좋은 → VP_MOD [말이]~~
- ~~말이 → NP_SBJ [가득한데]~~
- ~~가득한데 → VP_VMOD [싸구려이다]~~
- 실제<XQFT>성분</XQFT>은 → NP_SBJ [싸구려이다]
- <POL=NG>싸구려</POL>이다 → VNP_ROOT

5. LSTM 기반 감성분석

기본적으로 언어자원에 기반한 감성분석 모듈은 입력 텍스트가 'ㅁㅏㅁ에 안드러(맘에 안 들어)' 등과 같이 심각한 비정형성을 띄는 경우, OOV 문제가 대두된다. DECO-LGG의 의미 정보 및 의존 구문 분석 결과를 통해 자질기반 감성분석을 수행하는 경우, 제한적인 정형 데이터에 기반하여 학습된 의존 구문 분석기가 초래하는 토큰화(tokenize) 오류로 인한 의미 정보 소실 문제를 해결할 필요가 있다. 이는 아무리 언어자원을 도메인 특화하여 의미 정보를 제공할지라도, '한나눔'과 같은 의존 구문 분석기가 파싱 단계에서 단어 경계를 잘못 처리하면 의미 정보가 매핑되지 않기 때문이다. 본 장에서는 앞서 서술한 문제점을 딥러닝 기반의 감성 분류 기법을 사용하여 해결하는 처리 과정을 소개한다.

(27) ㄱ. 이번 제품 수분감이 영 별로였다.
ㄴ. 이번 제품 <XQFT>수분감</XQFT>이 영 <POL=NG>별로</POL>였다.
ㄷ. 의존 구문 분석 결과:
- 이번 → NP_NMOD [제품]
- 제품 → NP_NMOD [수분감이]
- 수분감이 → NP_SBJ [로였다]
- 영별 → NP_NMOD [로였다]
- 로였다 → VP_ROOT

의존 구분 분석기의 토큰화 오류로 인한 의미 정보 소실 문제는 (27) 예시를 통해 확인할 수 있다. (27ㄴ)에서는 DECO-LGG를 통해 '별로'의 부정 극성 정보가 처리되었으나, 의존 구문 분석을 진행한 (27ㄷ)에선 '영 별로였다'를 '영

별(NP_VMOD)', '로였다(VP_ROOT)'로 잘못 토큰화하여, '별로'에만 할당될 수 있는 부정 극성 정보가 소실될 수밖에 없음을 알 수 있다.

이처럼 다양한 이유로 처리 범주를 벗어난 OOV 발생문을 분석하기 위해 시퀀스 정보를 효율적으로 처리하는 LSTM 모델을 활용하였다. 이는 문장의 극성 처리를 위해 사용되며, 이때 의존 구문분석의 토큰화 오류로 인해 오피니언 의미 정보가 소실된 (27ㄱ)과 같은 텍스트가 입력문으로 들어올 때 감성분석을 진행한다. 단순 RNN 구조의 경우 시퀀스가 길어질수록 가중치의 값이 특정 값에 수렴하게 되는데, 이는 학습 경과에 따라 초기에 입력되었던 데이터의 영향력이 감소하는 장기의존성 문제로 이어진다. 이러한 단순 RNN 모델의 한계점을 보완하기 위해 Hochreiter과 Schmidhuber(1997)에 의해 제안된 LSTM 모델은 망각(forget), 입력(input), 출력 게이트(output gate)를 통해 필요맥락의 정보를 선별함에 따라 장기의존성 문제를 효과적으로 해결한다(손진광, 2017).

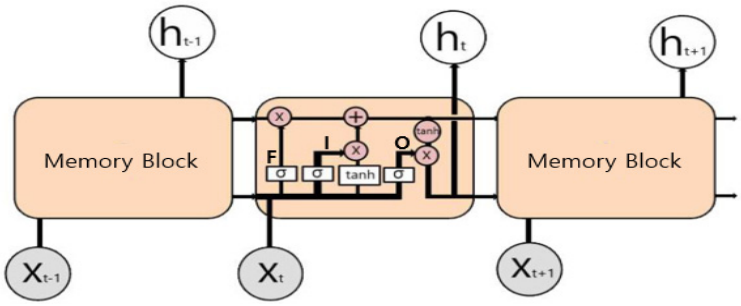

그림 9. LSTM의 메모리 셀 구조

<그림 9>를 통해 확인할 수 있듯이, LSTM 구조는 입력된 토큰 Xt가 망각(F), 입력(I), 출력(O) 게이트를 순차적으로 통과하도록 프로그램 되었다. 각 게이트를 통과한 데이터들은 셀 상태(cell state)에 저장된다. 각각의 게이트는 특정한 연산을 수행하는데, 망각 게이트에서는 이전 시점의 정보를 잊어버릴지를 계산하며, 시그모이드(sigmoid)층과 하이퍼볼릭 탄젠트(tanh)층의 두 부분으로 구성된 입력 게이트를 통과한 값과 망각 게이트를 통과한 값이 덧셈 연산을 통해 저장된다. 이 값은 출력 게이트의 하이퍼볼릭 탄젠트 함수에 전달되어 단기상태 ht를 구성하고 이는 다음에 입력될 데이터의 메모리 블록(memory block)에 전달된다. 이러한 처리 방식은 학습 토큰의 연쇄 상에서 문맥을 통한 중요 정보만 선별하여 저장할 수 있기 때문에 시퀀스 정보 처리에 효과적이다.

해당 모델은 OOV에 의해 분석이 도출되지 않은 문장의 감성 분류를 수행하기 위해 사용되었으며, 이때 개체명 및 자질명의 경우, 문장의 주 서술어를 감성표현의 핵으로 가정한 의존 구문분석을 수행하여 페어링되었다.

6. 실험 및 성능평가

6.1 실험 데이터

LSTM 학습을 위한 데이터 수집에는 한국외국어대학교 DICORA 연구센터의 키워드 기반 사용자 후기글 크롤러 DECO R-Crawler가 사용되었으며, 네이버 쇼핑몰에 등록되어있는 화장품 리뷰 150만개의 텍스트를 메타정보와 함께

수집하였다. 이후 필터링 과정을 통해 추출된 31,000개의 정제된 텍스트에 리뷰 별점을 토대로 극성 레이블을 부여함에 따라 학습 데이터로 사용하였다. 전체 학습 데이터의 10%를 검증 데이터로 구성하여 학습 모델을 자체 평가한 결과, 90.6%의 정확도(accuracy)를 보였다.

테스트 데이터는 총 200개 문장 규모로 구성되었으며, 해당 문장에 작업자가 정답 태그를 수동으로 주석하는 방식 구성되었다. 평가문의 예시에는 하나의 문장 내에 많은 양의 개체명 표현들이 포함되어 있는 경우들을 비록하여 다단어 극성 표현들이 나타나는 경우들이 포함되었으며, 이에 따라 의존 파서를 활용한 개체명/자질명에 대한 평가에 적절한 문장들이 구성되었다. 다음을 보자.

(28) ㄱ. E[콜라겐 쿠션] 쓰다가 다시 F[커버력]이 S[좋은] E[라임 쿠션] 쓰게 되네요.
ㄴ. 기존에 E[나스] 쓰다가 노재팬으로 인해 E[더샘]으로 넘어왔는데 F[품질] S[좋아요].
ㄷ. E[라네즈 파운데이션] 나왔길래 주문했는데 S[촉촉한_F[보습력] 느낌은 없어요].

위 문장들을 통해 알 수 있듯이, 오피니언을 전달하는 문장 내에 단순 극성의 정보만이 처리될 것이 아니라, 개체명(E)/자질(F)/감성(S) 어휘 정보들을 복합적으로 처리하여 감성분석을 진행해야 비로소 리뷰 작성자의 의견을 정보성 있게 처리할 수 있다. 이해를 도모하기 위해 DECO-LGG에서 처리되는 대상 개체명/자질명/감성표현을 대괄호로 표현하였으며, 특히 (28ㄷ)에서는 도메인 의존성을 지닌 '촉촉하다'와 같은 극성 표현의 자질명가 문장내 명시적으로 드러나지 않은 '보습력'을 나타냄을 표현하였다. 해당 정보는 앞서 기술한 DECO-LGG를 통해 처리되며, DecoFESA 플랫폼의 처리 과정에서 효과적으로 활용된다.

6.2 실험을 위한 DECO-LGG 자원

실험 데이터의 평가에 직접적으로 적용된 LGG 언어자원은 <그림 10>과 같은 메인 그래프를 통해 호출 및 적용되도록 구성되었으며, 자원의 하위 구성은 총 190여개의 서브 그래프를 통해 구분되었다.

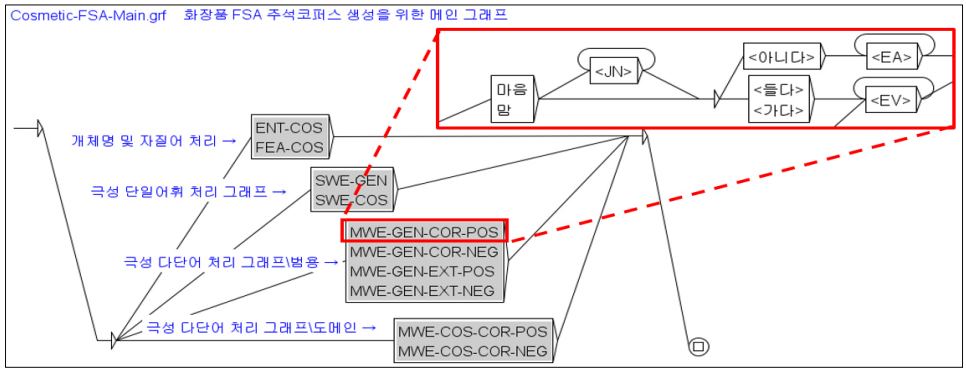

그림 10. 화장품 도메인 코퍼스의 주석을 위한 메인 LGG

본 연구에서는 총 700여개의 브랜드에서 나타나는 화장품 개체명 다단어 표현을 모두 패턴문법 그래프를 통해 인식할 수 있도록 구성하였으며, 총 60여개의 화장품 도메인 자질명과 그 변이형에 대한 형태들을 처리하고 정규화하도록 구성

하였다. 극성 어휘 처리와 관련해서 구축된 LGG는 13,000여 개의 단일어 극성 어휘와 2,700여개의 범용/도메인 다단어 극성 표현들을 모두 처리할 수 있도록 하였다. 이에 더하여 본 연구에서 구축한 자원은 극성 표현의 변이형 및 이와 공기하는 부정표현과 강화/약화부사들을 처리할 수 있게 디자인되었기 때문에, 실제 처리할 수 있는 패턴의 수는 단순 타입 수에 비해 폭발적으로 확장될 것으로 예상된다.

6.3 실험 결과

DecoFESA 시스템을 통해 처리된 성능 결과는 <표 3>과 같이 나타나는데, 극성 분류 과제에서는 부정문 처리에 0.80의 F1-score를 보인 반면, 긍정문 분류에서는 0.95의 준수한 성능을 보여주었으며, 평가 개체명/자질명 페어링 과제에서는 각각 0.845와 0.83의 정확도를 보였다. 여기서 분석에 실패하는 예를 보면 (29)와 같이 심각한 비정형성을 나타낸 문장에서 주로 관찰되었다. 비정형 텍스트 데이터에 대한 자질기반 감성분석을 진행할 때 이러한 문제를 해결하는 전처리의 중요성을 다시 한번 확인할 수 있다.

표 3. DecoFESA 실험 결과

Polarity	Precision	Recall	F1-score
Negative	0.69	0.94	0.80
Positive	0.99	0.91	0.95

Pairing	Accuracy
개체명	0.845
자질명	0.83

(29) ㄱ. 국민착붙파데 두번사세요.
ㄴ. (‘ROOT’ → VPROOT_국민착붙파데두번사세요.)

7. DecoFESA 플랫폼과 감성분석 결과의 시각화

DecoFESA는 앞서 제시된 하이브리드 방식의 자질기반 감성분석 알고리즘을 일반 연구자들이 손쉽게 사용할 수 있도록 구성된 윈도우 어플리케이션(application)으로, 다음과 같은 UI가 기본 화면으로 제공된다.

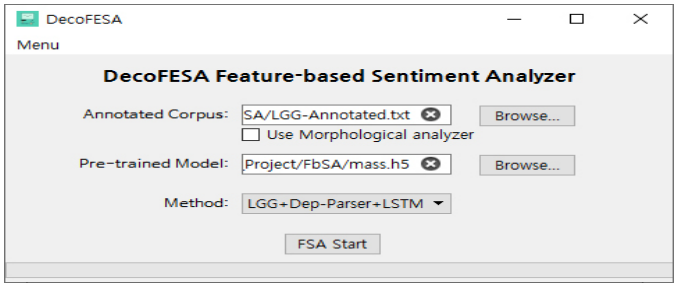

그림 11. DecoFESA 플랫폼의 기본 UI

DecoFESA 플랫폼은 기본적으로 두 가지 입력 값을 필요로 하는데, 그중 첫 번째는 위 플랫폼을 통해 자질기반 감성분석을 수행할 텍스트 파일(.txt) 형태의 코퍼스이다. 이때 코퍼스는, 플랫폼의 전체 프로세스가 언어자원을 통한 의미처리를 중점으로 하므로, 자질기반 감성분석을 위한 개체명/자질명/감성표현 의미 정보 처리가 완료된 코퍼스여야 한다. 그렇지 않은 경우 올바른 감성분석 결과를 도출하기 어렵다. 두 번째 입력값은 OOV을 처리하기 위한 LSTM 학습 모델로, 해당 위치에 들어가는 파일은 모델 파일(.h5)로, 레이블 정보가 할당된 동명의 레이블 파일(.label)이 동일 디렉토리에 존재해야만 정상적으로 작동한다.

두 가지 파일을 모두 설정한 후에는 감성분석에 사용할 알고리즘을 선택해야 한다. 그중 하나는 본고의 논의에 따라 의존 구문분석의 과정을 통해 개체명/자질명을 감성표현과 페어링하는 {LGG+ Dep-Parser+LSTM} 모듈이며, 다른 하나는 의존 파싱 과정을 건너뛴 상태에서 LGG를 통해 의미처리된 개체명 및 자질명 중 오른쪽 문맥 언어의 특성을 통해 가장 마지막에 나타나는 개체명 및 자질명을 선택하는 {LGG+LSTM} 모듈이다. 두 가지 모듈은 모두 개체명/자질명/감성표현이 문장에 한 개씩 존재하는 경우에는 의존 파싱 과정을 뛰어넘어 감성분석 결과를 도출하지만, 오피니언 핵심 어휘들이 중복되는 경우 첫 번째 모듈을 사용해야만 의존 구문 분석을 수행하게 된다.

{LGG+Dep-Parser+LSTM} 모듈을 사용하여 감성분석을 수행한 결과는 <그림 12>에서와 같이 두 가지로 제시된다. 해당 결과물 중 의미 정보와 의존 구문분석 기반한 자질기반 분석 결과는 왼쪽과 같은 형태로 팝업되며 오른쪽에 제시된 처리 결과는 의존 파서의 디버깅(debugging)에 활용하기 위한 부가적인 처리 결과로, 자동 팝업되지 않는다.

	Target	Feature	Polarity
0	커버 파운	GEN	POSITIVE
1	컨실러	GEN	POSITIVE
2	파운데이션	성능	POSITIVE
3	마몽드	디자인	POSITIVE
4	쿠션	GEN	POSITIVE
5	쿠션	GEN	POSITIVE
6	마몽드	가격	POSITIVE
7	파운데이션	GEN	POSITIVE
8	파운데이션	커버력	POSITIVE
9	나스	커버력	POSITIVE
10	파운데이션	커버력', '지	POSITIVE

	형태소 분석본	파서 분석본	파서 부사 제거
0	모공/NNG+파운데이션/NNG+을	M0_NP_OBJ_모공파운데이션을_0_	M0_OBJ_모공파운데이션을_0_L2
1	개/NNG+이/JKS ㄴ/NNG+적/X	M4_NP_SBJ_개이_0_L4 M4_NP_AJT	M4_SBJ_개이_0_L4 M4_AJT_ㄴ적
2	파운데이션/NNG+성능/NNG 지	M1_NP_NMOD_파운데이션성능_0	M1_NMOD_파운데이션성능_0_L3
3	마몽드/NNG+는/JX 전혀/MAG	M2_NP_SBJ_마몽드는_0_L3 M2_AF	M2_SBJ_마몽드는_0_L3 M2_ADV_
4	저/NP+가/JKS 쓰/VV+어/EC+ㅂ	M1_NP_SBJ_저가_0_L3 M1_VP_MC	M1_SBJ_저가_0_L3 M1_MOD_써ㅂ
5	저/NP+랑/JKB 정말/MAG 잘/M	M0_NP_AJT_저랑_0_L2 M1_AP_AD	M1_ADV_정말_1_L2 M3_ADV_잘_
6	저/NP+의/JKG 피부톤/NNG+이	M1_NP_MOD_저의_0_L3 M1_NP_	M2_ROOT_맞고_2_L1 M7_SBJ_가
7	요/MM 파운데이션/NNG+은/JX	M1_DP_NMOD_요_0_L3 M1_NP_S	M1_NMOD_요_0_L3 M1_SBJ_파운
8	성분/NNG+도/JX 괜찮/VA+고/	M1_NP_SBJ_성분도_0_L3 M1_VP_	M1_SBJ_성분도_0_L3 M1_VMOD_
9	커버력하/NNG+이/VCP+면/EC	VNP_ROOT_커버력하이면_0_L1 M	M0_ROOT_커버력하이면_0_L1 M
10	워낙/MAG 유명템/NNG+이/VC	M5_AP_ADV_워낙_0_L5 M5_VNP_	M5_ADV_워낙_0_L5 M5_MOD_유

그림 12. 의존 구문 분석기를 활용한 자질기반 감성분석의 출력 결과

또한 현재 수행한 자질기반 감성분석 대상 텍스트에 대한 정답지를 입력하면, 이를 기반으로 현재 감성분석 성능평가를 정보 추출(information retrieval)을 통해 평가해주는 기능이 내장되어 있어 활용도가 높아지는 장점을 가진다.

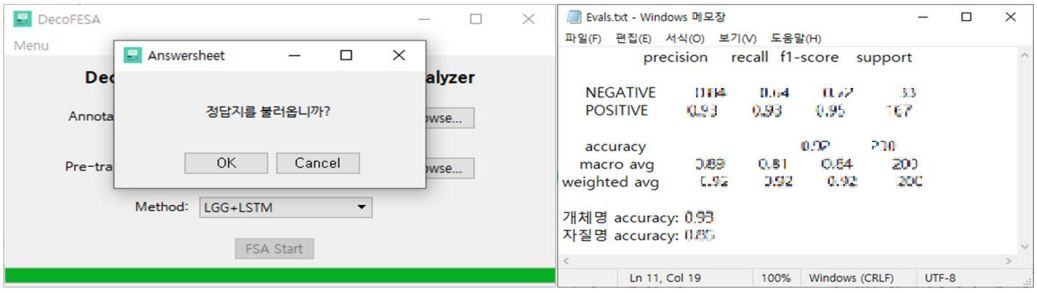

그림 13. 정답지를 기반으로 자질기반 감성분석 결과를 평가하기 위한 다이얼로그 및 평가 결과

감성분석의 경우 대용량의 오피니언 문서에 나타나는 감성 정보를 효과적으로 요약하는 시각화가 중요하다. DecoFESA 플랫폼에서는 평가대상에 따른 자질기반 극성 정보를 연결망 그래프(network graph)로 표상해주는 시각화 모듈을 제공한다. 해당 모듈은 시각화하고자 하는 대상 개체명의 개수를 비롯하여 각 노드 간 연결(edge) 길이 및 곡률과 그래픽 심볼의 기본 크기, 폰트 크기 등을 설정할 수 있는 파라미터(parameter)를 제공한다. 해당 설정값들에 따라 실제로 표상되는 자질기반 감성분석 결과의 연결망 그래프는 <그림 14>와 같다.

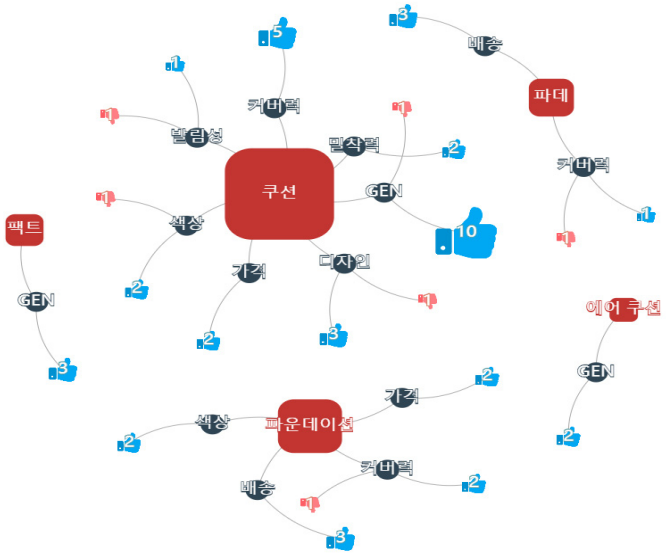

그림 14. 자질기반 감성분석 결과의 시각화

연결망 그래프를 통해 사용자 평가문에서 나타나는 평가 개체명/자질명/감성표현의 연결 관계가 위와 같이 시각화되기 때문에 다양한 개체명의 자질 범주에 대한 사용자의 평가를 효과적으로 살펴볼 수 있다.2)

8. 결 론

본 연구에서는 DECO-LGG 언어자원을 기반으로, 평가대상을 지칭하는 개체명 어휘 및 자질정보를 인식하는 의미 태깅을 자동으로 수행한 후, 이를 구문 분석 모듈과 연동시켜 오피니언 문장의 구문 의존 관계를 분석하는 과정을 진행하였다. 이를 통해 평가대상/자질/극성 정보를 처리하고, OOV가 발생한 문장의 경우 LSTM 모델을 활용하는 하이브리드 방식의 자질기반 감성분석을 진행하는 DecoFESA 감성분석 플랫폼을 구현하였다.

본고에서 제안하는 접근법은 점진적이고 명시적인 부트스트랩핑 방식의 처리시스템의 장점을 지님과 동시에, 딥러닝 기술을 활용하여 범용적 일반화 처리의 강점을 보여준다. 또한, 이를 연결망 그래프의 형식으로 시각화하여, 효과적인 감성분석 요약에 이르는 통합적 감성분석 시스템인 DecoFESA로 플랫폼화하는 실용성을 겸비하였다. 향후 화장품 도메인뿐 아니라 다른 도메인에 대한 효과적인 자질기반 감성분석을 위해, 도메인 의존적인 단일어/다단어 차원의 다양한

2) 평가 개체명 및 감성표현이 존재하나 자질명이 명시적/암시적으로 나타나지 않는 평가문의 경우, 가상의 자질 “GEN”을 설정함에 따라 이를 일괄적으로 처리할 수 있도록 정규화 하였다.

의미 정보를 기술하는 언어자원을 구축하고 이를 기계학습의 학습 데이터 및 감성분석 처리에 활용하는 연구에 활용될 수 있을 것으로 기대된다.

참고문헌

김한샘. 2003. 자연언어처리를 위한 관용표현 연구. 『한국어 의미학』 13, 43-68.

남지순. 2012. 오피니언 극성을 전환하는 한국어 부정표현 자동 인식을 위한 연구. 『언어와언어학』 57, 61-94.

남지순. 2018. 『코퍼스 분석을 위한 한국어 전자사전 구축방법론』. 서울: 도서출판 역락.

손진광. 2017. RNN LSTM과 ACO를 이용한 감성분석을 통한 컨텐츠 추천 시스템에 관한 연구. 『한국정보과학회 학술발표논문집』, 1033-1035.

최석재. 2012. 어휘의 부류와 감정 표현 관용구의 의미. 『한국어학』 55, 367-395.

최성용 · 남지순. 2018. 소셜 미디어 텍스트의 의존명사 내포 비정규토큰의 부분문법그래프 패턴 사전 구축. 『한국사전학』 32, 114-150.

함영균 · 임경태 · 박정열 · 윤용운 · 최기선. 2012. Linked Data를 위한 세종 품사태그셋 온톨로지 구축. 『한국정보과학회 학술발표논문집』, 109-111.

Alghunaim, A., M. Mohtarami, C. Cyphers, and J. Glass. 2015. A Vector Space Approach for Aspect Based Sentiment Analysis. *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 116-122.

Choi, D., J. Park, and K. Choi. 2012. Korean Treebank Transformation for Parser Training. *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, 78-88.

Garcia-Pablos, A., M. Cuadros, and G. Rigau. 2018. W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications* 91, 127-137.

Gross, M. 1997. *The Construction of Local Grammars: Finite-State Language Processing*. Cambridge: The MIT Press.

Gross, M. 1999. Nouvelles Applications des Graphes d'Automates Finis a la Description Linguistique. *Lingvisticae Investigationes* 22, 249-262.

Hai, Z., K. Chang, and J. Kim. 2011. Implicit Feature Identification via Co-occurrence Association Rule Mining. *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text processing* 6608, 393-404.

He, R., W. S. Lee., H. T. Ng, and D. Dahlmeier. 2018. Effective Attention Modeling for Aspect-Level Sentiment Classification. *Proceedings of the 27th International Conference on Computational Linguistics*, 1121-1131.

Hochreiter. S, and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9.8, 1735-1780.

Hu, M. and B. Liu. 2004. Mining Opinion Features in Customer Reviews. *Proceedings of the 19th National Conference on Artifical Intelligence*, 755-760.

Huang, J., O. Etzioni., F. Zettlemoyer, K. Clark, and C. Lee. 2012. RevMiner: An Extractive Interface for Navigating Reviews on a Smartphone. *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, 3-12.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, B. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotion*. Cambridge: Cambridge University Press.

Mirtalaie, M. A., O. K. Hussain, E. Chang, and F. K. Hussain. 2018. Extracting Sentiment Knowledge from pros/cons Product Reviews: Discovering Features along with the Polarity Strength of Their Associated Opinions. *Expert Systems with Applications* 114, 267-288.

Nam, J.-S. 2014. A Nouvel Dichotomy of the Korean Adverb Nemwu in Opinion Classification. *Studies in Language* 38.1, John Benjamins Publishing Company, 169-207.

Paumier, S. 2003. *De la Reconnaissance des Formes Linguistiques à l'Analyse Syntaxique*. Ph.D. Thesis, France: Université de Marne-la-Vallée.

Piao, S., P. Rayson., D. Archer, A. Wilson, and T. McEnery. 2003. Extracting Multiword Expressions with a Semantic Tagger. *Proceedings of the ACL Workshop on MWEs*, 49-56.

Samha, A. K., Y. Li, and J. Zhang. 2015. Aspect-Based Opinion Mining from Product Reviews Using Conditional Random Fields. *Proceedings of the 13th Australasian Data Mining Conference*, 119-128.

Song, M., H. Park, and K. Shin. 2019. Attention-based Long Short-Term Memory Network Using Sentiment Lexicon Embedding for Aspect-level Sentiment Analysis in Korean. *Information Processing and Management* 56.3, 637-653.

Thelwall, M., K. Buckley, and G. Paltoglou. 2012. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology* 63.1, 163-173.

Xue, W. and T. Li. 2018. Aspect Based Sentiment Analysis with Gated Convolutional Networks. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2514-2523.

Zhao, Y., B. Qin, S. Hu, and T. Liu. 2010. Generalizing Syntactic Structures for Product Attribute Candidate Extraction. *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 377-380.

황창회(제 1저자), 대학원생
경기도 용인시 처인구 모현 외대로 81
한국외국어대학교 언어인지과학과
E-mail: hch8357@naver.com

유광훈(제 2저자), 대학원생
경기도 용인시 처인구 모현 외대로 81
한국외국어대학교 언어인지과학과
E-mail: rhkdgns2008@naver.com

남지순(교신저자), 교수
경기도 용인시 처인구 모현 외대로 81
한국외국어대학교 언어인지과학과
E-mail: jeesun.nam@gmail.com