반자동 언어데이터 증강 방식에 기반한 FbSA 모델 학습을 위한 감성주석 데이터셋 FeSAD 구축

윤정우○, 황창회, 최수원 & 남지순
한국외국어대학교, DICORA 연구센터/언어인지과학과
skyjw1211@gmail.com, hch8357@naver.com, soown0607@gmail.com, jeesun.nam@gmail.com

Building Sentiment-Annotated Datasets for Training a FbSA model based on the SSP methodology

Jeong-Woo Yoon○, Chang-Hoe Hwang, Su-Won Choi & Jee-Sun Nam
DICORA, Hanguk University of Foreign Studies

요 약

본 연구는 한국어 자질 기반 감성분석(Feature-based Sentiment Analysis: FbSA)을 위한 대규모의 학습 데이터 구축에 있어 반자동 언어데이터 증강 기법(SSP: Semi-automatic Symbolic Propagation)에 입각한 자질-감성 주석 데이터셋 FeSAD(Feature-Sentiment-Annotated Dataset)의 개발 과정과 성능 평가를 소개하는 것을 목표로 한다. FeSAD는 언어자원을 활용한 SSP 1단계 주석 이후, 작업자의 주석이 2단계에서 이루어지는 2-STEP 주석 과정을 통해 구축된다. SSP 주석을 위한 언어자원에는 부분 문법 그래프(Local Grammar Graph: LGG) 스키마와 한국어 기계가독형 전자사전 DECO(Dictionnaire Electronique du COréen)가 활용되며, 본 연구에서는 7개의 도메인(코스메틱, IT제품, 패션/의류, 푸드/배달음식, 가구/인테리어, 핀테크앱, KPOP)에 대해, 오피니언 트리플이 주석된 FeSAD 데이터셋을 구축하는 프로세싱을 소개하였다. 코스메틱(COS)과 푸드/배달음식(FOO) 두 도메인에 대해, 언어자원을 활용한 1단계 SSP 주석 성능을 평가한 결과, 각각 F1-score 0.93과 0.90의 성능을 보였으며, 이를 통해 FbSA용 학습데이터 주석을 위한 작업자의 작업이 기존 작업의 10% 이하의 비중으로 감소함으로써, 학습데이터 구축을 위한 프로세싱의 소요시간과 품질이 획기적으로 개선될 수 있음을 확인하였다.

주제어: 감성주석 데이터셋, 자질기반 감성분석, 반자동 언어데이터 증강, DECO 전자사전, LGG 패턴문법

1. 서론

본 연구는 한국어 '자질기반 감성분석(FbSA: Feature-based Sentiment Analysis)' 시스템의 머신러닝 언어 모델을 개발하는 데에 필요한 대규모 학습데이터를 구축하기 위해서, [1]에서 제안된 반자동 언어데이터 증강(SSP: Semi-automatic Symbolic Propagation) 방법론에 입각하여 감성주석 데이터셋 FeSAD(Feature-Sentiment-Annotated Dataset)를 개발하는 과정과 그 성능에 대해 소개하는 것을 목표로 한다.

FbSA를 위한 학습데이터를 구축하는 데에 있어 고려되어야 할 사항은 다음 두 가지 관점에서 정리될 수 있다. 첫째는 '대상 텍스트'의 문제이다. 사용자들의 주관적인 의견, 평가, 감정 등이 실현되는 가장 중요한 자연언어 텍스트 원천으로서 '소셜미디어(social media) 텍스트'는 실제로 그 도메인과 플랫폼에 따라 텍스트의 특징이 서로 차이를 보인다. 상품후기글과 블로그, 트위터 및 신문기사 댓글 등에 나타나는 표현들은 서로 같지 않아서, 가령 정치기사 댓글에 나타나는 감정표현을 보면, 반어법적 야유글의 비중이 높아 '수사적 의문문'의 비중이 비정상적으로 높게 나타나며, 부정적(negative) 오피니언의 비중이 높게 나타난다. 반면 상품후기글에 나타나는 감정표현을 보면, 상품 카테고리에 따라 이에 수반되는 감정표현이 달라지기 때문에 가령 코스메틱 도메인에서 나타나는 '촉촉하다(+), 가루가 날리다(-)'와 같은 감정표현들은 다른 상품 후기글에서는 관찰되지 않는 도메인 의존적인 특징을 보인다. 또한 트위터(Twitter)와 같이 글자수에 제약이 있는 플랫폼에서, 띄어쓰기와 같은 맞춤법이 파괴되어 발생하는 노이즈(noise) 현상은 블로그와 같은 장문의 텍스트를 업로드하는 플랫폼에서는 상대적으로 이러한 '형식적' 노이즈의 비중이 적게 나타나는 것을 관찰할 수 있다. 그러나 140자의 제한된 환경에서, 사용자는 주장하고자 하는 개인의 주관적 의견/감정 표현에 집중하는 반면, 개인 블로그와 같은 환경에서는 '비감성적(no-opinion)' 문장 비중이 높아, 실제 감성분석을 위한 데이터 생성에 또다른 유형의 노이즈를 발생하게 된다. 일반적으로 소셜미디어 텍스트로 총칭하는 다양한 문서들에 대해서 이와 같은 도메인, 플랫폼별 차이점을 고려한다면, 실제로 어떠한 유형의 데이터를 학습데이터로 설정하는가가 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 구성하는 것이 가능하나, 자질층위의 정보가 부착되어 있는 학습데이터는 자동으로 획득하는 것이 용이하지 않기 때문에 클라우드소싱과 같은 방법으로 직접 수동 구축이 수반되어야 한다.

다만 이 과정은 시간과 비용적 측면에서뿐 아니라, 문장의 개별 요소들에 대한 언어학적 의미-형식적 어노테이션을 수행할 수 있는 전문성이 확보되어야 한다는 점에서 한층 더 난이도가 높은 태스크가 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 플랫폼들을 그 대상으로 하였으며, 이때 동일 플랫폼에서도 후술하는 3장에서 보는 바와 같이, 여러 다른 카테고리 도메인의 텍스트를 대상으로 설정하였다. 둘째로 자질기반 감성주석을 수작업으로 수행하는 접근법의 비효율성을 극복하기 위해서 DECO 한국어 전자사전[2]과 LGG(Local-Grammar Graph) 프레임[3]에 기반한 '반자동 언어데이터 증강(SSP)' 방식을 활용하여 1단계에서 SSP의 어노테이션이 수행되면, 2단계에서 매뉴얼한 방식으로 2차 어노테이션이 수행되도록 하였다. 현재 SSP 방식에 기반하여 수행된 1단계 어노테이션은 '코스메틱 도메인(COS)'의 경우 F1-Score가 0.93, '푸드 도메인(FOO)'의 경우 0.90의 성능을 보였다. 즉 이 과정을 통해 2단계에서 투입되는 작업자의 매뉴얼 어노테이션의 비중은 종래 방식의 1/10 규모로 축소되어 진행될 수 있음을 확인하였다.

본 연구에서 제안하는 SSP기반 감성주석 데이터셋 FeSAD 구축과정은 그림 1과 같이 도식화될 수 있다.

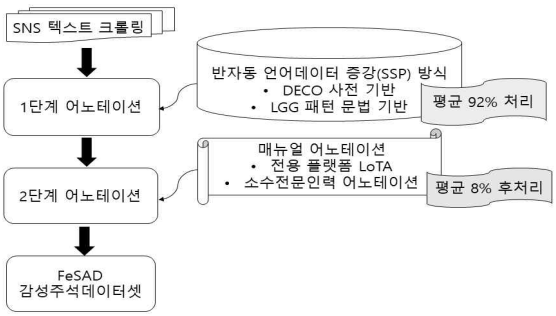

그림 1 SSP 기반 FeSAD 구축 프로세싱

2장에서는 감성주석 코퍼스와 FbSA와 관련된 선행 연구들에 대해 논의하고, 3장에서는 FeSAD 데이터셋을 구성하는 텍스트 도메인에 대해 논의한다. 4장에서는 본 연구에서 구현한 FeSAD의 주석체계와 데이터셋 구성에 대해 논의한 후, 5장에서는 SSP 방식의 성능을 실험한다. 6장에서는 본 연구의 의의 및 향후 연구 방향에 대해 논의한다.

2. 관련 연구

감성 분석과 관련한 연구는 다양한 분야에서 이루어졌다. 국내의 감성 분석 연구는 초기에는 사전을 기반으로한 감성분석이 주를 이루었으나, 2010년대 중반에 접어들어 기계학습 기반의 방법론, 그중에서도 딥러닝 방법론을 중심으로 많은 연구들이 이루어지고 있다. 특히 딥러닝 방법론은 대량의 주석 데이터가 요구되기 때문에, 이에 따라 감성주석 데이터의 중요성도 더 강조되었다.

감성주석 데이터에 있어, 주관성 표지와 의견 및 감정표현과 관련한 다양한 주석들을 제안한 MPQA(Multi Perspective Question Answering) 코퍼스[4]와, 국내에서 MPQA를 바탕으로 한국어에 맞게 적용한 KOSAC[5]을 볼 수 있다. MPQA와 KOSAC 데이터의 경우, 문장 내 반영된 태도, 사용자 생성문에 기반한 데이터가 아니기에 사용자 생성문이 갖는 비정형성을 고려하지 못하며, 다양한 도메인별 특징이 고려되지 않다는 한계점을 가진다.

사용자 생성문에 대한 감성 분석 데이터로는 SemEval-16, 17에서 사용된 SNS 플랫폼 중 하나인 Twitter 감성분석 평가 데이터와 Twitter 기반 감성 주석 코퍼스인 MUSE 코퍼스가 있다[6]. Twitter의 특성상 키워드를 통해 여러 분야의 데이터를 수집하는 것이 가능하기에 다양한 도메인별 언어적 특징들이 포착될 수 있는 장점이 있으나, 이 경우 수집된 코퍼스가 특정 어휘를 필수적으로 포함하여야 한다는 점에서 도메인별 전체적인 어휘 양상을 담기 어렵다는 한계가 있다.

사용자 생성문의 특징과 도메인 정보를 포착하기 위한 데이터셋의 구축 외에 도메인별 언어 자원의 분석 방법이나, 구축 방법론에 집중하여 도메인별 감성분석의 성능을 높이고자 한 연구들도 제안되었다. 학습 모델의 개선을 통해 도메인 간 성능 차이를 줄이고자 [7]에서는 특정 도메인의 감성분석을 타 도메인의 주석 코퍼스를 적용했을 때 SCL(structural correspondence algorithm)을 사용하여 도메인 차이로 인해 발생하는 감성분석의 오류를 줄이고자 하였다. 그 외 도메인 감성사전 구축을 위해, 기존의 범용 감성사전 내 어휘들에 대해 도메인별 가중치를 부여하여 도메인 감성사전을 구축하는 방법론과 관련된 연구도 진행되었다[8]. 그러나 이 두 연구 모두 통계적인 접근방식으로, 다단어 감성표현과 같은 언어적 특징들을 추출하는 데에는 한계가 있다.

FbSA를 위한 주석코퍼스에 대한 연구로는 SemEval-14, 15, 16에서 사용된 FbSA의 평가 데이터가 있다.[9] 이 데이터는 레스토랑과 노트북 후기글에 대한 영문 텍스트로 이루어진 데이터이다. 여기서는 각 텍스트에 대한 자질-극성 분류 결과가 XML 형식으로 주석된다. 또 다른 영어 FbSA 데이터로는 [10]에서 활용된 호텔 리뷰데이터가 있다. 해당 데이터셋은 호텔 리뷰글을 미리 설정한 자질 분류에 따라 텍스트를 분류하고, 극성 분류는 수집된 데이터의 메타 정보 중 각 리뷰별로 부여된 사용자의 1-5까지의 점수를 활용하는 방법을 사용하였다. 국내의 경우, 자질기반 감성분석 주석 코퍼스에 대한 본격적인 연구는 찾아보기가 어렵다. [11]에서는 자동차 도메인에 대한 영어, 한국어 자질기반 감성분석 코퍼스를 구축하였는데, 감성 평가의 대상을 생산업체로, 자질을 해당 생산업체의 대표 생산 모델로 설정하였다. 사용자 리뷰글에 대해 자질 분류를 수행한 후, 작업자들에게 사용자 리뷰와 해당 리뷰글의 자질 분류 정보를 제공하여, 리뷰글에 대한 감성 분류를 수행하도록 하였다.

이상에서 살펴본 기존 FbSA용 데이터셋에서는 자질명 및 감성 평가에 대해 단순 분류 주석 체계가 제안되었으나, 본 연구는 분류 정보에 대한 주석과 더 나아가 텍스트 내 구체적인 표현 정보를 FbSA 주석 체계에 포함하였으며, 이를 반자동으로 증강하는 방법론에 입각하여, 효율적인 방식으로 신뢰할만한 대규모의 주석 데이터셋을 구축하는 프로세싱을 제안하였다.

3. FeSAD 텍스트 도메인 특징

FbSA를 위한 감성주석 데이터셋 FeSAD는 한국외대 디코라연구센터(http://dicora.kr)에서 구축되어온 감성주석코퍼스(SAC) 데이터셋의 일환이다. FeSAD 이전의 데이터셋으로, 현재 DICORA센터에 구축되어 있는 감성주석코퍼스의 구조를 보면 표 1에서와 같다.

표 1. DICORA 감성주석코퍼스 SAC 데이터셋 유형

번호	유형	명칭	플랫폼	도메인	규모(문장)
1	문장주석	MUSAC	댓글/카페/블로그/쇼핑몰	쇼핑/교육/문화 등 19개	200,000
2		WESAC	트위터	정치/경제/사회/문화	50,000
3	자질주석	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 [6]에서 MUSE 코퍼스로 소개된 바 있다. 여기에는 트위터뿐 아니라 정치댓글/상품후기/게임/스포츠/교육/관광 등의 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 감성주석데이터를 바탕으로 하여, SSP 방식의 비중을 부트스트랩(bootstrap) 방식으로 증가시킴으로써 표 2와 같은 구성의 데이터셋을 구성하였다.

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 진행되었기 때문에 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.

그외, 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 코퍼스 일부를 추출하여 해당 데이터셋(KPOP)을 구성하였다.

이상에서 구성된 FeSAD 데이터셋의 전체 규모는 42,000여 문장이며, 도메인의 의미적 특징에 따라 관련 플랫폼이 결정되는 방식으로 진행되었다.

4. FeSAD 주석 체계 및 데이터셋 구성

4.1. DECO-LGG 기반 주석 체계

본 연구에서 FbSA를 위해 주석되는 자질기반 원소는 [12]에서 정의한 오피니언 5원소쌍(Opinion Quintuple)에서, 보통 상품후기글에서 메타정보(meta-information)로 실현되는 '평가자(opinion holder)'와 '평가시간(opinion time)'을 제외한 다음의 3가지 성분이 중심이 된다.

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

위에서 e는 '개체명' 부류로서, 도메인에 따라 그 의미적 특징과 어휘적 구성이 달라지는 열린 목록을 구성하는 반면, f는 '자질명(또는 속성명)'으로서, e와는 달리 제한된 유형의 일반 명사구들이 해당된다. s는 감성극성으로서 긍정/부정 등의 극성어 표현과, 일련의 극성전환장치들(PSD: Polarity-Shifting Device)[13][14]에 기반한 시퀀스들을 포함한다.

4.1.1. DECO 사전에 기반한 '단일어' 주석

DECO 사전 층위의 주석은 '단일어(simple word)' 어휘에 대해 이루어지며, 도메인에 무관한 범용의 어휘 정보를 담은 DECO 사전과 도메인 특화된 DECO-DOM 도메인사전의 의미 태그들을 참조하여 주석된다.

DECO 전자사전에는 30만여개의 표제어와 각 표제어별 형태, 통사, 의미, 감성 정보들이 명시되어 있으며, 각 표제어에는 활용클라스 정보가 부착되어 있어, 이를 통해 해당 활용어미 트랜스듀서가 호출되도록 설계되어 있다. DECO 사전과 호환되는 Unitex 플랫폼[15]에서 코퍼스 분석을 위해 곧바로 적용될 수 있으며, 이때 실제 텍스트에 실현되는 모든 표면형 어절들에 대한 올바른 형태소단위 분석이 가능하게 된다.

오피니언 트리플 주석을 위해 DECO 사전에 등재된 개체명 및 감성어휘 분류정보를 사용할 수 있다. 여기 등재된 11가지 개체명 분류 체계(EntLex)와 4가지 극성어휘 분류 체계(PolLex)를 이용하여 개체명과 극성어 태그가 부착된다.

앞서 논의한 바와 같이, 실제 텍스트에 실현되는 자질표현 및 감성표현에는 '가격'이나 '좋다/나쁘다'와 같이 범용의 사전에서 정의될 수 있는 유형 외에도, '발색력'이나 '가루가

날리다'와 같이 도메인 특화된 자질어/감성어 표현이 실현되므로, 이를 위한 '도메인사전 DECO-DOM'이 별도로 구축되어 적용되어야 한다. 가령 현재 코스메틱 도메인사전에 등재되어 있는 개체명 표제어는 9,959개이며, K-POP 도메인의 경우 5,789개, 배달앱 푸드 도메인의 경우 5,834개의 규모를 이루고 있다.

자질어 표현의 경우도 각 도메인별 특화된 양상을 보이므로, 가령 배달앱 푸드의 경우는 '맛, 양'과 같은 자질어가 실현되지만 이러한 자질어 성분은 인테리어제품에 대한 후기글에서는 자질어로 사용되지 않는다. 이러한 방식으로 각 자질어 및 감성어휘들이 해당 도메인사전에 별도로 저장되어, 실제 감성주석을 위한 언어자원으로 활용될 때에는 DECO 범용사전과 해당 도메인사전(예: DECO-COS(코스메틱), DECO-FOO(배달앱 푸드))이 페어링하여 적용되는 방식이 사용된다.

4.1.2. LGG 패턴문법에 기반한 다단어(MWE) 패턴 주석

4.1.2.1. 개체명 및 자질어 MWE 주석

위에서 단일어에 대한 주석을 위해 DECO사전의 표제어 태그 정보가 사용되었다면, 텍스트에서 실현되는 복합명사구, 복합동사구, 또는 부정소나 정도부사가 결합한 통사적 연쇄 등으로 실현되는 일련의 다단어(MWE: Multi-Word Expression) 연쇄에 대한 주석을 위해 LGG 프레임이 사용된다. LGG 프레임은 일련의 언어 시퀀스를 방향성그래프 형식으로 기술하여, 이를 Unitex 플랫폼에서 유한상태 트랜스듀서(FST: Finite-State Transducer)로 자동변환한 후 텍스트 분석 및 주석에 적용하는 접근법이다.

가령 코스메틱 도메인에는 MWE 형식으로 구성된 개체명의 출현이 빈번한데, 예를 들어, "이자녹스 UV 선크림"과 같은 개체명의 경우, 도메인사전에 등재된 원소들의 결합 관계를 그림 2와 같은 LGG 문법으로 기술하여 해당 시퀀스에 일련의 주석을 삽입하는 것이 가능하다.

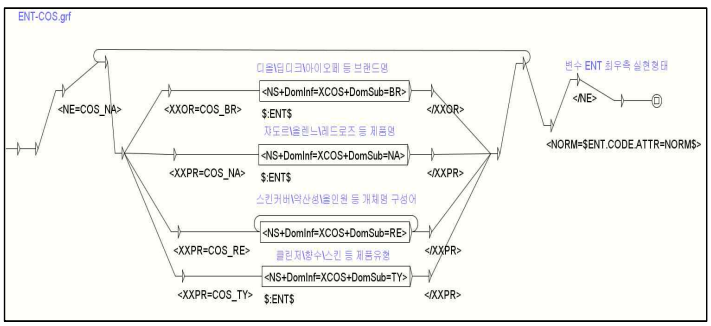

그림 2. 코스메틱 도메인 MWE 개체명 주석 LGG

4.1.2.2. 범용/도메인특화 감성표현 MWE 주석

실제 텍스트에서 등장하는 감성표현은 앞서 단일어 구성형태뿐 아니라, 다양한 유형의 MWE 형태로 실현될 수 있다. 감성MWE 유형은 크게 두 가지 범주로 분류할 수 있다.

첫째는, 감성을 표현하는 서술어(서술명사/서술용언) 자체가 여러 개의 토큰으로 구성되는 경우로서, 전형적인 관용표현, 연어, 복합구성으로 분류되는 일련의 언어표현들이 여기 해당한다. 다음에서 나타나는 서술어 성분들은 이러한 예를 보인다.

(2) ㄱ. 이번 갤럭시 S21 디자인이 맘에 쏙 드넹
ㄴ. 이번 앨범은 멤버들이 뼈를 갈아 넣은 거 같애

둘째는, '부정소(Negator)'가 부사나 보조용언 형태로 결합되어 기저의 서술어 고유의 극성을 반대로 전환하거나, 또는 '강화(Amplifier)/약화(Downtoner)'의 정도부사가 삽입되어 기저의 극성의 정도를 변환시키는 통사적 구성을 이루는 경우이다.

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

위의 예에서 보는 바와 같이, (3ㄴ)의 '예쁘다'는 '긍정(PO)' 술어(+1점)로서, 여기에 (3ㄱ)처럼 강화 부사 '엄청'이 결합하면 '강한긍정(SP)'(+1.5점)이 된다. 반면 (3ㄹ)처럼 부정소 '안/않다'가 결합하면 극성이 전환되어 부정 극성 술어(=안 예쁘다)가 되는데, 이러한 {부정소+긍정} 결합형의 경우는 '약한부정(WN)'(-0.5점)이 되도록 정의하였다.

같은 방식으로 (3ㅁ)의 '촌스럽다'는 '부정(NG)' 술어(-1.0점)로서, 여기에 (3ㅂ)처럼 강화 부사 '엄청'이 결합하면 '강한부정(SN)'(-1.5점)이 된다. 반면 (3ㄷ)처럼 부정소 '안/않다'가 결합하면 극성이 전환되어, 긍정 극성 술어(=촌스럽지 않다)가 되는데, 이 경우에도 {부정소+부정} 결합형의 경우는 '약한긍정(WP)'(+0.5점)이 되도록 정의하였다.

이상과 같은 6가지 극성분류 정보는 그림 3과 같은 LGG 패턴문법을 통해 주석될 수 있다.

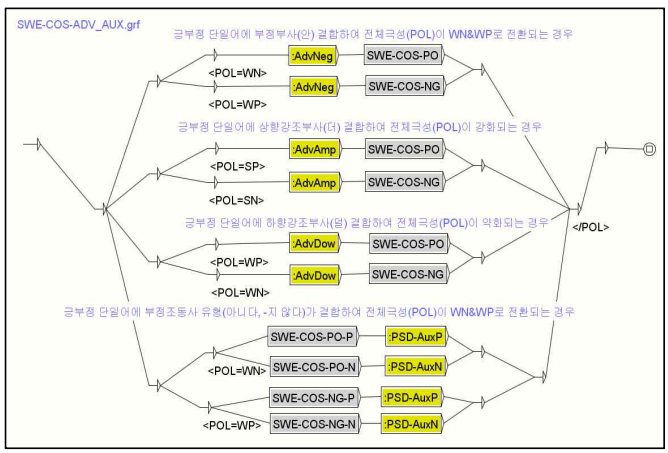

그림 3. 극성전환 태그가 부여되는 감성MWE 표상 LGG

4.2. 반자동증강을 위한 DECO-LGG 언어자원 규모

반자동 언어데이터 증강(SSP)을 위한 DECO-LGG 리소스의 패턴 규모는 각 도메인별 상이한 분포를 보인다. 표 3은 각 도메인별 긍정과 부정의 감성 MWE의 규모를 보이는데, 가령 '패션/의류'의 경우 "다리가 길어 보이다"와 같이 동사 '보이다'를 포함한 복합술어의 비중이 추가되기 때문에 상

대적으로 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 복합술어구의 구조를 기술하는 과정을 통해 상대적으로 보다 세분화된 MWE 표현들이 표상되었다.

표 3. DECO-LGG 도메인별 긍정/부정 MWE 패턴 규모

번호	유형	하위범주	긍정MWE	부정MWE
2	LoSAC 도메인	코스메틱	283,103	187,698
3		IT제품	249,972	106,715
4	의식주 도메인	패션/의류	947,336	733,492
5		푸드/배달	227,989	272,495
6		인테리어	224,384	117,701
7	문화 도메인	금융/핀테크	1,447,848	2,512,534
8		K-POP	299,279	681,663

표 3의 긍정/부정 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 규모를 보인다. 즉 명사구 또는 동사구/형용사구 형태로 실현되는 감성서술어의 어휘변이형 패턴들만을 고려한 결과로서, 여기에 범용의 감성 MWE 및 통사적 구성의 MWE들을 추가하면 그 수는 보다 큰 폭으로 확장된다.

4.3. DecoLOTA 플랫폼과 매뉴얼 후처리 주석 결과

DECO-LGG에 기반한 SSP 방식의 주석데이터셋이 구성되면, 이에 대한 매뉴얼 후처리 작업이 진행된다. 이 작업은 디코라 연구센터에서 개발된 DecoLOTA 플랫폼[16]을 통해 수행된다. DecoLOTA 플랫폼에서는 DECO-LGG에 기반해 SSP 주석된 데이터셋을 입력문으로 호출하여, 이를 EXCEL 테이블 형식으로 변환한다. 각 문장의 토큰들이 세로로 정렬되고, 그 좌우에 해당 태그들이 부착되어 있어 작업자가 주석을 수정/추가/삭제하는 작업을 수행할 수 있다. 이 작업이 종료되면 이를 <XML> 방식의 주석데이터로 변환하여 학습 데이터를 생성한다.

본 연구에서 FeSAD 데이터셋에 해당 원소에 정보가 주석되는 형식은, XML(Extended Markup Language)-MERGE 방식으로, 다음과 같은 방식으로 실현된다.

```
이 집 <NE=FOO>치킨</NE>이 <FT=FOO>맛</FT>은
<POL=SP>최고네요</POL>.
```

주석 방식은 <SeqType=value>, </SeqType>형식의 좌우 태그로 구성되며, 이때, '<NE=FOO>치킨</NE>'에서 'FOO(푸드/배달음식)' 도메인 영역의 개체명 분류 'NE'에 대한 주석임을 확인할 수 있다.

4.4. 두 쌍 이상의 Opinion Triple 페어링

이상과 같이 SSP 주석이 수행된 경우, 문장 내에 여러개의 오피니언 트리플(Opinion Triple) 쌍이 존재하는 경우들이 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 오피니언 트리플이 관찰되는 경우, 이들 사이의 페어링(pairing) 작업이 수행된다.

매뉴얼 작업 단계에서 이러한 페어링 작업이 동시에 진행되는데, 본 연구에서는 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 방식으로 수행되는데, 가령 오피니언 트리플의 '개체명(e)'이 복합구성으로 문장 내에 분리된 형태로 실현되는 경우, 이를 위해 '서브개체명(b)'의 카테고리를 추가할 수 있도록 하였다. "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 4와 같은 인덱싱 주석 방식을 통해 이들을 페어링하는 과정을 수행한다.

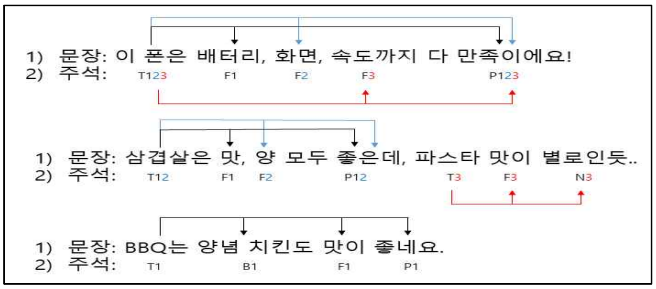

그림 4. 두 쌍 이상의 오피니언 트리플의 페어링 예시

그림 4에 나타난 "삼겹살은 맛, 양 모두 좋은데, 파스타 맛이 별로인듯" 문장의 3개 쌍의 트리플은 다음과 같은 인덱싱 방식을 통해 어노테이션된다.

```
<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>..
```

5. DECO-LGG기반 SSP 방식의 주석 성능 평가

본 연구에서 제안한 2-STEP 주석 방식은, 1단계에서 DECO-LGG기반 SSP 방식의 주석을 수행한 후, 2단계에서 매뉴얼 주석을 수행하는 방식으로 진행된다. 이때 1단계에서 사용되는 언어자원의 구조와 규모에 대해 앞 장에서 기술된 바 있으며, 이 장에서는 이와 같이 진행된 언어 자원 기반 SSP 반자동 증강방식의 성능을 평가한다.

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

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

표 4. 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. 실제로 고빈도로 실현되는 특정 주석 유형의 경우, 작업자의 피로감으로 인해 일관성이 유지되지 못하는 현상들이 나타났고, 굳어진 관용적 MWE 표현들의 경우, 작업자의 언어적 용법에 대한 지식이 부족할 때에는 올바르게 주석되지 못하는 현상들이 관찰되었다. 그런데 이러한 반복적 유형이나 다단어 시퀀스 유형은 SSP 방식으로 처리하는 경우, 최적화된 성능을 보이는 것으로 나타났다.

6. 결론

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 방식으로서, 1단계 작업만으로도 평균 0.915의 높은 주석 성능을 보임을 확인하였다. 이를 통해 FbSA용 학습데이터 주석을 위한 작업자의 작업이 기존 작업의 10% 이하의 비중으로 감소함으로써, 학습데이터 구축을 위한 프로세싱의 소요시간과 품질이 획기적으로 개선될 수 있음을 확인하였다.

현재 본 연구에서 구축한 범용 언어자원과 도메인별 DECO-LGG 사전/패턴문법은, 향후 다양한 영역에서의 FbSA를 위한 머신러닝용 학습데이터를 생성하는 데에 중요한 자원으로 사용될 것으로 기대된다.

참고문헌

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

[2] 남지순. 코퍼스 분석을 위한 한국어 전자사전 구축 방법론. 도서출판 역락. 2018.

[3] Gross, M. The Construction of local grammars. Finite-State language processing. Roche & Schabes(eds.), the MIT Press. 1997.

[4] Wiebe, J., Wilson, T., & Cardie, C. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3), 165-210. 2005.

[5] 김문형, 장하연, 조유미 & 신효필. KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견분석 코퍼스. 한국정보과학회 학술발표논문집 650-652. 2013.

[6] 조동희, 신동혁 & 남지순. MUSE 감성주석코퍼스 구축을 위한 분류 체계 및 태그셋 연구. 우리말연구, 47, 5-47. 2016.

[7] Blitzer, J., Dredze, M. & Pereira, F.C. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. ACL. 2007.

[8] Song, M., Park, H. & Shin, K. Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean. Inf. Process. Manag, 56, 637-653. 2019.

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

[10] Ray, B., Garain, A. & Sarkar, R. An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. Applied Soft Computing. 2021.

[11] Hyun, D., Cho, J. & Yu, H. Building Large-Scale English and Korean Datasets for Aspect-Level Sentiment Analysis in Automotive Domain. COLING. 2020.

[12] Liu, B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. 2012.

[13] 남지순. 오피니언 극성을 전환하는 한국어 부정표현 자동 인식을 위한 연구. 언어와언어학 57, 61-94. 2012.

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

[15] Paumier, S. Unitex Users' Manual. France: UPEM. 2003.

[16] 황창회 & 남지순. DecoLoTA. DICORA-TR-2020-01. 한국외대 디코라연구센터. 2020.