DECO-LGG 언어자원 및 의존파서와 LSTM을 활용한 하이브리드 자질기반 감성분석 플랫폼 DecoFESA 구현

황창회, 유광훈○ & 남지순
한국외국어대학교, DICORA 연구센터/언어인지과학과
hch8357@naver.com, rhkdgns2008@naver.com, jeesun.nam@gmail.com

DecoFESA: A Hybrid Platform for Feature-based Sentiment Analysis Based on DECO-LGG Linguistic Resources with Parser and LSTM

Changhoe Hwang, Gwanghoon Yoo○ & Jeesun Nam
DICORA, Hankuk University of Foreign Studies

요 약

본 연구에서는 한국어 감성분석 성능 향상을 위한 DECO(Dictionnaire Electronique du COreen) 한국어 전자사전과 LGG(Local-Grammar Graph) 패턴문법 기술 프레임에 의존파서 및 LSTM을 적용하는 하이브리드 방법론을 제안하였다. 본 연구에 사용된 DECO-LGG 언어자원을 소개하고, 이에 기반하여 의미 정보를 의존파서(D-PARS)와 페어링하는 한편 OOV(Out Of Vocabulary)의 문제를 LSTM을 통해 해결하여 자질기반 감성분석 결과를 제시하였다. 부트스트랩 방식으로 반복 확장될 수 있는 LGG 언어자원 및 알고리즘을 통해 수행되는 자질기반 감성분석 프로세스는 전용 플랫폼 DecoFESA를 통해 그 범용성을 확장하였다. 실험을 위해서 네이버 쇼핑몰의 '화장품 구매 후기글'을 크롤링하였으며, DecoFESA 플랫폼을 통해 현재 구축된 DECO-LGG 언어자원 기반의 감성분석 성능을 평가하였다. 이를 통해 대용량 언어자원의 구축과 이를 활용하기 위한 어휘 시퀀스 처리 알고리즘의 구현이 보다 정확한 자질기반 감성분석 결과를 제공할 수 있음을 확인하였다.

주제어: 자질기반 감성분석, DECO 전자사전, LGG 언어자원, 의존구문파서, LSTM, DecoFESA 플랫폼

1. 서론

본 연구에서는 한국어 감성분석의 성능 향상을 위해, DECO(Dictionnaire Electronique du COréen) 한국어 전자사전[1]과 LGG(Local-Grammar Graph) 패턴문법 기술[2] 프레임을 통해 의미 처리를 완료한 텍스트에 의존파서 및 LSTM(Long-Short Term Memory)를 적용하는 하이브리드 방법론을 제안하였다. 이를 위해 본 연구에 사용된 DECO-LGG 언어자원을 소개하고, 이에 기반하여 반자동으로 의미 처리되는 데이터와 의존 구문 분석기를 연동하고, OOV(Out Of Vocabulrary)로 나타나는 문장들을 LSTM을 사용하여 처리하는 '자질기반 감성분석(Feature-based Sentiment Analysis: FbSA)' 결과를 제시하였다. 이러한 과정을 부트스트랩 방식으로 반복 확장할 수 있도록 전 과정을 순차적으로 수행하고 그 결과값을 시각화하는 모듈을 제공하는 '감성분석 전용 플랫폼 DecoFESA'를 구현하였다.

실제로 사용자생성문 텍스트에 나타나는 감성표현을 보면, 단일어 형태로 이루어진 형태들 외에도 다단어(Multi-Word Expression: MWE)[3][4] 형태로 실현되는 경우가 매우 빈번하다. 가령 '좋다'와 같은 긍정어휘보다 '마음에 든다'와 같은 긍정 MWE의 출현이 훨씬 빈번한 텍스트 유형을 발견할 수 있는데, 이와 같은 MWE의 경우 일반적인 사전 표제어로 등재하거나 의미적 특징에 의한 자동 조합 또는 유추가 불가능한 '어휘적 특이성(lexical idiosyncrasy)'의 속성을 보이기 때문에, 이들에 대한 별도의 언어자원이 제공되지 않으면 감성분석의 성능을 저해하는 중요한 장애물이 된다[5].

사용자생성문 텍스트에서 나타나는 개체명 및 자질명 또한 다음과 같이 각 도메인별로 MWE의 양상으로 나타나기 때문에 이에 대한 별도의 처리가 요구된다.

(1)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 FbSA를 수행하기 위한 어휘 및 어구의 양상이 다르게 실현된다. 이를 자세히 살펴보면, 스마트폰 리뷰인 (1ㄱ)에서는 스마트폰 제품명에 해당하는 '아이폰 Xs'가 분석 대상인 개체명(named-entity)으로 나타나고, (1ㄴ)는 '뿌링클'이, (1ㄷ)에서는 '보브 실키핏 립스틱'이 2개 이상의 어휘가 결합하는 MWE으로 나타나고 있다. 대상의 고유한 특성을 나타내는 자질명[6] 또한 (1ㄱ)에서는 '화면'이, (1ㄴ)와 (1ㄷ)에서는 '시즈닝'과 '발색'이 각각 자질명으로 나타나고 있다. 이러한 실현 양상에 따라 FbSA를 수행하기 위해서는 코퍼스 도메인에 따른 개체명 및 자질명에 대한 고려가 선행되어야 하며, 범용 극성 표현과 더불어 도메인에 특화된 극성 표현을 감

지할 수 있도록 언어자원이 제공되어야 한다.

본 연구에서는 이러한 단일어 및 MWE를 효과적으로 표상하고 이를 텍스트 분석에 적용할 수 있도록, 유한상태 트랜스듀서(FST) 형식으로 구현되는 LGG 프레임을 사용하였다. LGG는 유니텍스(Unitex) 플랫폼[7]에서 방향성 그래프 형식으로 기술된 후 자동으로 FST로 컴파일되어 텍스트 분석에 적용 가능한 문법이 된다. 이는 순환전이망(recursive transition network) 문법의 하나로, 한국어의 어휘, 통사, 의미 정보 및 극성 정보를 처리하는 기계가독형 활용형사전 DECO를 기반으로 작동한다.

구축된 LGG 패턴문법이 FST문법으로 컴파일되면, 코퍼스에 적용되어 문장의 의미처리를 수행한다. 처리된 정보들은 의존파서의 분석 과정에 활용되어 문장 안에서 나타나는 대상-자질-감성표현을 연결해주는 정보로 활용된다. 이후 감성극성이 분석되지 않는 OOV를 LSTM 알고리즘을 활용한 오피니언 분류기가 처리하게 하여 문장의 최종 극성을 결정할 수 있도록 구성하였다. 이러한 일련의 과정을 도식화하면 아래와 같다.

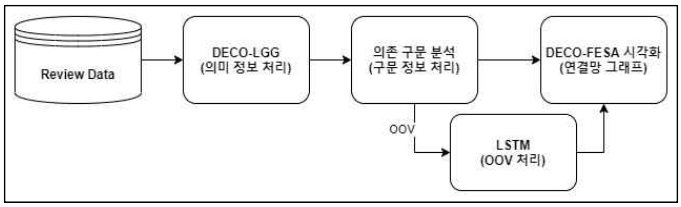

그림 1. DecoFESA 시스템 프로세스

2. 관련 연구

FbSA와 관련된 연구는 자질 어휘 추출 및 긍·부정 극성 정보 분석의 방식의 차이로 분류될 수 있다. 자질 어휘 추출의 방식으로 크게 사전 및 구문 정보에 기반한 방법과 기계학습 방식으로 나눌 수 있다.

사전 및 구문정보 기반 접근법의 장점은 저빈도라도 규칙에 입각하면 자질을 찾을 수 있다는 점이다. [8]에서는 구문 기반 방법과 사전을 활용하여 자질 표현을 추출하였는데 “The button is inconvenient”와 같은 의존 구문 정보를 “Aspect → Dep1→ VB - B VB ∈ {is, are} Dep1 ∈ {nsubj}”와 같이 일반화하고, 이를 특정 상품에 대한 상의어, 하의어 정보를 가진 사전을 활용하여 노이즈를 제거하는 방식으로 자질 추출 과정을 진행하였다.

지도적/비지도적 기계학습 또한 FbSA를 위한 자질 추출 방법론으로 활용되는데, [9]에서는 시퀀스 정보별 레이블에 기반한 기계학습 방법인 조건부 무작위장(Conditional Random Fields)을 활용하여 자질 표현을 추출하는 방법을 제안한 바 있다. CRF의 경우 토큰마다 라벨링을 하는 형태소 태깅과 같은 작업에 주로 사용되는 방법으로, 해당 연구에서는 자질을 추출 및 처리를 위해 ‘B(eginning)/M(iddle)/E(nding)’ 레이블을 수기 주석한 코퍼스를 CRF 모델에 학습시킨다.

비지도적 방법을 사용해 자질 어휘를 추출한 연구로는 [10]이 있다. 해당 연구에서는 레스토랑에 대한 사람들의 의견을 체계적으로 표현한 대규모 사용자 리뷰 모음에서 기능 의견 쌍을 추출하기 위한 ‘RevMiner’를 제안하였다. 이는 자동 부트스트래핑 방법을 사용한 속성-값 추출기(attribute-value extractor)를 기반으로 유사도에 대한 스코어링과 극성 계산을 수행한다. 또한 자질 추출을 위해 이미 설정해 놓은 5개의 클러스터를 정의하고, 해당 관련 자질 어휘를 그룹화하여 인터페이스에 시각화하는 방식을 활용하였다.

앞선 연구들과 같은 다양한 자질 추출 방법론에 기반하여 수행되는 FbSA 연구에 기계학습을 통한 방법론이 사용되는데, 먼저 지도적 학습과 관련된 연구로 [11]가 있다. 해당 연구는 음식점 관련 사용자 리뷰를 학습 데이터로 사용하여, 문장단위 구문분석을 수행함에 따라 얻어진 의존 관계를 토대로 어텐션(attention) 가중치를 할당하는 LSTM 모델을 제안하였다. 여기서는 동일 도메인에 대한 FbSA에서 84.61%의 정확도를 보였다.

비지도적 기계학습에 가까운 FbSA 연구로는 [12]이 있는데, 이는 연속단어 임베딩(continuos word embedding)과 최대 엔트로피 분류기를 결합한 토픽 모델링을 활용하여 FbSA를 수행한다. 해당 연구에서 제시된 모델은 사용자가 입력한 최소한의 토픽 시드 어휘들을 기반으로 도메인 자질 분류/자질-감성어휘 분리/극성 분류를 동시에 수행한다.

한국어에 대한 FbSA 관련 연구 성과는 거의 찾아보기 어려우나, 자질 추출 및 처리를 거치는 감성분석 방법론이 제시된 연구로는 [13]을 꼽을 수 있다. 어텐션 기반 LSTM을 사용하여 자질 어휘 및 감성 표현을 수기 주석한 한국어 화장품 리뷰를 형태소 레벨로 학습시키는 해당 연구는 학습된 모델의 분석 결과를 기반으로 자질-극성 어휘를 페어링하는 FbSA를 진행한다.

위 언급된 FbSA 연구들은 공통적으로 오피니언 대상 표현에 해당하는 개체명 정보가 메타데이터 등의 형태로 고정되거나, 평가 대상을 자질 어휘에 한정하는 방식으로 진행되었다. 따라서 텍스트에 존재하는 평가 대상으로서의 개체명 처리가 FbSA 과정에서 배제되었으며, 오로지 수기로 주석된 텍스트 데이터 내의 한정된 표현들만이 학습에 포함된다는 한계점이 있었다.

본 연구는 사전 및 구문 패턴을 기반으로 특정 도메인 코퍼스에서 나타나는 감성분석 관련 요소들을 반자동으로 감지하고 주석하는 언어자원을 통해 텍스트가 평가하고자 하는 개체명 대상을 인식할 수 있도록 구성된다. 자질 어휘의 경우, 범용 사전을 통해 분석된 결과와 더불어 빈도 기반으로 추출된 다양한 자질 어휘들을 의미적 속성에 따라 범주화하여 자질-카테고리 기반의 FbSA를 수행될 수 있도록 도메인 사전을 구축하였다. 텍스트에서 나타나는 감성분석 핵심성분들의 통사적 관계에 따라 오피니언 핵심요소들을 페어링하기 위한 구문 분석 과정에는 의존 파서(parser)가 활용되었으며, OOV가 발생한 문장의 경우 LSTM 모델이 극성 분류를 수행하도록 구성하였다.

3. 자질기반 감성분석과 DECO-LGG

특정 도메인의 사용자 생성문에서 나타나는 평가대상, 자질 어휘, 감성표현을 처리하기 위한 코퍼스 주석 언어

자원을 구축하기 위해 본 연구에서는 DECO 한국어 전자사전을 활용하였다. DECO 사전에 수록되어 있는 30만여 개의 표제어에는 기본적인 형태-통사 정보들과 더불어 다양한 의미-개체명-도메인-감성 정보 등이 할당되어 있다. 실제 DECO 사전에 수록되어 있는 표제어의 예를 들면, '갤럭시.NS01+..+XXPR+XITP+..'나 '화질.NS03+..+ XQFT+..', '예쁘다.AS16+..+QXPO+…' 등과 같다.

여기서 개체명과 관련된 정보는 <XX-AA> 형식의 태그 형태로 할당되며, 자질 어휘와 관련된 정보는 <XQFT> 태그로, 도메인 관련 정보는 <X-AAA> 형식으로 부착되어 있다. <QX-AA> 태그는 극성을 나타내는 태그로서, 이들은 각각 DECO 사전에서 사용한 EntLEX 개체명분류 체계와 DomLEX 도메인분류 체계, 그리고 PolLEX 극성분류체계에 의해 분류된 정보를 나타낸다.

3.1. 평가대상(Target)과 자질(Aspect) 어휘

본 연구에서는 도메인 의존적인 FbSA 연구를 수행하기 위하여 화장품 리뷰 텍스트를 그 대상 도메인으로 선정하였다. 여기서 평가대상 및 자질 어휘를 주석 처리하기 위한 언어자원을 구축하기 위해, DECO 사전에서 사용하는 Ent/DomLEX 분류 표제어를 활용하였다. EntLEX 분류체계는 11가지의 개체명 대분류를 포함하는데, 우선 화장품 브랜드를 포함하는 <XXOR> 태그와 화장품 개별제품명을 포함하는 <XXPR> 태그 내 어휘들을 검토하였다.

그러나 실제 코퍼스에서 나타나는 도메인 의존적인 다단어 표현들을 기술하기 위해서, 색인된 코퍼스 용례를 통해 분석대상 어휘를 추출하여 이를 바탕으로 MWE 개체명 및 자질 어휘 주석 처리를 위한 도메인 사전과 패턴문법을 추가로 구성하였다. 이렇게 구축된 도메인 언어자원을 토대로, 코퍼스 수집시 관찰된 686개의 개체명 메타정보를 분석하고 MWE 화장품 개체명의 결합패턴을 정형화하였다. 메타정보를 통해 습득한 MWE 개체명 구성어휘는 각각의 개체명 구성어휘가 실현 가능한 위치와 의미에 따라 분류되어 도메인 사전에 추가되었다. 이러한 과정을 바탕으로 구축된 MWE 화장품 개체명 패턴그래프는 그림 2와 같은 방식으로 구성되었다.

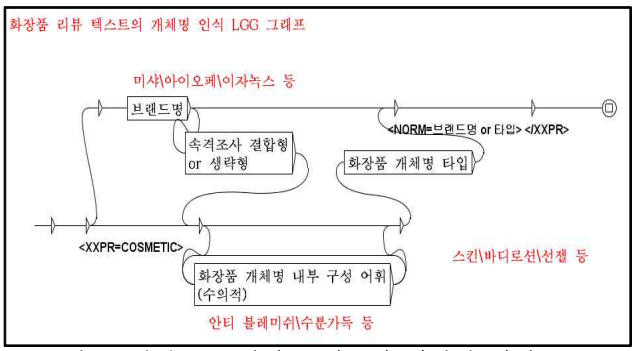

그림 2. 화장품 도메인 코퍼스의 개체명 인식 LGG

그림 2의 개체명 인식 LGG는 해당 도메인에서 나타나는 개체명의 결합적 특징을 구조화하여 기술하고 있어, 단일어로 나타날 수 있는 개체명 유형과의 불필요한 중복을 피할 수 있도록 하였다. 즉 조사 및 다른 범주의 어휘가 나타날 때까지 최장결합의 형태를 제한하지 않는 방식으로 구성된다. LGG는 FST문법으로 컴파일되어 코퍼스 분석에 적용되고, 이때 이렇게 인식된 단일어/다단어 개체명에는 XML 형식의 화장품 개체명 태그가 할당되며, 이를 통해 추후 FbSA 프로세스에서 해당 표현을 개체명으로 인식할 수 있게 한다.

자질 어휘 또한 개체명과 마찬가지로 DECO사전에 등재된 <XQFT> 태그를 통해 표제어를 확보하였으며, 코퍼스에서 나타나는 일반명사들의 빈도를 기술통계적으로 분석하여 그 외연을 확장하였다. 현재 도메인에서는 '가격/성분/향기/용량/보습력'과 같은 어휘들이 이러한 부류에 속한다. 그런데 이와는 반대로 자질어가 실현되지 않고 암시적(implicit) 자질이 함축되는 경우들이 있는데, 다음을 보자.

(2)ㄱ. 그래도 생각보다는 저렴한 편이에요. [가격]
ㄴ. 앰플이 싼 게 그 모양이지 뭐 [가격]
ㄷ. 촉촉한 타입이예요. [수분감]

위의 예문에 나타난 극성 형용사들은 자질어와 공기하지 않았으나 암시적인 자질어를 함축하고 있다. 이러한 암시적 자질어를 인식하기 위하여 코퍼스에서 명시적 자질어와 공기하여 나타나는 극성 형용사들을 분석한 후, 이들을 다시 문서단어행렬(DTM)을 통해 빈도분석하여, 유의미한 '명시적 자질어-암시적 자질어' 관계를 추출하는 이중 증식(double propagation) 방법론이 활용되었다. 주로 극성 어휘에 부여되는 암시적 자질어는 추후 FbSA 처리 과정 상에서 문장에 명시적 자질어가 실현되지 않은 경우 자질어로 기능하게 하였다.

3.2. 단일어/다단어 감성표현(Sentiment Expression)

도메인 코퍼스에서 나타나는 단일어와 다단어 감성표현 인식을 위해 DECO 사전의 감성어휘 분류체계의 하나인 PolLEX 극성어휘 분류체계를 활용하였다. PolLEX 분류체계에서 극성어휘는 표 1과 같이 4가지 유형으로 분류되어 있다. 이에 더하여, 특정 도메인에서만 의미를 갖는 도메인 의존적 감성표현을 기술하기 위해 무극성 어휘의 빈도정보 및 연어 관계 등을 추출 및 분석하여 이를 자원화하였다. 가령 '마음에 들다(+)'같은 범용의 MWE 외에 '수분감이 있다(+)'와 같은 도메인 의존적 MWE 유형이 기술되었다.

표 1. DECO-PolaLEX 극성어휘 분포

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

이러한 극성 표현들은 문장 내에서 일련의 극성 전환어(Polarity Shifting Device: PSD)에 의해 기저 극성이 반대로 전환될 수 있다. PSD에는 장/단형 부정문을 구성하는 부정소와 문맥 강화/약화 부사 유형이 포함되며, 이들은 각각 그림 3과 그림 4와 같은 방식으로 기술되었다. 해당 경로에 따라 감지된 극성 표현들은 의미 처리 시에 반대의 극성으로 변환되어 마크업되도록 기술되어 있다.

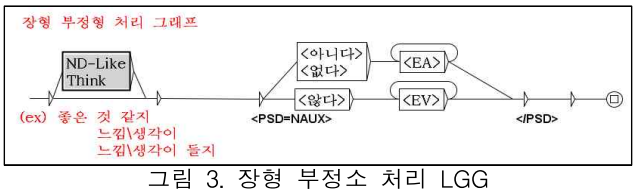

그림 3. 장형 부정소 처리 LGG

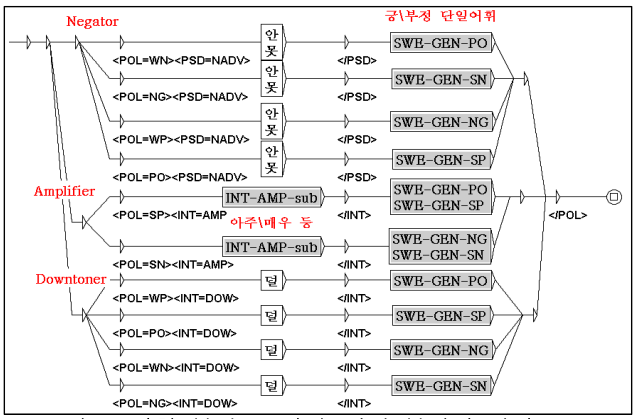

그림 4. 단형 부정소, 강화&약화 부사어 처리 LGG

4. DECO-LGG와 의존파서 기반 감성분석

본 장에서는 DECO-LGG을 통해 제공된 평가대상/자질/극성 어휘를 문장 내 구문의 의존 관계를 처리하여 감성 분석을 진행한다. 본 연구에서는 KAIST의 '한나눔 의존 구문 분석기'를 사용하였다. 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. 이러한 장점을 활용하기 위해 의존 구문분석기에 기반하여 문장 단위에서 나타나는 대상/자질 어휘 페어링과 구문 깊이(depth)에 따른 극성어휘 가중치 처리를 진행하였다.

4.1 의존파서 기반 감성 및 자질/대상어휘 페어링

위에서 소개한 DECO-LGG에서 평가대상/자질/극성 어휘들에 대한 의미정보를 처리한 후, 의존 구문분석을 통해 문장 단위 내에서 핵이 되는 서술어를 선별한다. 이후 이와 관계된 의존 구문 맵핑 및 문장성분을 분석하여 평가대상 및 자질 어휘를 페어링한다. 해당 과정은 입력 문장에 DECO-LGG를 적용하여 FbSA의 핵심 의미정보인 "평가대상(XXPR)", "자질(XQFT)", "극성(QX--)" 정보를 처리한 후, 이를 구문분석을 통해 문장성분과 핵심 술어 간의 거리, 즉 최상단으로 설정된 핵심 술어와 문장 구성요소 사이의 깊이를 분석함에 따라 서술어의 필수 논항을 제외한 부가어(adjunct)의 의미 정보를 탈락시키기 위한 문장단위 대상/자질 정보 처리 과정으로, 이를 요약하면 그림 5와 같다.

필수논항과 부가어의 구문 의존관계 구별을 위해선, 핵어의 직접지배를 받는 구문의 의존관계에 깊이우선 탐색(depth-first search) 알고리즘을 적용하였다. 이를 통해 대상/자질 어휘가 하나의 문장 내에서 다수 출현할 때, 여러 구문의 수식을 받는 부사구를 탈락 처리하여 필수논항 내 의미 정보를 우선적으로 처리할 수 있게 된다.

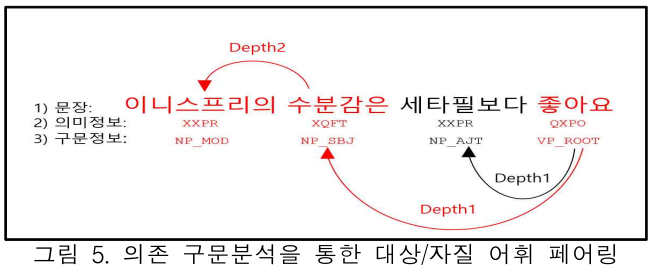

그림 5. 의존 구문분석을 통한 대상/자질 어휘 페어링

4.2 의존파서 기반 극성어 가중치 처리

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

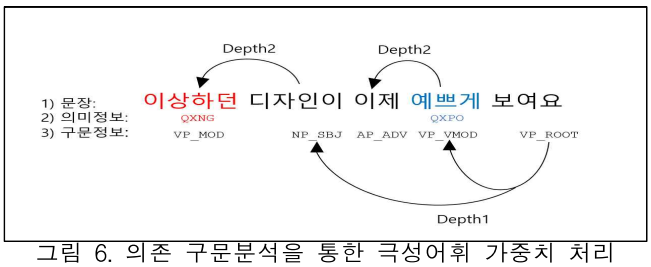

그림 6. 의존 구문분석을 통한 극성어휘 가중치 처리

그림 6은 서로 다른 극성어휘가 공기하는 문장에 대한 처리 예시이다. 부정 어휘와 긍정 어휘가 동시에 존재하는 해당 문장은 긍정의미 정보(QXPO)가 부여된 "예쁘게"의 깊이가 핵어와 더 가까움에 따라 문장의 주어(NP_SBJ)를 수식하는 부정극성(QXNG) 관형어 "이상하던"보다 더 높은 가중치를 할당 받는다. 이때 가중치는 어휘의 극성점수를 깊이 분의 1로 곱한 값으로 계산된다.

5. LSTM 기반 감성분석

DECO-LGG의 처리 범주를 벗어난 OOV 문제를 해결하기 위해 시퀀스 정보를 처리하는 LSTM을 활용하였다. 단순 RNN(Recurrent Neural Network) 구조의 경우 시퀀스가 길어지게 될수록 가중치의 값이 특정 값에 수렴하게 되는데, 이는 시간이 지남에 따라 이전에 입력된 데이터의 영향력이 감소하는 장기의존성 문제로 수렴된다. 이러한 단순 RNN 모델의 한계점을 보완하기 위해 제안된 LSTM은 망각(forget), 입력(input), 출력 게이트(output gate)를 통해 필요맥락의 정보를 선별함에 따라 장기의존성 문제를 효과적으로 해결하는데[14], 각각의 게이트를 열고 닫음으로써 초기 입력된 데이터의 손실 문제를 최소화하는 방식이다.

그림 7에서 보이는 바와 같이, LSTM 구조 상에서 데이터는 망각(F), 입력(I), 출력(O) 게이트의 순서대로 통과하며, 각 게이트를 통과한 데이터들은 셀 상태(cell state)를 통해 저장된다. 각 단계를 요약하면, 먼저 망각 게이트에서는 이전 시점의 정보를 잊어 버릴지를 계산하고, 그다음 시그모이드(sigmoid)층과 하이퍼볼릭 탄젠트(tanh)층의 두 부분으로 구성된 입력 게이트를 통과하는

데, 이를 통해 계산된 값은 최종적으로 망각 단계를 통과한 값과 연산이 되어 저장된다. 이러한 연산 결과는 출력 게이트를 거침에 따라 이후 위치한 메모리 블록(memory block)에 전달된다.

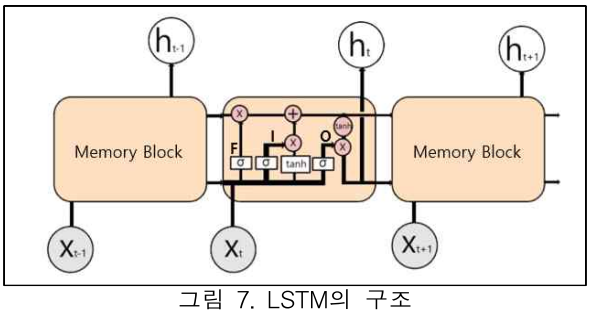

그림 7. LSTM의 구조

본 연구에서는 OOV가 발생한 문장의 극성을 분류하기 위해 LSTM 알고리즘을 통해 학습된 모델을 활용하였으며, 이때 평가 대상에 해당하는 개체명/자질 어휘는 오피니언 문장의 모문 서술어를 감성표현의 핵심 술어로 간주하여 의존 파싱한 결과에 의해 결정된다.

6. 실험 및 성능평가

6.1 실험 데이터

평가를 위한 데이터로 본 연구에서는 한국외국어대학교 DICORA 연구센터의 키워드기반 후기글 크롤러 DECOR-Cralwer를 사용하여 네이버 쇼핑몰에 등록되어있는 화장품 리뷰 총 150만개를 수집하였으며, 수집된 데이터 중 31,000개의 리뷰를 별점을 토대로 레이블링한 후 LSTM의 학습 데이터로 입력하였다. 입력된 텍스트 데이터의 10%를 검증 데이터로 구성하여 학습 모델을 자체 평가한 결과, 90.6%의 정확도(accuracy)를 보였다.

테스트 데이터는 총 300개 문장 규모로 구성되었으며, 해당 문장에 작업자가 수동 태그를 붙이는 방식으로 정답 문서를 구성하였다. 평가문의 예시에는 하나의 문장 내에 많은 양의 개체명 표현들이 포함되어있는 경우들을 비롯하여 극성 MWE가 나타나는 경우들이 포함되었으며, 이에 따라 의존파서를 활용한 대상 개체명/자질명에 대한 평가에 적절한 문장들이 (3)과 같이 구성되었다.

(3)ㄱ. 다른 T[썬크림]들은 눈이 시린데 T[비레머디스] 제품은 피부에 N(자극 없)어서 P(너무 좋)아요
ㄴ. T[랑콤]꺼 쓰다가 제 기준 P(더 좋)은 T[시세이도]로 갈아탔습니다.
ㄷ. T[밀티 클렌저] 완전 P(촉촉해요_F[수분감]).

위 예문에서와 같이, FbSA 처리 결과는 문장내 평가 대상(T)/자질(F)/극성(P) 어휘 정보들이 복합적으로 처리되어야 한다.[1] 또한 (3ㄷ)은 "촉촉하다"와 같은 극성 표현이 문장 내 명시적으로 드러나지 않은 자질 정보 "수분감"을 내포한다는 점에서 이러한 정보를 처리하기 위한 기술이 요구된다. 본 연구에서는 이를 앞서 기술한 DECO-LGG 자원을 통해 처리함에 따라 DecoFESA 플랫폼에서 분석할 수 있도록 설정하였다.

1) DECO-LGG에서 처리되는 대상/자질어를 대괄호로, 극성어는 소괄호로 표현하였다.

6.2 실험을 위한 DECO-LGG 자원

실험 데이터의 평가에 직접적으로 적용된 LGG 언어자원은 그림 8과 같은 메인 그래프를 통해 호출 및 적용되도록 구성되었으며, 자원의 하위 구성은 총 190여개의 서브그래프를 통해 구분되었다.

본 연구에서는 총 658개의 브랜드에서 나타나는 화장품 개체명 MWE를 모두 패턴의 형태로 인식할 수 있도록 구성하였으며, 총 60여개의 화장품 도메인 자질명과 그 변이형에 대한 형태들을 처리하고 정규화하도록 구성하였다. 극성어휘 처리와 관련해서 구축된 LGG는 13,000여개의 단일어 극성어휘와 2,700여개의 범용/도메인 MWE들을 모두 처리할 수 있도록 하였다. 이에 더하여 본 연구에서 구축한 자원은 극성표현의 변이형 및 이와 공기하는 부정표현과 강화/약화 부사들을 처리할 수 있게 디자인되었기 때문에, 실제 처리할 수 있는 패턴의 수는 단순 타입 수에 비해 더욱 확장될 것으로 판단된다.

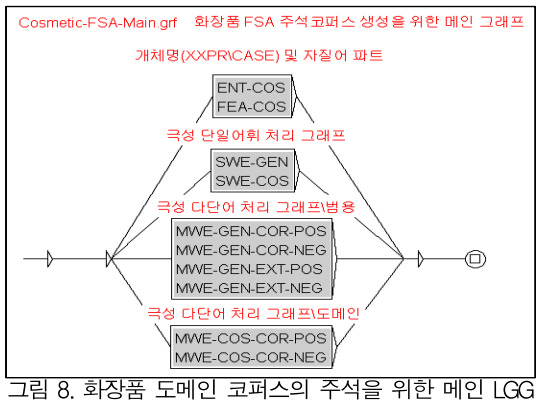

그림 8. 화장품 도메인 코퍼스의 주석을 위한 메인 LGG

6.3 실험 결과

DecoFESA 시스템을 통해 처리된 성능 결과는 표 2에서 보이는 바와 같다. 부정 문장을 처리한 f-measure의 경우 0.83, 긍정 문장의 경우 0.99로 준수한 성능을 보여주었으며, 평가 대상어/자질어의 경우 각각 0.75과 0.88의 정확도를 보였다.

표 2. DecoFESA 실험 결과

Polarity	Precision	Recall	f-measure
NEGATIVE	0.76	0.90	**0.83**
POSITIVE	0.99	0.98	**0.99**
Accuracy			
대상어	**0.76**		
자질어	**0.88**		

여기서 분석에 실패하는 경우는 다음 (4)처럼 비정형 텍스트의 특징이 잘 나타나는 문장이 주를 이루었다.

(4)ㄱ. 배송이 맘에안들어요
ㄴ. (ROOT → VNP_ROOT_배송이맘에안들어요)

이에 따라, 비정형 텍스트 데이터에 대한 FbSA의 수행에 앞서, 이러한 문제를 해결하기 위한 구체적인 전처리 방안이 프로세스 내부에 고려되어야 함을 알 수 있었다.

7. DecoFESA 플랫폼과 감성분석 결과의 시각화

감성분석의 경우 대용량의 오피니언 문서에 나타나는 감성 정보를 효과적으로 요약하는 시각화가 중요하다. DecoFESA 플랫폼에서는 평가대상에 따른 자질기반 극성 정보를 연결망 그래프(network graph) 형식으로 제공한다. 그림 9는 평가대상 "크리니크"에 대한 자질인 "성분/향/효과"에 대한 극성 정보를 효과적으로 보여주고 있다. 자질 어휘가 문장에 드러나지 않는 경우에는 자질명이 "GEN"으로 지정되어 극성정보를 입력받는다.

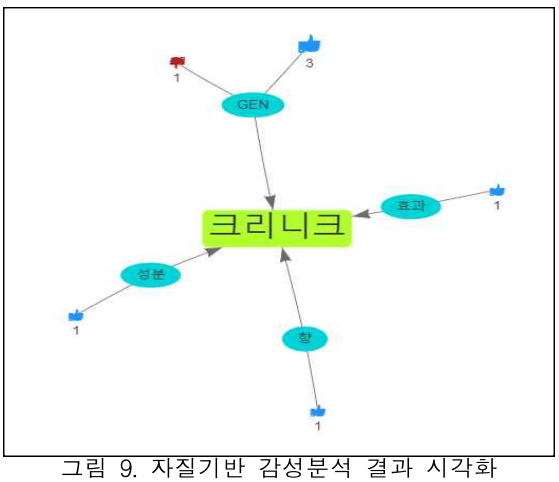

그림 9. 자질기반 감성분석 결과 시각화

8. 결론

본 연구에서는 평가대상 및 자질 어휘, 감성 극성 관련 의미 정보를 반자동으로 처리하는 DECO-LGG 언어자원과, 처리된 의미 정보 및 문장 성분 간의 의존관계를 바탕으로 평가대상/자질/감성표현을 페어링하는 감성분석 알고리즘을 제시하였다. 또한 OOV 문제를 LSTM을 통해 개선하는 하이브리드 FbSA 방법론 기반 DecoFESA 플랫폼을 구현하여 분석 결과를 가시적으로 제공하였다.

본 연구에서 제안하는 접근법은 점진적이고 명시적인 부트스트랩핑 방식의 처리시스템의 장점을 지님과 동시에 딥러닝 기술이 적용됨에 따라 일반화처리에서의 강점을 지니고 있다. 이러한 방법론은 통합적 감성분석 시스템 DecoFESA 구현을 통해 손쉽게 FbSA의 전 과정을 수행할 수 있게 하였으며, 분석 결과를 연결망 그래프의 형식으로 요약하여 제시해준다는 점에서 큰 실용성을 지닌다.

해당 방법론을 발전시켜 보다 효과적인 FbSA를 수행하기 위해서는 향후 타 도메인에서 나타나는 단일어/다단어 오피니언의 의미 정보를 처리하기 위한 언어자원이 구축되어야 할 것이며, 이를 기계학습의 학습 데이터 및 감성분석 처리에 활용하기 위한 연구가 지속되어야 할 것으로 보인다. 또한 검증된 데이터셋에 기반한 추가적인 평가를 통해 공신력을 확충해나가야 할 것이다.

참고문헌

[1] 남지순. 코퍼스 분석을 위한 한국어 전자사전 구축방법론. 도서출판 역락 (2018)

[2] Gross, M.. The Construction of Local Grammars. Finite-State Language Processing, The MIT Press (1997)

[3] 김한샘. 자연언어처리를 위한 관용표현 연구. 한국어의미학, 13, 43-68 (2003)

[4] 최석재. 어휘의 부류와 감정 표현 관용구의 의미. 한국어학, 55, 367-395 (2012)

[5] Piao, S., Rayson, P., Archer, D., Wilson, A. & McEnery, T., "Extracting multiword expressions with a semantic tagger", In Proceedings of the ACL Workshop on MWEs, 49-56 (2003)

[6] Liu, B., Sentiment analysis: mining opinions, sentiments, and emotion, Cambridge University Press (2015)

[7] Paumier, S.. Unitex Users' Manual. UPEM (2003)

[8] Mirtalaie, M, A., Hussain, O, K., Chang, E., Hussain & F, K.. 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 (2018)

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

[10] Huang, J,. Etzioni, O., Zettlemoyer, F., Clark, K. & Lee, C.. RevMiner: An Extractive Interface for Navigating Reviews on a Smartphone. Proceedings of the 25th annual ACM symposium on User interface software and technology (2012)

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

[12] García-Pablos, A., Cuadros, M. & Rigau, G.. W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis.Expert Systems with Applications, 91, 127-137 (2018)

[13] Song, M., Park, H. & Shin, K.. Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean. Information Processing & Management, 56(3), 637-653 (2019)

[14] 손진광. RNN LSTM과 ACO를 이용한 감성 분석을 통한 컨텐츠 추천 시스템에 관한 연구. 한국정보과학회 학술발표논문집, 1033-1035 (2017)