FeSAD

Jeong-Woo Yoon, Jee-Sun Nam

The Journal of Studies in Language

The Journal of Studies in Language 37(3), 335-358 (2021)

자질기반 감성분석 모델 학습을 위한 주석데이터셋 FeSAD 구축 방법론

윤정우* · 남지순**

한국외국어대학교

A Methodology of Constructing Sentiment-Annotated Datasets for Training a FbSA model

Yoon, Jeong-woo* and Nam, Jee-sun**

Hankuk University of Foreign Studies

*First Author / **Corresponding Author

https://doi.org/10.18627/jslg.37.3.202111.335

pISSN : 1225-4770
eISSN : 2671-6151

Received: October 11, 2021

Revised: November 05, 2021

Accepted: November 15, 2021

ABSTRACT

The Journal of Studies in Language 37.3, 335-358. This study introduces the process of constructing FeSAD (Feature-Sentiment-Annotated Dataset) based on the Semi-automatic Propagation methodology (SSP), which aims for Korean FbSA (Feature-based Sentiment Analysis). FeSAD was constructed by a 2-step annotation process: the SSP methodology is applied first, and then human annotators revise the annotated results. The linguistic resources for SSP consist of the LGG (Local Grammar Graph) patterns and the DECO (Dictionnaire Electronique du COreen) Korean machine-readable dictionaries. In this study, we evaluated the performance of the SSP-based approach with Cosmetics and Food domain texts. The result shows 0.93 and 0.90 F1-score for each domain, which reveals that the SSP-based approach could reduce the amount of the human annotators' task by under 10%. **(Hankuk University of Foreign Studies)**

Keywords: Feature-Based Sentiment Analysis, Feature-Sentiment-Annotated Dataset, Semi-automatic Symbolic Propagation, Local Grammar Graph, DECO dictionary

1. 서 론

본 연구는 기계학습 기반의 한국어 '자질기반 감성분석(FbSA: Feature- based Sentiment Analysis)'의 효과적인 처리를 위한 감성주석 학습데이터셋 FeSAD (Feature-Sentiment-Annotated Dataset)의 구축 방법론을 소개하는 것을 목표로 수행되었다.

- 본 연구는 HCLT-2021 학술대회에서 구두로 발표된 내용을 토대로 확장 · 보완된 연구로서, 2021년도 한국외국어대학교 교내연구지원 프로그램에 의해 수행되었다.

FbSA 학습 데이터 구축 방안은 크게 두 가지 측면에서 고려되어야 한다. 첫째는 '대상 텍스트'의 문제이다. 사용자들의 오피니언이 실현되는 가장 중요한 자연언어 텍스트 원천으로서 '소셜미디어(social media) 텍스트'는 실제로 그 도메인과 플랫폼에 따라 텍스트의 특징이 서로 차이를 보이기 때문이다. 가령 상품후기글 플랫폼에 업로드되는 텍스트 유형과 트위터나 신문기사 댓글 플랫폼에 업로드되는 텍스트 유형들은 그 형식과 내용면에 있어서 많은 차이를 보인다. 일반적으로 소셜미디어 텍스트로 총칭하는 다양한 문서들에 대해서 이와 같은 도메인, 플랫폼별 차이점을 고려하는 문제는, 실제로 어떠한 유형의 텍스트를 학습데이터로설정하는가를 결정하는 요인이 되므로, 이는 자질기반 감성분석용 학습데이터를 구축하는 데에 중요한 이슈가 된다. FbSA용 학습데이터를 구축할 때 고려해야 할 두번째 문제는 '구축 방법'에 대한 것이다. 문장 전체에 대한 긍정/부정의 극성 판별을 통해 정보를 획득하는 '문장층위의 감성분석(Sentence-level Sentiment Analysis)'에서는 사람들의 평점이나 별점 정보 등을 통해 대량의 학습데이터를 획득하는 것이 가능하나, 자질층위의 정보가 부착되어 있는 학습데이터는 자동으로 수집하는 것이 용이하지 않기 때문에 크라우드소싱과 같은 방법을 통한 직접 수동 구축이 수반되어야 한다. 다만 이 과정은 시간과 비용적 측면에서뿐 아니라, 문장의 개별 요소들에 대한 언어학적 의미-형식적 주석을 수행할 수 있는 전문성이 확보되어야 한다는 점에서 한층 더 어려운 작업이 된다. 이와 같은 어려움으로 인해, 상대적으로 발달해 있는 영어 데이터셋과 달리, 한국어와 같은 개별 언어들의 경우, FbSA를 위한 감성주석 데이터셋이 본격적으로 개발되어 있지 않은 상황이다.

본 연구에서 제안하는 감성주석 데이터셋 FeSAD는 바로 위의 두 가지 관점에 대한 성찰에서 출발하였다. 우선 '대상 텍스트' 구성 측면에서 다양성을 확보하기 위해 여러 유형의 플랫폼들과 도메인의 텍스트를 대상으로 설정하였다. 둘째로 자질기반 감성주석을 수작업으로 수행하는 접근법의 비효율성을 극복하기 위해서, DECO 한국어 전자사전(남지순, 2018)과 LGG (Local-Grammar Graph) 프레임(Gross, 1997)에 기반한 '반자동 언어데이터 증강(SSP: Semi-automatic Symbolic Propagation)' 방식(남지순, 2021)을 활용하였다.

SSP 방식에 사용되는 DECO-LGG 언어 자원은 자원의 유지 및 보수를 위한 효율성과 확장 가능성을 고려하여, 도메인간 공유되는 범용 언어자원과 도메인 특화 언어자원으로 모듈화하여 구성되었다. 도메인 특화 언어자원은 <그림 1>과 같이 현재 본 연구에서 제안하는 7가지 도메인별로 각각 구성되어, 최종적으로 이러한 언어자원들의 조합을 통해 FeSAD 데이터셋이 구축되었다.

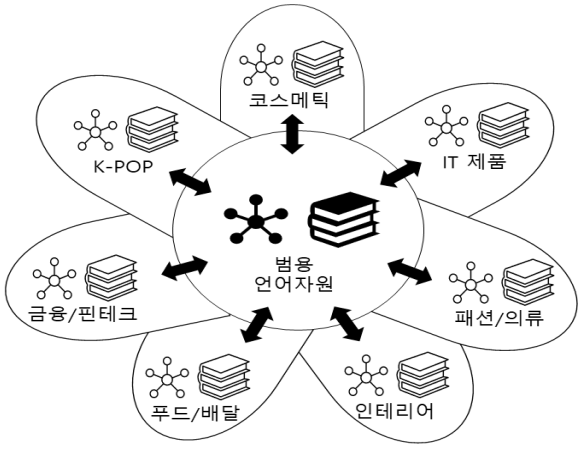

그림 1. 모듈화되어 있는 SSP 언어자원의 구조

2장에서는 감성주석 코퍼스와 FbSA와 관련된 선행 연구들에 대해 논의하고, 3장에서는 도메인, 플랫폼의 차이에서 기인하는 소셜미디어 텍스트의 차이와 FeSAD 데이터셋의 특징에 대해 논의한다. 4장에서는 본 연구에서 제안하는 FeSAD의 주석체계에 대해 소개한 후, 5장에서는 SSP를 위한 DECO-LGG 언어 자원과 SSP 적용 방법론에 대해 논의한다. 6장에서는 SSP 접근법의 성능을 실험하며, 7장에서는 본 연구의 의의 및 향후 연구 방향에 대해 논의한다.

2. 관련 연구

2.1 감성 사전 관련 연구

감성분석과 관련된 연구는 다양한 방법론을 통해 제안되었으나, 초기 국내 감성분석 연구는 감성사전 기반 감성분석이 주를 이루었다. 감성 사전을 통한 접근법은 분석 대상 텍스트의 긍/부정 극성 정보를 어휘 단위로부터 유추할 수 있으므로, 이를 통해 문장, 문서와 같은 오피니언 텍스트의 감성분석을 수행할 수 있다는 점에서 강점이 있으며, 여전히 다양한 연구가 이루어지고 있다.

감성 사전과 관련된 연구는 사전의 적용 범위에 따라 도메인 특화 감성 사전과 범용 감성 사전 연구로 나뉠 수 있는데, 범용 감성 사전에 대한 연구로, 신동혁 외(2016)에서는 영어 감성 어휘망(SentiWordNet)과 DECO 한국어 전자사전을 활용한 한국어 감성 사전 DecoSelex 구축 방법론을 제시하였다. 박상민 외(2018)에서는 기계학습 기법 중 하나인 Bi-LSTM을 통해 한국어 표준국어대사전의 뜻풀이를 분석하여 구축된 범용 감성 사전 KNU 한국어 감성사전을 소개하였다. 이와 같은 범용 감성 사전은 대규모의 언어 자원에 기반하여 구축되었으며 이에 따라 효과적인 사전 구축을 위해 기계학습 기법이 사용되기도 하였으나 학습데이터에 대한 주석 작업이 요구되는 단점이 있으며, 범용적인 극성 표현을 주로 다루기 때문에 특정 도메인에 한정되어 나타나는 도메인 의존 극성에 대한 처리에 있어 한계를 보인다.

도메인 특화 감성 사전에 대한 연구로, 이상진과 조은경(2020)에서는 도서 리뷰 도메인에 대하여 형용사, 부사 감성 어휘의 구축과 Word2Vec 기법을 통한 사전 확장 방안을 제안하였고, 조수지 외(2021)에서는 기업 애널리스트 보고서를 바탕으로 구축된 재무 분석용 감성 사전 KOSELF을 소개하였다. 황창회 외(2018)에서는 다단어 감성 표현(MWE: Multi-Word Expression)에 대해 범용, 도메인 감성 표현으로 구분하여 코스메틱(Cosmetic) 도메인에 대한 LGG 형식의 감성 언어자원 구축 방법론을 제시하였다. 이와 같은 도메인 감성 사전은 실제 도메인 데이터에 대한 분석 작업과 도메인 지식이 요구된다는 점에서 특정 도메인에 한정되어 연구가 진행되었으며, 이 경우 통일된 스키마의 다양한 도메인을 포괄하지 못한다는 점에서 한계가 지적될 수 있다.

실제로 사전 내 등재되는 감성 표현은 데이터의 도메인에 따라 다양한 언어적 양상을 보인다. 도메인 텍스트별 감성 표현의 언어적 특징에 대한 연구들도 진행되었는데, Lei와 Liu(2011)에서는 'Within a month, a valley formed in the middle of the mattress'의 valley와 같이 특정 도메인에서 나타나는 긍정/부정적 극성을 갖는 사실 표현들을 분석하기 위해 긍정/부정 극성어의 수식가능성을 활용한 방법론을 제시하였다. Ahn et al.(2012)에서는 5개 도메인(코스메틱, 호텔, 병원, 휴대폰, 영화)에서 나타나는 감성 표현들 중, 문맥에 관계 없이 특정 극성을 갖는 절대 극성 술어와 문맥에 따라 극성이 달라지는 상대 극성 술어 유형을 정리하고, 일부 상대 극성 술어에 대해 자질과의 공기 관계에 따른 극성 변화를 확인하였다. 이준환 외(2013)에서는 맛을 표현하는 형용사들에 대한 설문 조사를 통해 유사도를 측정하고 이를 군집화 및

계층화를 통해 비슷한 유형의 맛 평가 형용사들을 분류하였으며, '맛있다'와 '맛없다'와 같은 극성을 갖는 표현들과의 연관성을 통해 선호하는 맛과 비선호하는 맛에 대한 표현들을 확인하였다.

2.2 감성 주석 데이터 관련 연구

감성분석 연구를 위한 또다른 접근 방법으로 감성 주석 데이터에 기반한 기계학습 기반의 방법론을 볼 수 있다. 이 경우 대량의 주석 데이터가 요구되기 때문에, 이에 따라 감성주석 데이터의 중요성이 더 강조되었다. 영어 데이터의 경우, 영문 텍스트에 주관성 표지와 의견 및 감정표현과 관련한 다양한 주석들을 제안한 MPQA (Multi Perspective Question Answering) 코퍼스가 있다(Wiebe et al., 2005). 한국어의 경우, 한국어의 언어적 특징을 반영하여 MPQA 코퍼스 형식으로 구축한 KOSAC (김문형 외, 2013)이 제안되었다. MPQA와 KOSAC 데이터는 뉴스 기사글과 같은 정형문을 바탕으로 구축되었기 때문에 사용자 생성문이 갖는 비정형성을 고려하지 못하며, 데이터가 포함하는 도메인의 범위가 한정적이라는 한계를 갖는다.

사용자 생성문에 대한 감성 분석 데이터로는 SemEval-16, 17에서 사용된 트위터(Twitter) 감성분석 평가 데이터와 트위터 기반 감성 주석 코퍼스인 MUSE 코퍼스 중 SESAC (Sentence-level Sentiment Annotated Corpora)이 있다(조동희 외, 2016). 트위터는 플랫폼 내에서 키워드 기반 검색을 제공하며, 이를 통해 다양한 분야의 데이터 수집이 가능하다. 그러나 키워드 검색 기반으로 수집된 코퍼스는 특정 어휘를 필수적으로 포함하여야 한다는 점에서 수집된 도메인별 전체적인 어휘 양상을 담기 어렵다는 한계가 있다.

자질-감성 주석 데이터셋에 대한 연구로는 SemEval-14, 15, 16에서 FbSA 모델의 학습 및 평가 데이터로 활용된 FbSA 데이터를 대표적으로 들 수 있다(Pontiki et al., 2014). 이 데이터는 레스토랑과 노트북 후기글에 대한 자질-감성 정보가 주석된 데이터이다. 또 다른 영어 FbSA 데이터로는 Ray et al.(2021)에서 활용된 호텔 리뷰 데이터가 있다. 해당 데이터셋은 호텔 후기글을 '청결', '서비스' 등의 평가 자질을 기준으로 분류하고, 각 후기글별 부여된 사용자의 별점을 활용하여 자질-감성 주석 데이터를 생성하였다. 국내의 경우, 자질 정보가 문장 내 직접 주석된 데이터로는 조동희 외 (2016)에서 제안한 MUSE 코퍼스 중 TOSAC (Token-level Sentiment Annotated Corpora) 데이터가 있다. 해당 데이터는 문장 내 평가 대상, 자질, 감성 정보 주석을 XML 형식으로 구조화하였다. Hyun et al.(2020)에서는 자동차 도메인에 대한 영어, 한국어 자질-감성 주석 데이터를 구축하였는데, 평가의 대상을 자동차 생산업체로, 자질을 해당 생산업체의 대표 생산 모델로 설정하였다. 각 사용자 후기 문장을 자질을 기준으로 분류한 후, 사용자들로 하여금 각 문장 별 감성분류를 수행하도록 하였다.

이상에서 살펴본 기존의 감성 주석 데이터와 언어자원 연구에서는 특정 도메인 및 플랫폼에 한정되었으며 자질명 및 감성 평가에 대해 단순 분류 주석 체계가 제안되었다. 본 연구에서는 다양한 도메인, 플랫폼에서 수집된 데이터를 활용하였으며, 단순 분류 정보에 대한 주석뿐 아니라 텍스트 내에서 나타나는 구체적인 표현 정보를 FbSA 주석 체계에 포함하였으며, 이를 반자동으로 증강하는 방법론에 입각하여, 효율성과 정확성을 확보한 대규모의 자질-감성 주석 데이터셋을 구축하는 프로세싱을 제안하였다.

3. 주석 데이터셋 구축을 위한 플랫폼 및 도메인별 텍스트 특징 분석

3.1 소셜미디어 플랫폼에 따른 텍스트별 특징 분석

소셜 미디어 데이터는 다양한 플랫폼에서 생성되는데, 플랫폼에 따라 생성되는 텍스트의 형식적, 그리고 의미적 측면에서의 특성이 서로 차이를 보인다.

트위터(Twitter)의 경우, 140자 글자 수 제한[1]으로 인해 짧은 길이의 텍스트가 실시간으로 생산되며, 띄어쓰기 오류나 줄임말이 빈번하게 나타난다. 또한 뉴스 댓글이나 상품 후기글과 달리, 사용자들 간에 공유되는 뉴스 기사나 상품 정보가 없어, 작성자가 직접 글과 관련된 컨텐츠에 대한 링크를 공유하기도 한다. 이와 같은 트위터 플랫폼의 특징에 의해, 감성 분석을 수행할 때 맞춤법 교정과 링크 같은 노이즈 제거, 정규화 등과 같은 일련의 전처리 단계가 요구된다. 아래 예문은 트위터 텍스트의 사례를 보인다.

(1) ㄱ. LIE를 올리시고 안녕히 주무시라고 트루럽님이 말씀하셨다. ㄹㅏㅏㅏㅏㅏㅇㅣ 거짓말..... 장난없어.. ㅠ 내가 하성운 4K **과질**로 끊김없이 볼라고 **컴터** 빵빵하게 맞췄다구 ㅠ^ㅠ 하성운 HASUNGWOON Lie 라이 직캠 211002 FOREST& 부산/Truelv https://youtu.be/9gF3vjiA5bQ 출처 @YouTube

ㄴ. 이 스튜디오의 이 스타일 **넘나많은** 연예인이 해버려서 **뭐그렇게됫지만** 어쨋든 투바투가 **이런느낌으로** 하면 진짜 **너무좋을거같아** ㅜ

(1ㄱ)의 '과질'은 '고화질'의 줄임말이며, 컴터는 컴퓨터의 줄임말이다. 또한 관련 내용을 언급하기 위해 유튜브 영상 링크가 사용된 것을 확인할 수 있다. 그 외 'ㅠ^ㅠ'와 같이 모음, 특수 문자를 이용한 이모티콘과 'ㄹㅏㅏㅏㅏㅏㅇㅣ'와 같이 모음이 연속적으로 삽입된 표현들도 나타나며, 이러한 표현들에 대한 정규화 처리가 요구된다. (1ㄴ)에서도 '너무나'의 줄임말인 '넘나'나, '넘나 많은', '뭐그렇게됫지만'와 같은 띄어쓰기의 오류 현상을 찾아볼 수 있다.

반면, 뉴스 기사 플랫폼에 나타난 사용자 댓글은 트위터에 비해 상대적으로 느슨한 글자 수 제한으로 인해, 상대적으로 긴 글이 작성되며 맞춤법, 띄어쓰기 오류와 같은 노이즈 또한 상대적으로 매우 한정적으로 나타나는 것을 관찰할 수 있다. 또한 내용적 측면에 있어서도, 조수선(2007)에서 언급한 바와 같이 기사와 관련없는 개인의 넋두리, 홍보성 글, 주제를 벗어난 인신공격과 같은 유형의 글들이 많은 비중을 차지할 수 있다. 이와 같은 유형의 글들은 감성 분석에서 다루고자 하는 대상에서 벗어난다는 점에서, 트위터와는 다른 유형의 '의미적' 노이즈가 발생하는 것을 확인할 수 있다. 또한 뉴스 기사 플랫폼에서는, 댓글에 다는 댓글인 대댓글 기능을 제공하며, 대댓글은 댓글에 대한 작성자의 평가글이라는 점에서 별도의 처리가 필요함을 예측할 수 있다. 다음 예시를 살펴보자.

(2) ㄱ. 하늘만은 제발 건드리지말자.. 하늘에 드론과 자동차들 날라다니는거 생각만해도 끔찍하다

ㄴ. 이게 왜 끔찍하지?ㅋㅋㅋㅋㅋㅋ

1) 한글 음절의 경우 140자 제한이나, 영어나 특수 문자가 같이 쓰인 경우 140자 이상의 입력이 가능하다.

(2ㄱ)은 한 기사글에 대한 댓글이며, (2ㄴ)은 (2ㄱ)에 대한 대댓글이다. (2ㄱ)은 기사에 등장한 정책에 대한 부정적 평가글인 반면, (2ㄴ)은 (2ㄱ)의 댓글 작성자에 대한 평가라는 점에서 일종의 메타 오피니언 유형이 된다.

상품 후기글이 업로드되는 플랫폼의 텍스트들을 보면, 트위터에 비해 글자수 제한이 느슨하여 긴 텍스트의 작성이 가능한 것을 볼 수 있다. 그러나 비슷한 내용의 짧은 텍스트가 반복적으로 등장하기도 하며, 글자 수는 많지만 단순 반복 표현들로 구성된 도배글 또한 빈번하게 관찰된다. 이러한 텍스트에 대한 필터링 과정 없이 주석 데이터를 구축하면 다양한 문장 유형에 대한 정보를 제공하지 못할 위험이 나타나기 때문에, 머신러닝을 위한 학습데이터로 구성하기 위해서는 이러한 반복 유형에 대한 일련의 전처리 과정이 필요하게 된다. (3ㄱ)과 (3ㄴ)은 이와 같은 유형의 짧은 패턴이 반복되는 도배성 텍스트의 예를 보인다.

(3) ㄱ. 맛있네요 ㅎㅎ/맛있어요~/맛있음!
ㄴ. 좋아요!좋아요!좋아요!좋아요!좋아요!좋아요!좋아요!

상품 후기글의 '내용적인' 측면에서 처리되어야 할 노이즈 유형을 판별하기 위해선 후기글의 구조에 대해 살펴볼 필요가 있다. 이도영(2021)에서는 마스크 상품 후기글의 내용 구성이 '구매동기-사용자 사용담-추천/재구매 여부'의 구조로 이루어짐을 언급한 바 있는데, 즉 후기글 구조에서 구매 동기와 관련된 부분이나 사용자 사용담 중 단순 정보를 제공하는 부분은 실제 상품에 대한 평가가 아니기에 감성 분석에서 제외되어야할 노이즈 유형에 해당된다. 가령 '친구 추천으로 샀는데, 주변 사람들이 좋다길래 샀어요'와 같이 긍정 표현이 사용되었더라도, 구매 동기에 내용을 표현하고 있거나, 작성자 자신이 상품에 대한 평가 주체가 아닌 경우, 별도의 처리 방식이 요구될 수 있다. 이러한 현상은 특히 후기글 플랫폼에서 나타나는 특징의 하나로 관찰된다.

3.2 대상 도메인의 토픽에 따른 텍스트 특징 분석

같은 플랫폼에서 작성된 데이터라 하더라도 평가하는 대상의 도메인, 즉 토픽(topic)이나 평가대상의 의미 카테고리의 유형에 따라, 거기서 실현되는 감성/오피니언의 표현이 다르게 나타난다. 가령 상품 후기글이 업로드되는 플랫폼에서 수집된 텍스트 유형이라 하더라도, 음식이나 의류, IT 제품에 대한 후기글의 오피니언 및 감성 표현은 각기 다른 양상을 보인다.

(4) ㄱ. 이 집 탕수육은 맛이 정말 끝내주네요
ㄴ. 전 여기 블라우스가 몸에 잘 맞고, 촉감도 부드러워서 좋아요.
ㄷ. 사용한 지 이틀만에 쇼트가 났네요 정말..

(4ㄱ)은 음식, (4ㄴ)은 의류, (4ㄷ)은 IT제품을 평가 대상으로 하는 상품 후기글의 예시이다. (4ㄱ)과 같이 음식에 대한 후기글에서는 음식의 맛에 대한 평가가 주로 나타나며, 따라서 여기서 사용되는 긍정적 감성표현은 (4ㄴ)과 같은 의류 후기글에 나타나는 긍정적 감성표현과 동일하지 않다. (4ㄴ)과 같은 의류 후기글에서는 옷의 사이즈와 핏, 재질 등이 평가

자질로 등장하므로, 이와 관련된 술어들이 감성 표현으로 나타난다. 또한 (4ㄷ)의 IT 제품 후기글에서 보듯이, 긍정/부정의 표현에 있어 '쇼트'와 같은 도메인 전문 용어가 사용되기도 한다.

이처럼 평가 대상의 도메인이 다른 경우, 평가에 사용되는 감성 술어가 서로 상이한 양상으로 나타나게 되며, 따라서 감성 분석을 위한 주석 데이터 구성에 있어 이와 같이 다양한 도메인을 고려하여 균형잡힌 접근을 진행하는 것이 필요하다.

3.3 FeSAD 데이터셋의 플랫폼/도메인별 데이터 구성

본 연구에서 자질기반 감성주석 데이터셋 FeSAD는 이상과 같은 관점을 고려하여 구축되었다. FeSAD는 한국외대 디코라연구센터(http://dicora.kr)의 감성주석코퍼스(SAC) 데이터셋 연구의 일환으로 진행되었다. 현재 DICORA센터에 구축되어 있는 FeSAD 이전의 감성주석코퍼스의 구성은 <표 1>과 같다.

표 1. DICORA 감성주석코퍼스 SAC 데이터셋 유형

번호	유형	명칭	플랫폼	도메인	규모(문장)
1	문장 주석	MUSAC	댓글/카페/블로그/쇼핑몰	쇼핑/교육/문화 등 19개	200,000
2		WESAC	트위터	정치/경제/사회/문화	50,000
3	자질 주석	TOSAC	댓글/카페/블로그/쇼핑몰	정치/IT상품/맛집/성형외과	25,000
4		LOSAC	카페/배달앱	코스메틱/IT	16,000

문장 단위의 감성주석이 부착되어 있는 MUSAC 데이터와 WESAC 데이터는 전체 약 25만 문장 규모로서, 그 일부는 조동희 외(2016)에서 MUSE 코퍼스로 소개된 바 있다. 여기에는 트위터뿐 아니라 정치댓글/상품후기/게임/스포츠/교육/관광 등 다양한 분야에 대한 극성(polarity) 분류가 '긍정/부정/중립/복합극성'의 형태로 부여되어 있다.

자질기반 주석데이터는 TOSAC과 LOSAC으로 구조화되어 있는데, TOSAC은 감성주석이 XML-TREE 방식으로 구조화되어 있는 반면, LOSAC은 XML-MERGE 방식으로 텍스트에 해당 주석이 삽입되는 방식으로 구조화되어 있다. TOSAC에서는 4개의 도메인에 대한 FbSA를 위한 언어정보가 20여개 유형으로 부착되어 있으며, LOSAC에서는 FbSA를 위한 개체명/자질명/감성표현 유형이 40여개 레이블 형식으로 부착되어 있다.

FeSAD는 위와 같은 유형의 감성주석 데이터 스키마를 토대로, <표 2>와 같은 방식으로 구성되었다.

표 2. FeSAD 자질기반 감성주석 데이터셋 구성

번호	상위분류	명칭	도메인	플랫폼	규모
1	LOSAC연계	COS	코스메틱	쇼핑몰후기	3,576문장
2		ITP	IT제품	쇼핑몰후기	3,581문장
3	의식주관련	CLO	의류/패션	쇼핑몰후기	7,202문장
4		FOO	푸드/배달	배달앱후기	7,200문장
5		HOU	인테리어	쇼핑몰후기	7,205문장
6	문화관련	FIN	금융/핀테크	핀테크앱후기	7,201문장
7		KPOP	K-POP	카페/트위터	7,216문장

현재 FeSAD 데이터셋을 구성하는 텍스트의 도메인은 모두 7가지이다. 이중 코스메틱(COS)과 IT제품(ITP)에 관련된 후기글 데이터셋은 앞서 LOSAC 데이터의 후속 버전으로 진행되었기에 상대적으로 작은 규모로 구성되었으며, 의식주 관련 후기글 텍스트는 의류/패션 분야의 상품후기글(CLO)과 푸드/배달음식 후기글(FOO), 그리고 인테리어제품 후기글(HOU)로서, 각 7천여 문장씩 구성되었다. '음식'과 '가구'는 국내 소비자들이 평가한 소비 중요도 분야별 순위에서 1, 2위를 유지해왔기 때문에 선정되었으며, '의류' 역시 별도의 온라인 쇼핑몰의 활성화와 더불어, 자주 소비되고 많은 리뷰 수집이 용이한 분야로 주목되고 있기에 선정되었다.

그 외, 유형(有形)의 상품이 아닌 무형(無形)의 서비스에 대한 평가글 감성 분석을 위해, 금융/핀테크 관련 텍스트로 플레이스토어(Playstore)의 토스, 카카오뱅크 등의 핀테크앱 후기글(FIN) 데이터를 포함하였고, K-POP 음악/그룹에 대한 텍스트로 트위터와 카페, 그리고 앞서 구축된 MUSAC 코퍼스에서 추출된 일부 텍스트를 중심으로 해당 데이터셋(KPOP)을 구성하였다. 이상에서 구성된 FeSAD 데이터셋의 전체 규모는 42,000여 문장이며, 도메인의 의미적 특징에 따라 관련 플랫폼이 결정되는 방식으로 진행되었다. FeSAD에서 사용된 실제 데이터의 예시는 아래와 같다.

(5) ㄱ. (COS): 보습력이 많이 아쉽네요
ㄴ. (CLO): 걱정했던 것과 달리 신축성 덕분에 잘 맞아요!
ㄷ. (FIN): 수수료 없이 이체하는거 쭉 해주세요♡♡

4. FbSA 데이터셋의 주석체계

4.1 3가지 오피니언 원소의 주석

본 연구에서 FbSA를 위해 주석되는 오피니언 원소는 Liu(2012)에서 정의한 오피니언 5원소쌍(Opinion Quintuple)에서 추출된 오피니언 트리플이다. 일반적으로 상품후기글에서 메타정보(meta-information)로 실현되는 유형인 '평가자(opinion holder)'와 '평가시간(opinion time)'을 제외한 다음의 3가지 성분이 중심이 된다.

(6) Opinion Triple: {e, f, s}

위에서 e는 '개체명(entity)' 부류로서, 도메인에 따라 그 의미적 특징과 어휘적 구성이 달라지는 열린 목록을 구성한다. 반면, f는 '자질명(feature)' (또는 속성명: aspect)으로서, e와는 달리 제한된 유형의 일반 명사구들로 실현된다. s는 '감성표현(sentiment)'으로서 긍정/부정 등의 극성어 표현과 일련의 극성전환장치들(PSD: Polarity-Shifting Device)(남지순, 2012; Nam, 2014)에 기반한 시퀀스들을 포함한다.

4.2 어휘 층위의 DEC 주석 체계: DEC-layer Annotation

단일어휘로 실현되는 오피니언 원소들은 해당 도메인과 무관한 '범용의 어휘 의미/감성 정보'를 담은 DECO 전자사전과 '도메인 특화된 의미/감성 정보'를 담은 DECO-DOM 도메인사전의 태그들을 참조하여 주석된다.

어휘 층위의 오피니언 트리플 주석(DEC-layer Annotation)은 DECO 사전에 등재된 개체명/자질명 및 감성어휘 분류 정보에 기초한다. 여기 등재된 11가지 개체명 분류 체계(EntLex)와 4가지 극성어휘 분류 체계(PolLex)를 이용하여 개체명과 극성어 태그가 부착되며, 자질명은 단일 태그를 갖는 자질명 분류체계(FeaLex)로 구성되어 있다. DECO 사전의 각 체계별 태그 형식은 <표 3>에서 보이는 바와 같다. DECO-DOM 사전의 주석 체계 역시 DECO 사전의 주석 체계와 태그 형식을 공유하며, 실제 사용된 태그들과 표제어의 예를 보면 <표 4>와 같다.

자질명과 감성 표현은 DECO 사전의 태그가 전체 참조되었지만, 개체명 태그는 일부만 참조될 수 있다. 이는 현재 구축된 7개 도메인에서 나타나는 평가대상 개체명에 대한 의미 분류가 <표 4>에서와 같이 일부 유형으로 한정되어 실현될 수 있기 때문이다. 반면, 일부 출현한 유형들의 경우 그 하위분류가 더 요구되는 경우들이 관찰되므로, 이런 경우 도메인 특화된 하위분류가 추가로 수행되었다. 가령 <표 5>는 개체명에 대해 도메인별 하위분류 정보가 추가적으로 주석된 예를 보인다.

단일어휘 층위의 DEC 주석에서는, 위에서 기술된 개체명, 자질명, 감성 표현 외에, 감성 술어의 극성을 변환시키거나 그 정도성에 영향을 미치는 일련의 성분들인 정도부사, 부정부사, 부정보조용언 등이 주석된다. 정도부사와 부정소의 주석 체계를 보이면 <표 6>에서와 같다.

표 3. DECO 사전 내 개체명, 자질명, 감성 표현 주석체계

주석 체계	대분류	소분류	태그
EntLex	인물	인명	XXPE
		직무/직업	XXHU
		조직명	XXOR
	공간	자연공간	XXGE
		인공공간	XXLO
	시간	명시적 시간	XXTI
		비명시적 시간	XXEV
	사물	구체물/이동 불가	XXCO
		구체물/이동 가능	XXTH
		구체물/상품명	XXPR
		추상물	XXCR
FeaLex	-	자질 명사	XQFT
PolLex	-	강한 긍정	QXSP
	-	긍정	QXPO
	-	강한 부정	QXSN
	-	부정	QXNG

표 4. DECO-DOM 도메인 사전 내 사용된 태그 및 예시

소분류	태그	예시
인명	XXPE	싸이, 이효리
조직명	XXOR	삼성, BBQ
구체물/상품명	XXPR	갤럭시, 치킨
추상물	XXCR	행복연금대출, Tell me
자질 명사	XQFT	맛, 사이즈
강한 긍정	QXSP	존맛, 존예이다
긍정	QXPO	촉촉하다, 직관적
강한 부정	QXSN	개빡치다, 폐기물급
부정	QXNG	금가다, 다운그레이드

표 5. DECO-DOM 사전내 기술된 개체명 분류와 그 하위분류 정보

태그	하위분류 태그	의미	예시
XXPE	NA	핵심 개체명	소녀시대
	ME	그룹 멤버명	태연
XXOR	BR	브랜드명	삼성
	CO	소속사명	JYP
XXPR	NA	핵심 개체명	갤럭시
	TY	상품 유형	S21
	RE	관련어	5G
XXCR	NA	핵심 개체명	행복연금대출
	SO	곡명	Tell me
	RE	관련어	뮤비

표 6. 통사적 구성 기술을 위한 정도부사 및 부정소 주석 체계

분류	태그	하위분류 태그	의미	예시
정도 부사	INT	AMP	상향 부사	아주
		DOW	하향 부사	덜
부정소	PSD	NADV	부정 부사	안
		NAUX	부정 보조용언	않다
		NEMU	극성도입어 {너무}	너무

현재 FeSAD 데이터셋에 해당 원소에 정보가 주석될 때, DEC 주석체계에서 오피니언 원소의 주석 태그들은 XML (Extended Markup Language)-MERGE 방식으로 다음과 같이 실현된다.

(7) <XXPR=ITP_BR>삼성</XXPR> <XXPR=ITP_NA>갤럭시</XXPR> <XXPR=ITP_TY>S21</XXPR>
<XQFT=ITP>디자인</XQFT>이 <INT=AMP>완전</INT> <QXPO=GEN>예쁘네요</QXPO>

주석은 <SeqType=value>, </SeqType>형식의 좌우 태그로 구성된다. SeqType에는 DECO 사전의 오피니언 트리플 원소 및 정도부사, 부정소 정보가 기술되며, value에는 태그의 분류에 따라 도메인 및 세부 정보가 기술된다. (7)의 예에서 개체명의 경우, 각 단일 어휘 성분별로 주석되어 있는 것을 볼 수 있으며, 이때 '<XXPR=ITP_BR>삼성</XXPR>'의 태그 예를 보면, '상품 개체명(XXPR)' 중 'ITP(ITP 제품 도메인)'의 '브랜드명(BR)'의 정보를 획득할 수 있다.

4.3 구/절 층위의 주석 체계: MID-layer Annotation

4.3.1 구/절 층위의 주석 필요성

실제 텍스트에서는, 앞서 논의한 오피니언 트리플 원소들이 반드시 하나의 단일어휘 형태로 실현되지 않는다. 이 경우, 구 또는 절 유형의 층위에서의 주석의 필요성이 대두되며, 이러한 주석 체계가 앞서 어휘 층위의 주석 체계와 구별되기 위해서는 별도의 태그 방식이 제시되어야 한다. 본 연구에서는 앞서 단일어휘 층위의 주석 체계(DEC-layer Annotation)과 구별되어, 구/절 층위의 주석 체계(MID-layer Annotation)로 명명된다.

개체명과 자질명의 경우, 단일 어휘 층위에서 주석된 대상들이 여러 개의 다단어(MWE: Multi-Word Expression)로 구성된 시퀀스로 등장할 때, 이들을 통합하여 하나의 개체명 구 및 자질명 구로 주석하는 것이 바람직하다. 구/절 층위에서의 개체명/자질명 주석 태그는, 각각 '<NE=도메인_NA>', '<FT=도메인>'과 같은 형식으로 주석된다(<표 7>). 반면, 감성 어휘의 다단어 연쇄 태그는 '<POL=극성값>' 형식으로 구성되며, 이때 내부 극성값은 {강한긍정(SP), 긍정(PO), 약한긍정(WP), 약한부정(WN), 부정(NG), 강한부정(SN)}의 6개 극성 분류 중 하나의 값을 표현하게 된다. 이 때, 6개 극성 분류는 기존 단일 어휘의 4개 극성 분류 체계에 '약한긍정(WP)'과 '약한부정(WN)'이 추가된 형태를 이룬다. <표 8>에서 보이는 바와 같다.

표 7. 구/절 층위 개체명, 자질명 주석 체계

분류	태그	내부정보 태그	예시
개체명	NE	NA	삼성 갤럭시 S21, 그램
자질명	FT	-	카메라 디자인, 디자인

표 8. 구/절 층위 감성 어휘 주석체계

분류	태그	내부정보 태그	예시
개체명	POL	SP, PO, WP, WN, NG, SN	너무 좋아요(SP), 예쁘진 않아요(WN)

이상의 방법을 통해 FeSAD 데이터셋에 MID 주석체계가 적용될 때, 위의 예시들에 대한 주석 결과의 예를 보이면 다음과 같다.

(8) <NE=ITP_NA>삼성 갤럭시 S21</NE> <FT=ITP>디자인</FT>이 <POL=WN>예쁘진 않아요</POL>

실제로 감성 표현의 다단어(MWE) 시퀀스는 다시 다음의 두 가지 유형으로 분류해서 고려할 수 있다. 첫째는 소위 관용표현, 연어, 복합어 등으로 정의될 수 있는 일련의 어휘적 구성체로서, '마음에 들다(=좋다)'와 같은 유형이 여기 해당하며, 둘째는 부정소, 또는 정도부사가 삽입된 구문 연쇄, 술어구 등으로 일련의 통사적 구성체인 '나쁘지 않다(→좋다)'와 같은 유형이 여기 해당한다. 전자의 경우는 단일어휘 층위의 주석체계에서와 같이 4가지 극성 유형으로 주석하는 것이 적절하나, 후자의 경우는 통사적 성분의 개입에 의해, <표 9>와 같이 6가지 유형으로 세분된 극성 점수를 할당하는 것이 바람직하다.

표 9. 구/절 층위 다단어 감성 표현 중 통사적 구성을 이루는 유형의 6가지 극성 점수

번호	통사적 결합 양상	극성 태그	의미	극성 점수
1	상향 부사 + 긍정 감성 표현	SP	강한 긍정	+1.5
2	긍정 감성 표현	PO	긍정	+1.0
3	부정소 + 부정 감성 표현	WP	약한 긍정	+0.5
4	부정소 + 긍정 감성 표현	WN	약한 부정	-0.5
5	부정 감성 표현	NG	부정	-1.0
6	상향 부사 + 부정 감성 표현	SN	강한 부정	-1.5

위의 두 가지 유형의 다단어 감성표현에 대해 다음에서 살펴보기로 한다.

4.3.2 다단어(MWE) 감성 표현의 주석체계

4.3.2.1 어휘적 특이성을 보이는 다단어 감성 표현

어휘적 특이성(Idiosyncrasy)을 보이는 다단어(MWE) 감성 표현은 어휘적 특이성에 기반한 연쇄로서, 여기서 어휘적 특이성이란 단일 어휘의 의미 조합만으로 그 복합체(구 또는 절)의 의미를 유추하기 어려운 성질을 말한다. 가령 '마음에 들다'와 같이 전체 복합 술어구는 긍정적 의미를 표상하지만, 그 내부 구성요소인 '마음'과 '들다'는 일련의 어휘적 제약관계를 보이고 있어, 가령 '마음'의 유의어인 '가슴'과 같은 어휘로의 치환이 불가능하며, '들다' 자체의 술어 의미만으로 '마음에 들다' 전체 시퀀스의 의미가 유추되기 어렵다는 특징을 보인다.

이와 같은 유형의 다단어 표현은 전체 '구'를 통해 의미가 도출되기에 사전에 단일 표제어 형태로 등재되기 어렵다. 또한 어휘적 특이성을 갖는 이러한 다단어 표현은 일반적인 단일 어휘와 마찬가지로 정도 부사 및 부정사와 결합하여 통사적 연쇄를 구성할 수 있기 때문에 더욱 복잡한 구조의 MWE를 구성할 수 있다. 이들은 구/절 층위에서'<POL=극성값></POL>'형식으로 주석된다.

본 연구에서는 이와 같은 유형의 다단어 감성 표현에 대해, 황창회 외(2018)에서 제안한 바와 같이, 'NN(체언-체언)' 결합형, 'NP(체언-술어)' 결합형, 'PP(술어-술어)' 결합형의 3가지 유형으로 분류하였다. NN 구성은 앞 체언이 뒷 체언을 수식하는 수식 관계나 뒷 체언에서 서술 명사가 등장하는 주술 관계를 이루는 체언들로 구성되며, NP 관계는 '주어-서술어' 관계, 혹은 '목적어-서술어' 관계를 이루는 체언과 술어의 조합들로 구성된다. PP 관계는 술어의 술어 수식이나, 부사에 의한 술어 수식 구문들, 그리고 NN, NP 구조에 포함되지 않는 관용 표현들을 포함한다. 이와 관련된 실제 예시를

보면 다음과 같다.

(9) ㄱ. 디자인이 <POL=PO>맘에 쏙 드넹</POL>
ㄴ. 역시 <POL=PO>국민 아이템</POL>!
ㄷ. <POL=PO>자주 듣고 있어용</POL>
ㄹ. 하... <POL=NG>엎친 데 덮친 격이네요</POL>...

(9ㄱ)은 NP 유형의 다단어 구성의 예를 보이며, (9ㄴ)은 명사 연쇄로 이루어진 NN 유형의 예를 보인다. (9ㄷ)은 PP 유형의 예를 보이며, (9ㄹ)은 일종의 관용적 표현으로 현재 PP 카테고리에서 함께 분류되어 있다.

4.3.2.2 통사적 구성을 이루는 다단어 감성 표현

통사적 구성을 이루는 다단어 감성 표현은, 부정소나 정도부사와 같은 성분을 포함하는 시퀀스 유형을 나타낸다. 이 경우, 6가지 극성 분류에 대해 0.5점 단위 점수가 적용되어, 전체 {+1.5 ~ -1.5점}의 극성 스코어가 부여된다.

(10) ㄱ. [+1.5] 디자인이 <POL=SP>엄청 예뻐요</POL>.
ㄴ. [+1.0] 디자인이 <POL=PO>예뻐요</POL>.
ㄷ. [+0.5] 디자인이 <POL=WP>촌스럽지 않아요</POL>.
ㄹ. [-0.5] 디자인이 <POL=WN>안 예뻐요</POL>.
ㅁ. [-1.0] 디자인이 <POL=NG>촌스러워요</POL>.
ㅂ. [-1.5] 디자인이 <POL=SN>엄청 촌스러워요</POL>.

위의 예에서 보는 바와 같이, (10ㄴ)의 '예쁘다'에 대해 '긍정(PO)' 술어로서 {+1점}을 부여한다면, 여기에 (10ㄱ)처럼 강화부사 '엄청'이 결합하는 경우, '강한긍정(SP)'이 되어 {+1.5점}이 된다. 반면 (10ㄹ)처럼 부정소 '안/않다'가 결합하면 극성이 전환되어 부정 극성 술어(=안 예쁘다)가 되는데, 이러한 {부정소+긍정} 결합형의 경우는 '약한부정(WN)'의 태그로 주석되어 {-0.5점}의 점수를 부여받는다.

같은 방식으로 (10ㅁ)의 '촌스럽다'는 '부정(NG)' 술어로서 {-1.0점}을 부여받고, 여기에 (10ㅂ)처럼 강화부사 '엄청'이 결합하면 '강한부정(SN)'이 되어 {-1.5점}이 된다. 반면 (10ㄷ)처럼 부정소 '안/않다'가 결합하면 극성이 전환되어, 긍정 극성 술어(=촌스럽지 않다)가 되는데, 이 경우에도 {부정소+부정} 결합형의 경우로 '약한긍정(WP)'의 태그로 주석되어 {+0.5점}의 점수를 부여받게 된다.

이상에서 오피니언 트리플에 대한 단일어/다단어 주석체계에 따라, 해당 요소들에 감성분석을 위한 태그 정보가 주석될 수 있다. 다음 장에서는 이와 같은 주석체계에 기반하여 본격적인 언어자원을 구축하고, 이를 기반으로 반자동 언어데이터 중강(SSP)을 수행하는 접근법에 대해 논의한다.

5. DECO-LGG 언어자원에 기반한 SSP 방법론

5.1 DECO 범용 전자사전과 DECO-DOM 도메인 사전

DECO 전자사전에는 30만여개의 표제어와 각 표제어별 형태, 통사, 의미, 감성 정보들이 명시되어 있다. 또한 각 표제어에는 활용클라스 정보가 부착되어 있어, 이를 통해 해당 활용어미 트랜스듀서가 호출되도록 설계되어 있다(남지순, 2018). DECO 사전은, 호환되는 Unitex 플랫폼(Paumier, 2003)을 통해 코퍼스 분석을 위해 곧바로 적용될 수 있는 기계가독형 사전으로서, 실제 텍스트에 실현되는 모든 표면형 어절들에 대한 올바른 형태소단위 분석이 가능하게 된다. DECO 사전 내에 기술되는 표제어의 예시는 다음과 같다.

(11) ㄱ. 가격.NS02+ZNZ+XQFT+…
ㄴ. 예쁘다.AS16+ZAZ+QXPO+…

DECO 사전 내 기술되는 표제어는 위와 같이 표제어와 관련 정보가 '.'으로 구분되며, 언어 정보들은 내부에서 다시 '+'로 연결된다. 표제어 뒤에서 첫 번째로 등장하는 정보인 'NS02', 'AS16'는 품사 정보와 활용 클라스 정보를 의미한다. 그 다음 연결된 'ZNZ'와'ZAZ'는 표제어의 하위 카테고리 정보를 의미하며, 'XQFT'(자질명), 'QXPO'(긍정극성)과 같이 앞서 4장에서 기술된 오피니언 트리플과 관련된 주석 정보 또한 내장되어 있다.

본 연구에서는 이러한 범용 사전과 더불어, 각 도메인별로 특화된 DECO-DOM 도메인사전을 구축하여 이를 함께 적용하는 방식을 사용하였다. 도메인 사전에는 도메인 특화된 자질표현, 감성표현과 함께 평가 대상이 되는 개체명이 기술된다. 이때, DECO 사전에 기술된 정보 외에 도메인 정보, 개체명 의미 세부 분류와 같은 추가적인 의미 정보가 기술된다.

(12) ㄱ. 피자헛.NS02+ZNW+DomInf=XFOO+DomEnt=XXOR+DomSub=BR+…
ㄴ. 충무김밥.NS02+ZNW+DomInf=XFOO+DomEnt=XXOR+DomSub=NA+…
ㄷ. 식감.NS02+ZNW+DomInf=XFOO+DomEnt=XQFT+…
ㄹ. 맹맹하다.AS24+ZAW+DomInf=XFOO+DomPol=QXNG+…

위의 예시는 푸드/배달음식 도메인 사전에 기록된 표제어 정보로, (12ㄱ)과 (12ㄴ)은 개체명, (12ㄷ)은 자질명, (12ㄹ)은 감성 표현의 표제어의 예를 보인다. DECO 사전과 마찬가지로 '표제어.의미정보'의 형식으로 기술되며, 의미정보의 처음 두 항목은 각각 품사형태 정보와 사전 정보를 나타낸다. 세 번째 항목부터는 DECO 사전과 다른 형식의 정보가 추가되는데, 도메인 특화 정보에 관련된 부분으로 '속성(Attribute)=값(value)' 형식으로 기술된다. DomInf 속성에는 도메인 정보가 기술되며, DomEnt에는 개체명 또는 자질명 태그가 부여된다. 개체명 태그에만 존재하는 DomSub 속성에는 4장에서 기술된 개체명별 하위분류 태그가 부착되며, (12ㄱ)에는 브랜드명(BR), (12ㄴ)에는 상품 개체명(NA) 태그가 부착되어 있는 것을 확인할 수 있다. (12ㄹ)에서는 DomPol 태그를 통해 감성 극성 정보가 부여되는 것을 확인할 수 있다.

5.2 다단어(MWE) 기술을 위한 LGG 패턴 문법

위에서 단일어에 대한 주석을 위해 DECO 사전의 표제어 태그 정보가 사용되었다면, 텍스트에서 실현되는 복합명사구, 복합동사구, 또는 부정소나 정도부사가 결합한 통사적 연쇄 등으로 실현되는 일련의 다단어(MWE) 연쇄에 대한 주석을 위해 LGG (Local-Grammar Graph) 프레임이 사용된다. LGG로 구축되는 다단어 패턴문법은 일련의 언어 시퀀스를 방향성 그래프 형식으로 기술하여, 이를 Unitex 플랫폼에서 유한상태 트랜스듀서(FST: Finite-State Transducer)로 자동 변환한 후, 텍스트 분석 및 주석에 적용하는 방식을 사용한다.

5.2.1 개체명/자질명 MWE를 표상하는 LGG 패턴문법

개체명 MWE의 경우, 가령 코스메틱 도메인에 실현되는 개체명을 보면, “이자녹스 UV 선크림”은 실제 텍스트에서 “UV 선크림, 이자녹스 선크림, 선크림” 등과 같이 여러 형태로 분리되어 실현될 수 있다. 따라서 현재 구성된 코스메틱 도메인의 도메인 사전(DECO-COS)에는 ‘이자녹스’, ‘UV’, ‘선크림’과 같이 각각 분리된 표제어들이 등재된다. 이러한 단일어 표제어들이 여러 조합에 의해 다단어 시퀀스를 구성하는 경우, 이들을 인식하고 인식된 전체 시퀀스에 대한 태그를 주석하기 위해 <그림 2>와 같은 형식의 LGG가 구축되었다.

<그림 2>는 코스메틱 도메인의 개체명 MWE를 주석하기 위한 LGG의 예를 보이는데, 여기서는 도메인사전의 태그 정보에 대한 참조식을 병렬적으로 연결하여 기술하고 있다. 실제로 참조된 태그 카테고리의 어휘 성분이 둘 이상 실현되는 구조를 인식하도록 하는 재귀 기술(recursive description)을 통해, ‘브랜드명(BR)+개체명(NA)’조합, ‘개체명(NA)+제품 유형(TY)’과 같은 다양한 패턴을 주석할 수 있도록 한다.

자질명 역시 실제 텍스트에서 단일 어휘 뿐만 아니라 MWE 형태로도 실현될 수 있다. 가령, 핀테크앱 도메인에서는 ‘로그인 화면, 로그인 속도’와 같이 두 개 이상의 성분으로 구성된 자질명 MWE가 나타나는데, 이와 같은 MWE 패턴을 주석하기 위해 <그림 3>과 같은 LGG가 구성되었다.

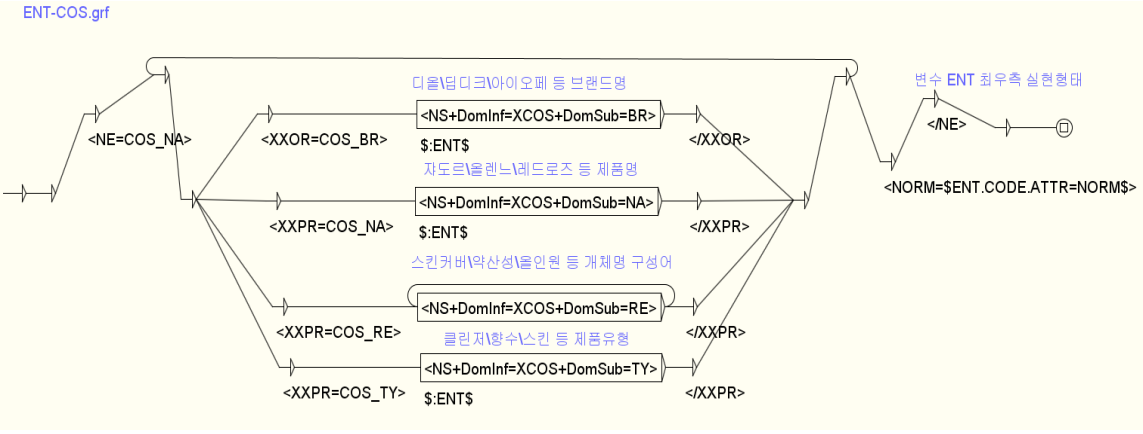

그림 2. 코스메틱 도메인의 개체명 MWE 주석을 위한 LGG

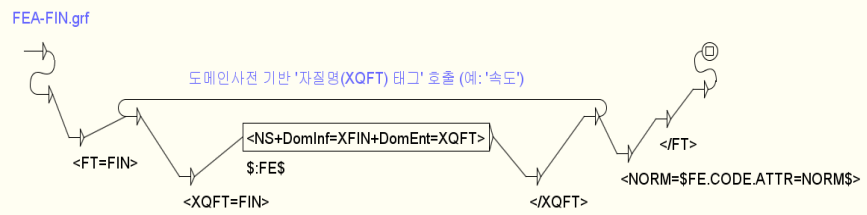

그림 3. 금융앱 도메인의 자질명 MWE를 주석하기 위한 LGG

5.2.2 감성 표현 MWE를 표상하는 LGG 패턴문법

5.2.2.1 어휘적 특이성을 보이는 감성 MWE의 LGG

어휘적 특이성을 보이는 감성 MWE는 4장에서 기술된 것처럼 내부 어휘들의 구문 구성 방식에 따라 3가지 유형으로 분류되며, 이와 같은 분류가 반영되어 LGG 프레임 내에서 <그림 4>와 같이 구조화된다.

그림 4. 어휘적 특이성을 보이는 감성 MWE의 구조화 예시

<그림 4>는 KPOP 도메인의 어휘적 특이성을 보이는 감성 다단어표현(MWE) 중 긍정 극성의 표현들을 호출하는 LGG로, {MWE-POP-PO-NN}, {MWE-POP-PO-NP}, {MWE-POP-PO-PP}는 각각 "NN(체언-체언)", "NP(체언-술어)", "PP(술어-술어)" 유형의 서브그래프들을 호출하는 경로를 보인다. <그림 5>은 위의 {MWE-POP-PO-NP}에서 호출하는 서브그래프의 일부 예를 보인다.

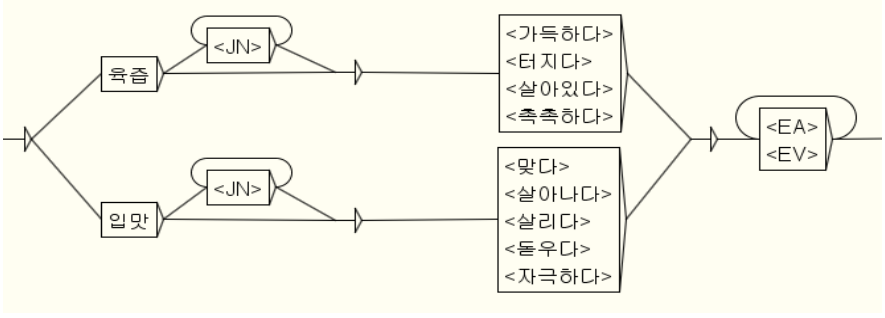

그림 5. 어휘적 특이성을 보이는 감성 MWE 중 '체언-술어'를 표상하는 서브그래프 일부 예

5.2.2.2 통사적 구성을 이루는 감성 MWE의 LGG

통사적 구성을 이루는 감성 MWE는, 앞서 논의한 바와 같이, 일련의 정도 부사 및 부정소와의 결합으로 이루어진다. <그림 6>는 이를 표상하는 LGG의 예를 보인다.

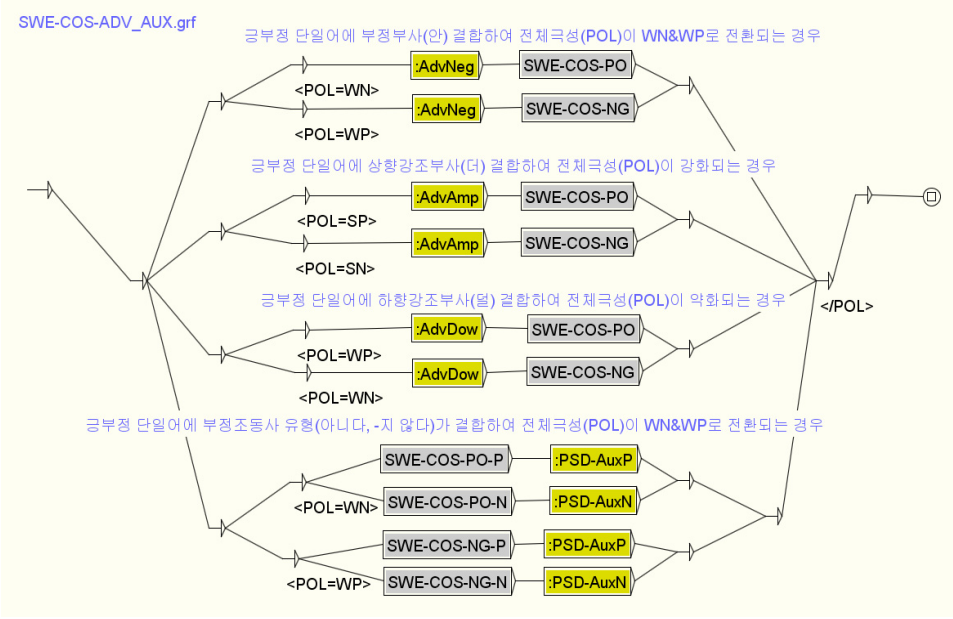

그림 6. 통사적 구성을 이루는 감성 MWE를 표상하는 LGG의 예시

<그림 6>에서 회색 박스로 기술된 {AdvNeg}, {AdvAmp}, {AdvDow}, {PSD-AuxP/N}은 각각 부정 부사, 강화 정도부사, 약화 정도부사, 부정 보조용언의 시퀀스를 기술한 서브그래프들을 호출한다. 이 그래프는 코스메틱(COS) 도메인의 통사적 감성 MWE의 예로서, 뒤따르는 {SWE-COS-PO}, {SWE-COS-NG}는 각각 코스메틱 도메인의 감성술어를 표상하는 그래프를 호출하고 있다. 위 그래프를 통해 이와 같이 인식된 MWE 시퀀스 전체의 앞뒤에 <POL=극성값> </POL> 방식의 태그가 주석된다.

5.2.2.3 문맥 의존 극성을 갖는 감성 MWE의 LGG

LGG 프레임을 통해 주석되는 MWE 감성표현의 또 다른 범주로는 공기어에 영향을 받는 문맥 의존 감성표현이 있다. 문맥 의존 감성 표현은 특정 문맥을 통해 비로소 극성이 발현되거나 전환되는 표현들을 의미한다. 문맥의존 감성 표현을 살펴보면, 다음과 같은 세 가지 유형의 구문으로 분류될 수 있다. 첫번째 유형은 무극성 술어가 '너무, 좀'와 같은 특정 부사와 결합하여 새로이 부정적 극성을 갖게되는 경우이며(남지순, 2012), 두번째 유형은 자질명과 일련의 기능동사/존재사가 결합하여 극성을 나타내게 되는 경우이다. 세번째 유형은 '빠르다, 크다'와 같은 무극성 술어가 특정 자질명과 공기하여 새로운 극성을 갖게 되는 경우이다.

(13) ㄱ. 옷이 너무 길어요.
ㄴ. 수분감이 있어요./ 수분감이 없어요.
ㄷ. 배터리 다는 속도가 빨라요.

(13ㄱ)은 '너무'와 '길어요'와 같은 측량 형용사 유형이 결합하여 부정 극성을 갖게 되는 예를 보인다. (13ㄴ)은 '수분

감’이라는 코스메틱 도메인의 자질명과 ‘있다/없다’와 같은 기능동사가 결합하여 일정 극성을 표현하게 되는 예를 보인다. (13ㄷ)에서는 서술어 ‘빠르다’가 ‘배터리 다는 속도’라는 자질 표현과 공기하여 부정 극성을 표현하게 되는 경우이다. 이들 시퀀스의 각 구성성분에는 어떠한 극성어휘도 내포되어 있지 않으나, 주어진 문맥에 의해 새로 극성이 부여되는 특징을 보인다. 이와 같이 문맥 의존 극성을 표상하는 LGG는 <그림 7>과 같은 형식으로 기술될 수 있다.

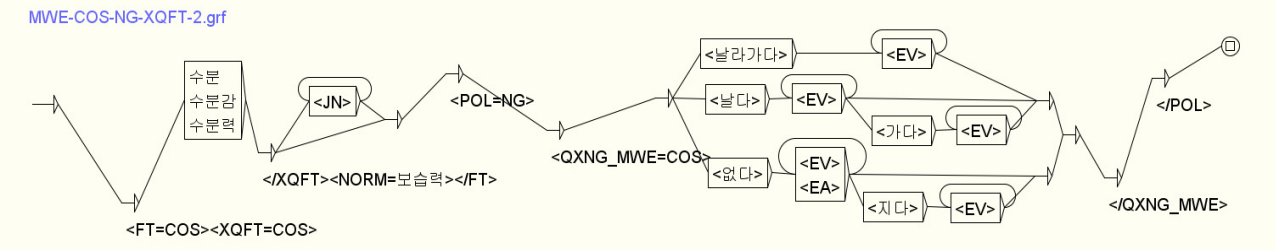

그림 7. 문맥의존 극성을 갖는 MWE를 표상하는 LGG의 예시

<그림 7>의 그래프는 ‘자질명-술어’결합을 통해 극성이 생성되는 (13ㄴ)과 같은 형태들을 표상하는 것으로, ‘수분, 수분감, 수분력’과 같은 자질명이 ‘날라가다, 날아가다, 없다, 없어지다’와 같은 술어와 결합하여 부정 극성을 표현하는 시퀀스를 표상하고 있다. 이 그래프를 통해 해당 MWE 시퀀스 전체의 앞뒤에 해당 태그가 주석될 수 있다.

5.2.3 패턴문법으로 표상된 감성 MWE 시퀀스 전체 규모

이상에서 구축된 DECO-LGG 리소스의 패턴 규모는 각 도메인별로 상이한 분포를 보인다. <표 10>은 각 도메인별 감성 MWE 중에서 어휘적 구성에 한정된 규모를 보이는데, 가령 ‘패션/의류’의 경우 “다리가 길어 보이다”와 같이 동사 ‘보이다’를 내포하는 복합술어가 높은 비중을 차지하기 때문에 상대적으로 전체 패턴 수의 규모가 확장된 것을 볼 수 있다. 또한 ‘금융/핀테크’의 경우, 어플 사용자들이 다양한 업무 및 기능에 대한 불만을 표현하는 방식이 질의, 항의, 요청 등의 여러 화행 형태로 실현되기 때문에 이에 대한 복합술어구 분석을 통해 상대적으로 보다 세분화된 MWE 표현들이 기술되는 결과를 가져왔다. <표 10>을 보면, 도메인별 각 극성별로 10만개에서 20만 정도의 패턴 규모를 보이고 있으며, ‘패션/의류’ 도메인과 ‘금융/핀테크’ 도메인에서 특히 확장된 규모를 이루고 있음을 확인할 수 있다.

표 10. DECO-LGG 도메인별 긍정/부정 MWE 패턴 규모

하위범주	긍정 MWE	부정 MWE
코스메틱	283,103	187,698
IT제품	249,972	106,715
패션/의류	947,336	733,492
푸드/배달	227,989	272,495
인테리어	224,384	117,701
금융/핀테크	1,447,848	2,512,534
K-POP	299,279	681,663

<표 10>의 어휘적 감성 MWE의 패턴 규모는, 현재 DECO-LGG 언어자원에서 범용으로 적용되는 감성 MWE 유형들(예: "마음에 들다")과 부정소를 포함한 통사적 구성(예: "세련되어 보이지 않다"), 또는 강화/약화 부사에 의한 극성전환(예: "완전 예뻐요" {긍정(+1) → 강한긍정(+1.5)})의 형태들은 제외한 것으로서, 이를 고려하면 <표 11>과 같은 결과를 획득할 수 있다.

표 11. 범용 및 도메인특화 DECO-LGG에 기술된 언어자원의 규모

번호	도메인	도메인사전	'통사적' 감성 MWE	'어휘적+통사적' 감성 MWE
CORE	범용 LGG	{DECO 범용 사전}	150,958,290	10,806,603
1	코스메틱	10,179	1,064,852	12,799,924
2	IT 제품	1,613	555,816	4,041,996
3	의류	1,863	1,562,970	15,663,825
4	배달음식	5,985	1,158,546	151,636,265
5	가구/인테리어	1,262	847,536	10,022,087
6	금융 어플	2,984	591,102	2,990,587,965
7	K-POP	5,920	1,444,424	382,165,641

<표 11>의 도메인사전 정보는 DECO-DOM 도메인사전 내 등재된 개체명, 자질명 표제어의 총합을 나타내며, 실제 텍스트에 적용시 다단어 개체명/자질어 패턴 결합 양상에 따라 더 큰 규모로 확장될 수 있다. '통사적' 감성 MWE는 DECO 범용사전 및 DECO-DOM 도메인사전에서 참조될 수 있는 모든 단일어휘 감성 표현들과 정도 부사 및 부정소 사이의 결합으로 생성된 경로의 총합을 보인다. 반면 '어휘적+통사적' 감성 MWE는 <표 10>에서 제시된 도메인별 어휘적 감성 MWE에 다시 통사적 성분들이 결합할 수 있는 유형을 모두 총합한 경로의 수를 보인다. <표 11>에서 예를 들어 코스메틱(COS) 도메인의 경우, 긍정/부정 어휘적 감성 MWE의 규모가 45만개 규모(<표 10>)를 보였다면, 이들이 다시 통사적 성분들과 결합하여 구성되는 패턴의 규모는 1,200만개 규모에 이르는 것을 확인할 수 있다. 반면 어휘적 감성 MWE가 제외된 통사적 감성 MWE의 규모는 100만개 규모로 나타났다.

이와 같이 각 도메인별로 기술되어 있는 DECO-DOM 도메인사전과 도메인 MWE 패턴문법의 자원은, 범용의 DECO 사전의 개체명/자질어 표제어와 범용표현에 대한 MWE 패턴문법의 자원과 함께 실제 텍스트 분석시에 적용되어, 해당 시퀀스가 인식되면, 이에 대한 적절한 태그를 주석할 수 있게 하는 언어자원으로 기능하게 된다.

5.3 SSP에 기반한 TWO-STEP 접근법

5.3.1 [STEP 1]: SSP 방법론

본 연구에서는 반자동 언어데이터 증강(SSP) 방법론에 기반하는 2단계(TWO-STEP) 주석 방식을 적용하여 FeSAD를 구축하였다. 여기서 제안된 2단계 주석 방식은 언어자원에 기반한 주석을 SSP 방식으로 진행하는 [STEP 1] 단계와, 이에 대한 작업자의 후처리 작업이 수행되는 [STEP 2]의 2단계로 구성된다. [STEP 1]에서는 작업자의 개입 없이 앞서 기술한 DECO-LGG 언어 자원이 Unitex 플랫폼을 통해 주석되며, [STEP 2]에서 작업자가 개입하여 [STEP 1]의 주석 결과물에 대한 검수 과정이 수행된다. 이 과정을 도식화하면 다음과 같다.

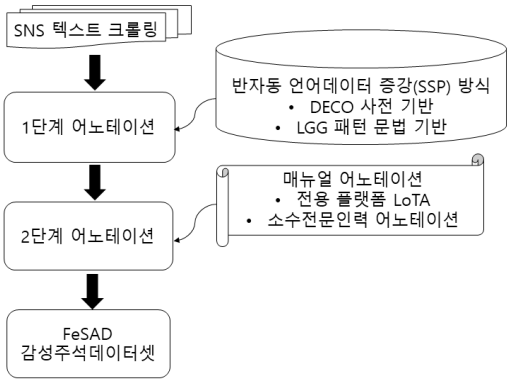

그림 8. SSP 기반 2단계(TWO-STEP) 주석 방법론

SSP 주석 방법론은, 수동으로 수행되는 주석 데이터 구축의 비효율성을 해소하는 반자동 접근법으로, 정교하게 구축된 언어 자원에 기반하여 오피니언 트리플에 대한 주석을 무한 증강하는 것을 가능하게 한다. 언어자원은 DECO 전자사전과 LGG 패턴문법 형식으로 구조화되며, DECO-LGG 언어 자원 내에 기술된 단일어 어휘 및 다단어 패턴에 대한 오피니언 정보를 텍스트에 적용하여 관련 태그를 주석하게 된다. 가령 긍정 술어 '좋다'에 대해, 이러한 언어 자원을 연동하게 되면, '좋아요, 너무 좋은 듯, 안 좋네, 좋지는 않네요'와 같은 다양한 언어 패턴에 대한 감성 주석이 자동으로 수행된다.

5.3.2 [STEP 2]: 후처리를 통한 FeSAD 구축

5.3.2.1 작업자 후처리 과정

DECO-LGG에 기반한 SSP 방식의 주석데이터셋이 구성되면, 이에 대한 매뉴얼 후처리 작업이 진행된다. 이 작업은 디코라 연구센터에서 개발된 DecoLOTA 플랫폼(황창회·남지순, 2020)을 통해 수행된다. DecoLOTA 플랫폼에서는 DECO-LGG에 기반하여 SSP 주석된 데이터셋을 입력문으로 호출하여, 이를 EXCEL 테이블 형식으로 변환한다. 각 문장의 토큰들이 세로로 정렬되고, 그 좌우에 해당 태그들이 부착되어 제공되므로, 이를 바탕으로 작업자가 주석을 수정/추가/삭제하는 작업을 수행할 수 있다. 실제 후처리가 수행되는 문장들의 예를 보면 다음과 같다.

(14) ㄱ. [STEP 1]: 배달넘느려요.
[STEP 2]: ⇒ <FT=FOO>배달</FT> <POL=SN>넘 느려요</POL>
ㄴ. [STEP 1]: 엄마가 카키색 <POL=PO>이쁘다고</POL> 해서 샀는데...
[STEP 2]: ⇒ 엄마가 카키색 이쁘다고 해서 샀는데...

위의 예는 언어 자원의 주석과정에서 나타난 오류 예문들이 작업자에 의해 수정된 결과를 보인다. (14ㄱ)의 경우, 띄어쓰기 부재로 인해 주석되지 않은 예시로, 작업자에 의해 수동으로 태그가 추가되었다. (14ㄴ)은 인용문과 관련된 유형으로, 이 부분은 화자가 궁극적으로 표현하는 부정적 극성 정보와 부합되지 않기 때문에, 예시에서처럼 부착된 주석이 제거되는 과정을 거치게 된다. 이 작업이 종료되면 EXCEL 작업 데이터를 <XML> 방식의 주석데이터로 변환하여 최종 학습데이터를 생성한다.

5.3.2.2 두 쌍 이상의 오피니언 트리플에 대한 페어링

이상과 같이 SSP 주석이 수행된 경우, 문장 내에 여러개의 오피니언 트리플(Opinion Triple) 쌍이 존재하는 경우가 발생한다. FbSA에서는 기본적으로 문장 전체에 대한 극성을 분류하는 것에 목적을 두지 않고, 각 자질(feature) 단위로 대응되는 감성 극성이 어떠한가를 분석하여, 각 자질 단위의 오피니언 트리플을 구축하는 것을 목표로 하므로, 문장 내에 여러 쌍의 명시적인 오피니언 트리플이 관찰되는 경우, 이들 사이의 페어링(pairing) 작업이 수행된다.

매뉴얼 작업 단계에서 이러한 페어링 작업이 동시에 진행되는데, 본 연구에서는 문장 분할(sentence split) 등의 전처리 과정을 통해 데이터를 FbSA에 최적화되도록 전처리 단계를 수행하지만, 이러한 유형의 문장 비중이 높은 도메인의 경우, 의존파서를 연동하여 페어링을 수행한 후 매뉴얼 작업을 진행하도록 하였다.

페어링 작업은 여러 양상을 보이는데, 가령 오피니언 트리플의 '개체명(e)'이 복합구성으로 문장 내에 분리된 형태로 실현되는 경우, 이를 위해 '서브개체명(b)'의 카테고리를 추가할 수 있도록 한다. "BBQ는 양념 치킨도 맛이 좋네요"에서 'BBQ'와 '양념 치킨'은 계층구조를 보이는 복합 개체명으로, 문장 내에서 조사를 함유한 2개의 별개 명사구로 실현되었기 때문에, 하나의 MWE로 주석되기가 어렵다. 이 경우, 'BBQ'은 '개체명(e)'으로, '양념 치킨'은 '서브개체명(b)' MWE로 각각 주석된다.

또한 하나의 개체명(e)에 대해 여러 개의 자질어(f)가 실현되거나 여러 개의 개체명에 여러 개의 자질어가 실현되어 2쌍 이상의 트리플이 나타난 복합문의 경우는 <그림 9>와 같은 인덱싱 주석 방식을 통해 이들을 페어링하는 과정을 수행한다.

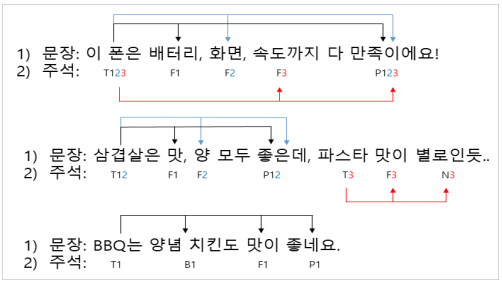

그림 9. 두 쌍 이상의 오피니언 트리플의예시

<그림 9>에 나타난 "삼겹살은 맛, 양 모두 좋은데, 파스타 맛이 별로인듯" 문장의 3개 쌍의 트리플은 다음과 같은 인덱싱 방식을 통해 주석된다.

<NE_1_2>삼겹살</NE_1_2>은 <FT_1>맛</FT_1>, <FT_2>양</FT_2>은 모두 <POL_1_2=PO>좋은데</POL_1_2>, <NE_3>파스타</NE_3> <FT_3>맛</FT_3>이 <POL_3=NG>별로인듯</POL_3>.

각 태그 내 포함된 인덱싱 정보를 활용하여 최종적으로 다음과 같은 3개 쌍의 오피니언 트리플을 추출할 수 있다.

(15) {삼겹살, 맛, POSITIVE/좋은데}, {삼겹살, 양, POSITIVE/좋은데}, {파스타, 맛, NEGATIVE/별로인 듯}

6. FeSAD 구축에 적용된 SSP 접근법의 성능 평가

앞서 논의한 바와 같이, 본 연구에서 제안한 TWO-STEP 주석 방식은, 1단계에서 DECO-LGG기반 SSP 방식의 주석을 수행한 후, 2단계에서 매뉴얼 주석을 수행하는 방식으로 진행된다. 이때 1단계에서 사용되는 언어자원의 구조와 규모에 대해서는 앞 장에서 기술된 바 있다. 이 장에서는 이와 같이 1단계에서 DECO-LGG 언어자원에 기반하여 SSP 반자동 증강방식으로 학습데이터를 증강한 경우, 그 데이터의 성능을 분석하여 본 연구의 SSP 기반 접근법의 성능을 평가한다.

성능 평가를 위해, 코스메틱(COS)과 푸드/배달(FOO) 도메인에서 새로운 후기글 텍스트를 크롤링하였다. 코스메틱 도메인의 후기글은 '네이버 쇼핑몰'의 화장품 범주에서, 푸드/배달음식 후기글은 '배달 어플 요기요'에서 크롤링하였다. 코스메틱 분야에서 126,377개 리뷰글, 그리고 푸드/배달음식 분야에서 127,878개의 리뷰글을 수집하여, 다시 랜덤 방식으로 각각 1,000개의 후기글 데이터를 추출하였다.

추출된 후기글 텍스트에 대해 2명의 작업자가 전체 정답 주석을 매뉴얼 방식으로 구성하였고, 이후 동일한 데이터에 SSP 방식으로 주석된 1단계 데이터셋과 비교 평가하였다. 이를 통해 SSP 주석 결과에 대해 다음과 같은 성능을 획득하였다.

표 12. 코스메틱/푸드 도메인의 SSP 성능 평가 결과

도메인	RECALL	PRECISON	F1-SCORE
코스메틱	0.92	0.94	0.93
푸드/배달	0.87	0.93	0.90

코스메틱 도메인과 푸드 도메인에서 SSP 주석의 정확율은 각각 0.94와 0.93으로, 두 도메인 모두 재현율(0.92와 0.87)에 비해 정확율이 높은 값으로 나타났고, 전체 F1스코어는 각각 0.93과 0.90으로 나타났다. 두 도메인의 성능 점수를 평균하면 0.915로서, 전체 주석작업의 9/10 정도의 작업이 SSP를 통한 1단계 주석 단계에서 처리될 수 있음을 예측할 수 있다.

실제로 동일한 작업에 대해 SSP를 진행하지 않고, 전체 작업에 대해 전문인력이 매뉴얼하게 주석을 수행하는 경우에 비해 소요되는 시간을 현저히 낮출 수 있었다. 뿐만 아니라, 고빈도로 실현되는 특정 주석 유형의 경우, 작업자의 피로감으로 인해 일관성이 유지되지 못하는 현상들이 나타나며, 굳어진 관용적 MWE 표현들의 경우, 작업자의 언어적 용법에 대한 지식이 부족할 때에는 올바르게 주석되지 못하는 현상들이 관찰되었다. 반면 이러한 반복적 유형이나 다단어 시퀀스 유형은 SSP 방식으로 최적화된 성능을 보이는 것으로 나타났다.

7. 결 론

본 연구에서는 FbSA 연구에서 머신러닝 언어모델을 학습시키기 위해 요구되는 대규모의 정교한 학습데이터를 구축하는 데에 있어서, DECO-LGG 언어자원에 기반한 반자동 언어데이터 증강(SSP) 방식에 입각하여 주석 작업을 2-STEP으로 진행하는 접근법을 제안하였다. 1단계에서 SSP 방식의 주석을 수행한 후, 2단계에서 매뉴얼 주석을 진행하는 방식

으로서, 1단계 작업만으로도 평균 0.915의 높은 주석 성능을 보임을 확인하였다. 이를 통해 FbSA용 학습데이터 주석을 위한 작업자의 작업 비중이 언어 자원의 적용을 통해 분담될 수 있으며, 학습데이터 구축을 위한 프로세싱의 소요시간과 품질이 획기적으로 개선될 수 있음을 확인하였다.

SSP의 핵심이 되는 DECO-LGG 언어 자원은, 범용과 도메인 특화 자원으로 이원화되어 구조화되었다. 따라서 새로운 도메인 분야로 확장하는 경우, 현재 코어가 되는 범용자원을 공통으로 사용할 수 있으며, 새로운 도메인에 필요한 자원만을 추가할 수 있도록 모듈화되어 있어, 작업의 효율성을 극대화할 수 있도록 설계되어 있다. 언어 자원에 기반한 주석 성능, 다양한 도메인 자원과 확장성을 고려했을 때, 현재 본 연구에서 제안한 방법론은 향후 다양한 영역에서의 FbSA를 위한 머신러닝용 학습데이터를 생성하는 데에 중요한 접근법으로 활용될 수 있을 것으로 기대된다.

참고문헌

김문형 · 장하연 · 조유미 · 신효필. 2013. KOSAC (Korean Sentiment Analysis Corpus): 한국어 감정 및 의견분석 코퍼스. 『한국정보과학회 학술발표논문집』, 650-652.

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

남지순. 2018. 『코퍼스 분석을 위한 한국어 전자사전 구축 방법론』. 도서출판 역락.

남지순. 2021. 자질기반 감성분석(FbSA) 모델의 인공지능 학습을 위한 지식베이스 · 패턴문법 기반 반자동 학습데이터 증강(SSP) 방법 및 장치. DICORA-TR-2021-10. 한국외대 디코라연구센터.

박상민 · 나철원 · 최민성 · 이다희 · 온병원. 2018. Bi-LSTM 기반의 한국어 감성사전 구축 방안. 『지능정보연구』, 24.4, 219-240.

신동혁 · 조동희 · 남지순. 2016. 한국어 감성 사전 DecoSelex 구축을 위한 영어 SentiWordNet 활용 및 보완 논의. 『한국사전학』 28, 75-111.

안애림. 2011. 한국어 오피니언 문장 분류 시스템을 위한 사전 및 구문 패턴 연구. 한국외대 석사학위논문.

이도영. 2021. 온라인 텍스트종류로서의 상품평 연구 - 독일 아마존 온라인 쇼핑몰의 '마스크' 상품평을 중심으로. 『독일언어문학』 0.93, 53-73.

이상진 · 조은경. 2020. 책 리뷰 말뭉치를 활용한 형용사, 부사 감성 사전 개발. 『언어과학연구』 92, 27-40.

이준환 · 정성환 · 노정옥 · 박근호. 2013. 한국어 맛 평가 형용사에 관한 연구. 『감성과학』 16.4, 493-502.

조동희 · 신동혁 · 남지순. 2016. MUSE 감성주석코퍼스 구축을 위한 분류 체계 및 태그셋 연구. 『우리말연구』 47.5-47.

조수선. 2007. 온라인 신문 댓글의 내용분석: 댓글의 유형과 댓글 게시자의 성향. 『커뮤니케이션학 연구』 15.2, 65-84.

조수지 · 김흥규 · 양철원. 2021. 기업 재무분석을 위한 한국어 감성사전 구축. 『한국증권학회지』 50.2, 135-170.

황창회 · 남지순. 2020. DecoLoTA. DICORA-TR-2020-01. 한국외대 디코라연구센터.

황창회 · 유광훈 · 남지순. 2018. 자질기반 감성분석을 위한 다단어표현 사전 DecoMWE 연구. 『언어와 정보 사회』 35, 407-458.

Ahn, A., E. Laporte, and J. Nam. 2012. Semantic Polarity of Adjectival Predicates in Online Reviews. ArXiv, abs/1211.4161.

Aragon, O. R., M. S. Clark, R. L. Dyer, and J. A. Bargh. 2015. Dimorphous Expressions of Positive Emotion: Displays of Both Care and Aggression in Response to Cute Stimuli. *Psychological Science* 26, 259-273.

Ataei, T. S., K. Darvishi, B. Minaei-Bidgoli, and S. Eetemadi. 2019. Pars-ABSA: An Aspect-based Sentiment Analysis Dataset in Persian. ArXiv, abs/1908.01815.

Blitzer, J., M. Dredze, and F. C. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 440-447.

Gross, M. 1997. *The Construction of Local Grammars. Finite-State Language Processing*, Roche and Schabes(eds.). the MIT Press.

Hyun, D., J. Cho, and H. Yu. 2020. Building Large-Scale English and Korean Datasets for Aspect-Level Sentiment Analysis in Automotive Domain. *Proceedings of the 28th International Conference on Computational Linguistics*, 961-966.

Lei, Z. and B. Liu. 2011. Identifying Noun Product Features That Imply Opinions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 575-580.

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

Mudalige, C. R., D. Karunarathna, I. Rajapaksha, N. D. Silva, G. Ratnayaka, A. Perera, and R. Pathirana. 2020. SigmaLaw-ABSA: Dataset for Aspect-Based Sentiment Analysis in Legal Opinion Texts. *2020 IEEE 15th International Conference on Industrial and Information Systems* (ICIIS), 488-493.

Nam, J.-S. 2014. A Novel Dichotomy of the Korean Adverb Nemwu in Opinion Classification. *Studies in Language* 38.1, 171-209.

Paumier, S. 2003. *Unitex Users' Manual*. France: UPEM.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation* 27-35.

Ray, B., A. Garain, and R. Sarkar. 2021. An Ensemble-based Hotel Recommender System Using Sentiment Analysis and Aspect Categorization of Hotel Reviews. *Applied Soft Computing* 98, 106935.

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

Wiebe, J., T. Wilson, and C. Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39.2-3, 165-210.

윤정우(제1저자), 대학원생
경기도 용인시 처인구 모현 외대로 81
한국외국어대학교 언어인지과학과
E-mail: skyjw1211@gmail.com

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