화장품 후기글의 자질기반 감성분석을 위한 다단어 표현의 유한그래프 사전 및 문법 구축

황창회○, 유광훈, 최성용, 신동혁, 남지순
한국외국어대학교 언어어인지과학과, 디지털언어지식콘텐츠연구센터(DICORA)
hch8357@naver.com, rhkdgns2008@naver.com, csy@hufs.ac.kr, sdh876@hanmail.net, namjs@hufs.ac.kr

Building Korean Multi-word Expression Lexicons and Grammars Represented by Finite-State Graphs for FbSA of Cosmetic Reviews

Chang-Hoe Hwang○, Gwang-Hoon Yoo, Seong-Yong Choi, Dong-Heouk Shin, Jee-Sun Nam
DICORA, Department of Linguistics and Cognitive Science, Hankuk University of Foreign Studies

요 약

본 연구는 한국어 화장품 리뷰 코퍼스의 자질기반 감성 분석을 위하여, 이 도메인에서 실현되는 중요한 다단어 표현(MWE)의 유한상태 그래프 사전과 문법을 구축하는 방법론을 제시하고, 실제 구축된 사전과 문법의 성능을 평가하는 것을 목표로 한다. 본 연구에서는 자연어처리(NLP)에서 중요한 화두로 논의되어 온 MWE의 어휘-통사적 특징을 부분문법 그래프(LGG)로 형식화하였다. 화장품 리뷰 코퍼스에 DECO 한국어 전자사전을 적용하여 어휘 빈도 통계를 획득하고 이에 대한 언어학적 분석을 통해 극성 MWE(Polarity-MWE)와 화제 MWE(Topic MWE)의 전체 네 가지 하위 범주를 분류하였다. 또한 각 모듈 간의 상호관계에 대한 어휘-통사적 속성을 반복적으로 적용하는 이중 증식(double-propagation)을 통해 자원을 확장하였다. 이 과정을 통해 구축된 대용량 MWE 유한그래프 사전 DECO-MWE의 성능을 테스트한 결과 각각 0.844(Pol-MWE), 0.742(Top-MWE)의 조화평균을 보였다. 이를 통해 본 연구에서 제안하는 MWE 언어자원 구축 방법론이 다양한 도메인에서 활용될 수 있고 향후 자질기반 감성 분석에 중요한 자원이 될 것임을 확인하였다.

주제어: 한국어 다단어 표현, 자질기반 감성분석, 유한그래프 사전과 문법. 토픽MWE. 감성MWE

1. 서론

본 연구는 화장품 후기글에 대한 자질기반 감성분석(Feature-based Sentiment Analysis; FbSA)에 활용하기 위한 한국어 다단어표현(Multiword Expression; MWE)의 사전 및 문법 자원을 구축하는 방법론을 제시하고, 실제 구축된 언어자원의 성능을 평가하는 것을 목표로 수행되었다. 본 연구의 MWE 사전과 문법은 '부분문법 그래프(Local-Grammar Graph; LGG)' [1]로 명명되는 순환 전이망(recursive transition network) 형식으로 구축되었다. 본 연구에서 구축된 LGG는 한국어의 어휘, 통사, 의미 정보 및 극성 정보(polarity information)와 활용형 정보를 제공하는 기계가독형 사전 전자사전 DECO[2]에 기반하여 MWE를 유한상태 트랜스듀서(Finite-State Transducer; FST)로 기술된다.

감성분석, 또는 오피니언 마이닝은 현대 SNS 플랫폼에서 관찰되는 다양한 사용자 생성문(user-generated sentence)을 분석하여, 어떤 이슈 또는 제품에 대한 대중의 주관적 의견을 파악하고 트랜드를 예측하기 위한 연구로, 이를 위해서는 분석 단위문의 감성 지향성(sentiment orientation), 일반적으로 긍정·부정의 극성 방향을 포착하는 것이 중요하다. 문제는 이러한 감성표현이 단일어가 아닌 다양한 유형의 다단어표현, 또는 관용구 등으로 실현되는 경우이다. 더욱이 자질기반 감성분석과 같이 평가대상과 그 자질에 대한 추출이 요구되는 경우, 이러한 대상어도 단일어로 실현되지 않고 여러 단어로 구성된 MWE로 실현된다는 점이 문제가 된다. 이러한 MWE를 올바르게 인식하지 못하는 경우, 감성분석 시스템의 신뢰도에 큰 영향을 미치기 때문이다.

MWE에 대한 엄밀한 정의는 쉽지 않은데, 일반적으로 다단어표현은 기존의 '관용표현, 복합구, 연어' 등에 대한 언어학적 정의들을 포함하는 중립적인 용어로 사용된다. 좁게는 여러 개의 어휘의 결합이 각 원소들의 의미 합으로 유추할 수 있는 일종의 합성성(compositionality) 원리에서 벗어나 하나의 독립적인 의미로 실현되는 표현들을 지칭하지만, 넓게는 빈번하게 함께 실현되는 일정 유형의 공기 관계의 어휘들의 연쇄를 포함한다. 다음 예를 보자.

(1) ㄱ. 이 스마트폰이 **마음에 차더라구요**.
ㄴ. **수분감이 살아있는** 키엘 크림.
ㄷ. **미쟝센 슈퍼 매트 왁스**
ㄹ. 이 팩트의 화사한 **컬러 밝기**가 최고예요.

(1ㄱ)의 '마음에 차더라구요'와 같은 표현은 구성 어휘들의 의미 결합을 통해 문장의 의미를 판별해낼 수 없는 관용 표현(idiomatic expression)의 일종으로, 감성단어 기반으로 문장의 극성을 분석하는 시스템에서는 이

러한 MWE의 극성을 포착할 수 없다. 이러한 표현들은 '마음에 들다/ (-이) 통하다/ (-에) 두다/ (-에) 붙다' 등과 같은 일련의 변이 형태로 실현될 수 있을뿐더러, 여기 결합하는 용언의 어미 활용이 다양하게 나타나기 때문에 실제로는 매우 다양한 표면 형태로 실현될 가능성이 높다. (1ㄱ)의 예가 일반적인 '만족감'을 표현하는 관용적 MWE의 예를 보인다면, (1ㄴ)의 예는 특정 도메인에 종속적인 극성 표현의 예를 보인다. 즉 화장품 후기글 코퍼스에서 '수분감이 살아있다'는 긍정적 오피니언을 표현한다. 따라서 이러한 부류는 (1ㄱ)과는 별도의 모듈에서 고려될 필요가 있다. (1ㄷ)에서 나타나는 문제는 외래어의 한글 표기에 있어서 언중 지식이 통일되어 있지 않기 때문에 더 어려운 문제가 된다. 화장품 후기글과 같은 텍스트 장르에서 평가의 대상이 되는 개체명(Named Entity)은 이와 같이 여러개의 명사가 연결된 MWE 형태를 보이는 경우가 매우 빈번한데, 여기서 외래어 표기의 변이 현상으로 인해 '미장센 슈퍼 매트 왁스', '미쟝센 수퍼 매트 왁스' 등과 같이 그 표기법이 더 다양해지는 문제가 발생한다. (1ㄹ)의 예는 자질기반 감성분석에서 추출해야 하는 자질(Feature) 표현의 경우에도 한 단어가 아닌 '컬러 밝기'와 같은 MWE의 형태로 실현되는 경우를 보인다. 이 경우도 외래어 사용 비중이 높게 나타나기 때문에 '컬러'는 '칼라' 등으로 표현되거나 또는 유의어 '색', '색감' 등으로 실현됨으로써 궁극적으로 보다 다양한 MWE를 구성하게 되는 문제를 보인다.

본고에서는 이러한 MWE들이 몇 개의 규칙으로 기술되거나 유추될 수 없는 '어휘특이성(Idiosyncrasy)의 현상을 보인다고 판단하여 이들을 체계적으로 기술하기 위한 MWE 사전과 문법을 구성하는 것을 목표로 한다. MWE의 사전 기반 처리 방법론은 일반적으로 해당 MWE를 단일 어형만을 제한적으로 인식하는 방법이 주를 이루었는데, 본 연구에서 제시하는 방법론은 그 내부에서 나타나는 부사 수식어구 삽입 및 조사 어미 결합에 따른 많은 변이형 등의 문법적 속성을 고려하여 이를 LGG 유한 그래프로 구축하여, MWE의 활용형들에 대한 기본형(Lemma)들을 제공하여 보다 효율적인 자질기반 감성분석(FbSA)이 이루어지도록 한다는 점에서 기존 연구들과는 차별성을 가진다.

본 연구에서는 MWE를 감성 단가(Polarity)를 가진 MWE(Polarity MWE; Pol-MWE)와, 명사의 연쇄 결합으로 나타나는 문장 내 화제(Topic) 관련 MWE(Topic MWE; Top-MWE)로 분류하였으며, 이를 다시 '일반 감성 MWE(General Pol-MWE; GMWE)', '도메인의존 감성 MWE(Domain-Dependant Pol-MWE; DMWE)', '개체명 MWE(Name-Entity Top-MWE; EMWE)' 및 '자질 MWE (Feature Top-MWE; FMWE)'로 하위분류하였다. 이와 같이 구축되는 사전은 DECO-MWE로 명명된다.

2. 관련 연구

본 연구에서 MWE는 합성성 원리가 미약하게나마 남아있더라도, 관습적인 측면에서 하나의 단위로써 굳어져 사용되는 정형화된 표현(formulaic sequence)[3]들을 포함한다.[1] 이러한 표현들 또한 처리 효용성의 관점에서 하나의 단위로 인지되는 것이 바람직하기 때문이다.

실제로 많은 MWE가 그것을 구성하는 개별 어휘의 의미들의 합으로 유추되지 않는다는 점에서 감성분석에 어려움을 제기한다. [4]는 MWE에 대한 처리가 수행되지 않는다면 이러한 문제가 언어의 파생 및 활용 양상에 따라 점점 더 확장되어 텍스트 의미 층위에서 심각한 문제가 야기될 수 있음을 지적하였지만, 한국어 MWE를 처리를 위한 언어학적 시도에는 더 많은 연구가 필요하다.

극성 관련 MWE 처리에 대한 감성분석 연구로, [5]에서는 영어 텍스트에 대해서 그들이 구축한 분류기인 SO-CAL(Semantic Orientation Calculator)를 이용하여 감성분석을 수행하였다. 하지만 MWE에 있어서 152개의 구동사(phrasal verb) 및 35개의 강조어 MWE만을 언급하였으며, 사전에 해당 MWE에 관한 정보를 수동으로 넣어주어야 한다는 점에서 제한점을 보였다.

명사 연쇄 형태의 MWE에 관한 연구에서는, 가령 [6]에서는 두 개 이상의 명사가 결합되어 있는 형태의 합성명사들을 MWE로 분류하였다. 이러한 합성 명사들은 FbSA의 개체명 및 자질의 형태로 코퍼스 내에서 나타나지만, 일반적으로 복합 명사의 연쇄를 개체명에 한정하여 살펴본 경우들을 제외하고는, 개체명 및 자질을 MWE로 분류하여 자원을 구성한 연구는 찾아보기 힘들다.

[7]에서는 21,235개의 감성 단가를 가진 어휘 중 코퍼스에서 나타나는 2,010개의 단일 어휘와, 단일 어휘 상 감성이 나타나지 않지만 맥락에 따라 극성을 나타내는 50개의 중립 어휘 및 50개의 화제 관련 어휘들을 시드(seed) 어휘로 활용하여 이를 확장함으로써 다단어 표현 추출을 수행하였다. 추출된 총 3,193개의 MWE는 코퍼스 적용시 약 60%의 정확성(precision)을 보였다. 이는 감성 단가를 계산하기 위한 극성 MWE에 국한된 연구이다.

본 연구에서는 감성분석의 신뢰도를 높이기 위하여 MWE를 인식할 수 있는 본격적인 언어자원의 구축이 중요하다고 판단하였고, 이를 위해 다음과 같은 방법론을 제안하였다.

3. MWE 사전 구축을 위한 연구 방법론

3.1 데이터 수집

본 연구에서는 화장품 리뷰 사이트인 '파우더룸'[2]의 후기글을 주요 데이터로 선정 및 수집하였다. 여기에는 한국의 화장품 시장이 급속하게 발전하여 2015년 이후 세계 10위권에 올라섰다는 점에서 시대적인 의의도 포함되었지만[8], 화장품 후기글에 특히 다양한 유형의 MWE 표현이 실현된다는 점에서 실제 실험을 위한 코퍼스 도메인으로 선정될 필요성이 대두되었다. 수집된 데이터는 총 796,689개의 토큰으로 이루어져 있으며, 총 56,354개의 문장으로 구성된다.

1) MWE의 범주 구분에 관한 매우 다양한 논의들이 존재하지만, 본고의 목적이 이러한 표현들을 FbSA에 활용하는 데에 있으므로, 잠정적이며 포괄적인 견해를 따랐다.

2) https://www.powderroom.co.kr

3.2 MWE 추출 및 처리의 방법론

대용량 MWE 언어자원 DECO-MWE는 정교한 사전 정보에 바탕하여 구축되는데, 해당 자원이 감성 분석 중에서도 가장 세밀한 단위의 분석이라 할 수 있는 FbSA를 위한 자원이기 때문이다. FbSA는 문서 및 문장을 대상으로 한 극성 분류 및 분석을 넘어 문장을 구성하는 어휘 단위에 할당되어 있는 자질(feature)을 대상으로 감성분석을 수행한다. [9]는 FbSA로 분석되어야 할 가장 중요한 5가지 요소를 평가 대상 개체(the name of an entity, e), 평가 대상에 대한 자질/속성(aspect, a), 극성 점수(sentiment, s), 평가자(opinion holder, h), 평가가 이루어진 시간(time, t)로 제시하였다.

여기서는 이들 중 화장품 코퍼스의 FbSA 수행을 위한 평가의 대상(e), 자질(a) 및 극성 표현(s)을 추출하기 위한 사전 구축의 방법론을 제시한다. 우선 극성 점수(s)에 해당하는 Pol-MWE를 추출하기 위하여, 본 연구에서는 '좋다', '나쁘다' 등과 같은 단일 단어 기반 감성분석 단계로 처리될 수 없는 문장을 판별하였다. 이를 위해 DECO 사전에 수록된 DecoPolClass 감성 어휘 및 어휘 연산 감성 분석법(Lexical algorithmic Sentiment Analysis)을 사용한 {PolaLexLGGbasedSSA} 모듈의 DecoSentiClassifier[10]를 활용하여 문장 단위로 '긍정/부정'의 극성을 표현하는 데이터를 추출하였다. 여기에 단일 단어 감성어휘가 존재하지 않는 문장들을 분류하였다. 즉 감성어휘가 포함되지 않았으나 특정 극성을 표현하는 문장은 분명히 일정 MWE 방식의 다른 장치에 의해 극성을 표현할 것으로 예측되었기 때문이다. 이에 해당 문장들을 추출하여 거기 실현된 고빈도의 용언들의 공기(collocation) 관계를 조사하여 MWE의 후보 시퀀스를 추출하였다.

개체명(e)의 MWE 추출과 관련하여, 화장품의 상품명을 나타내는 성분들은 구성요소 간의 선택적 결합으로 이루어지지만, 일반적으로 브랜드명은 제품 종류명(reference)에 선행되는 경향을 보이며, 두 요소 사이에 일정 명사 및 명사구가 상품명의 일부로서 실현되는 것이 가능함이 관찰되었다. 이에 따라, 이러한 명사구 패턴을 이루는 연쇄를 추출하여 EMWE 자원 구축에 1차적인 자료로 사용하였으며, 더불어 코퍼스 내에서 나타나는 상품명을 추출해내기 위한 문법적 추론 경로를 제시하였다.

자질(a) 명사 관련 MWE와 관련하여, 본 연구에서는 화장품 후기글에서 나타나는 자질명 목록을 연구자들의 언어학적 직관을 통해 구성하고, 이를 바탕으로 코퍼스 내 실현된 명사 빈도를 통계적으로 확인하였다. 또한 화장품 후기글 관련 웹 사이트에서 관련 자질 명사 목록을 추출하여 그 토대를 보완하였다. 이와 더불어 자질명사에 후치되는 요소들을 별도로 수집하여 언어학적 추론을 위한 패턴에 활용하였으며, 이들의 통사적 구성 및 용언의 의미 선택 제약(semantic selectional restriction)에 따라 자질 추정 명사와 해당 자질의 범주까지 추론하는 경로 또한 포함하였다.

4. DECO-MWE의 구성

데이터에 실현된 문장들을 자동으로 분석 및 추출하기 위하여, 본 연구에서는 프랑스 파리-이스트 대학교(UPEM)에서 구현된 다국어 리소스 구축 및 코퍼스 분석 플랫폼 UNITEX[11]의 전처리 모듈을 사용하였다. UNITEX 전처리 모듈에서는 언어 현상에 관련된 LGG를 사용자가 직접 유한상태 트랜스듀서로 구축하여 코퍼스에 적용하는 것이 가능하다.

LGG는 프랑스 전산언어학자 모리스 그로스(Maurice Gross)에 의해 방법론적 토대가 마련된 언어 기술 및 언어 처리 모델이다[1]. LGG 모델은 전이망 모델의 일환으로, 오토마타 방식의 개념으로 문법을 구현함으로써 문장 내부의 부분적인 언어 현상을 보다 유연하게 기술할 수 있다. 여기서는 그래프 방식으로 문법을 표상하고, 이를 재귀성 그래프를 활용하여 변이형을 포착하며, 이를 FST로 변환하여 텍스트를 사용자가 요구하는 형태로 처리하는 것이 가능한 UNITEX 플랫폼을 통해 진행되었다. 본 연구에서는 한국어 전자사전 DECO 태그셋에 기반하여 LGG 형식으로 DECO-MWE를 구현하였고, UNITEX 플랫폼을 사용하여 테스트 코퍼스 및 화장품 도메인 코퍼스 분석에 적용하였다.

4.1 극성 관련 MWE(Polarity MWE; Pol-MWE)

앞서 논의한 바와 같이, 극성이 나타난 문장 중 단일 감성어휘가 실현되지 않은 문장의 용언을 추출하여 살펴본 결과, '있다, 하다, 쓰다, 들다' 등의 무극성 용언들이 고빈도로 나타나는 것을 볼 수 있었다.

순위	표면형	레마형	빈도
1	사용하	사용하다	545.266
2	있	있다	489.242
3	하	하다	292.656
4	쓰	쓰다	228.071
5	구매하	구매하다	225.9
6	않	않다	213.075
7	느끼	느끼다	175.666
8	써보	써보다	161
9	사	사다	154.381
10	저	절다	146.375

표 1. 무극성 문장에서 나타나는 용언의 빈도목록

이러한 용언들의 공기관계를 통해 살펴본 코퍼스 내 예문은 (2), (3)과 같은 양상을 보였다.

(2) ㄱ. 색은 무난하구요, 제품 디자인이 너무 **마음에 드는** 제품이에요.
ㄴ. **끝내 주는** 크림이예요.
ㄷ. 이 제품에 푹빠져 면세점 갈 때마다 사오는 **인생 아이템**이 됐어요

(3) ㄱ. 자꾸 **잔여물이 남아요**.
ㄴ. 일반 파데랑 달리 **뭉치기 쉬워요**.(변경)
ㄷ. **가루 날림**이 있으니 주의하세요.

(2)에서 제시된 '마음에 드는', '끝내 주는', '인생 아이템'과 같은 표현은 각각의 어휘만으로는 극성을 판단할 수 없지만 관용적으로 모든 도메인에서 고정된 극성으로 활용되는 MWE(GMWE)들이다. 이는 (1ㄷ)에서처럼 명사와 명사의 결합 형태로도 나타날 수 있다.

이와 반대로, (3)에 나타난 '잔여물이 남다', '뭉치기 쉽다', '가루 날림'과 같은 표현들은 특정 도메인 내에서만 긍정/부정의 극성을 표현하는 MWE(DMWE) 부류이다. 이러한 표현들은 그 도메인 속성에 의존적이므로, 코퍼스에 따라 그 적용을 달리해야 한다.

본 연구에서는 모든 도메인에 활용이 가능한 부류를 GMWE, 특정 도메인에서만 의미를 갖는 부류를 DMWE로 설정하고 이들을 극성 MWE(Pol-MWE)의 하위 유형으로 설정하였다. 이때 해당 표현들이 문법적으로 어떻게 결합하는가에 따라 '체언+체언' / '체언+용언' / '용언+용언' / '그 외 굳어진 표현(frozen)'의 네 가지 통사적 구조로 분류하였다. 극성 태그의 측면에서, 각각의 MWE는 긍정 극성일 경우 <QXPO>, 부정일 경우 <QXNG>의 태그를 할당할 수 있도록 분리되었으며, 이들은 XML 방식으로 마크업되어 FbSA에 직접 활용되도록 하였다.[3)]

언어자원의 규모 및 적용 활용성을 최대화하기 위해, 본 연구에서는 화장품 리뷰 코퍼스에서 나타나는 GMWE뿐만 아니라 선행 연구 및 관용어 사전에서 확인된 관용구의 목록들을 추출하여 LGG로 구축하였다. 이에 따라 [12]에서 분류한 감성 MWE 256개와 Naver 관용어 사전에서 추출한 800여개가 포함되었으며, GMWE를 외부자료 기반 언어 자원과 실험코퍼스기반 언어 자원으로 구분하여 그 내부에 극성을 할당하는 방식으로 구성하였다. 현재 이러한 그래프문법의 실제 구성은 매우 복잡하고 방대한데, '체언+용언' MWE가 가장 높은 비중을 차지한다. 다음은 일부 예를 보인다.

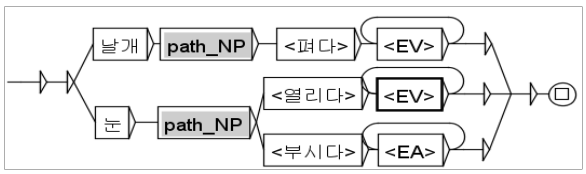

그림 1. GMWE 그래프의 일부

그림 1는 외부자료 기반 긍정 GMWE의 '체언 + 용언' 연속체 처리를 위한 LGG의 일부이다. 해당 그래프의 경로들을 통해 '날개를 펴다', '눈에 차다' 등의 감성 표현을 인식할 수 있다. 또한 복합 조사 및 부사 결합이 예상되는 형태와, 용언 어간에 명사화 접사가 붙은 형태들을 포착하기 위한 서브 그래프가 포함되어 있어, 인식할 수 있는 표면형은 더 늘어나게 된다.

DMWE를 기술하기 위해 코퍼스 내에서 극성이 나타나지 않는 비극성문들을 추출한 후, 선별된 문장 내부에 위치하는 체언 및 용언의 빈도를 통계적으로 분석하였다. 이를 기반으로 해당 어휘들 사이의 결합관계를 포착하여 극성을 나타내는 MWE 기본 구문을 그림 2와 같은 LGG로 표상하였다. 그림 2는 '각질(을 많이) 제거하다', '끈적임 현상이 적다'와 같은 긍정 표현의 MWE에 대한 활용형 구문을 기술하는 LGG이다.

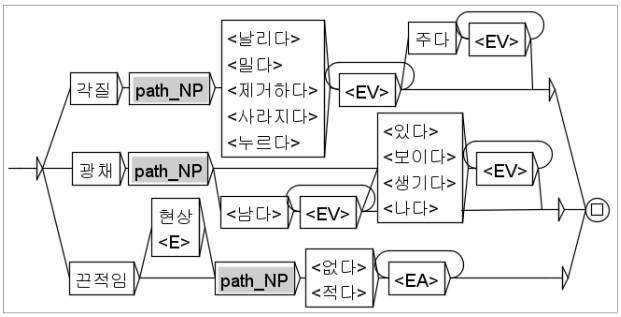

그림 2. DMWE 그래프의 일부

이와 같이 구축된 GMWE와 DMWE의 그래프들은 그림 3과 같은 방법으로 구조화된다.

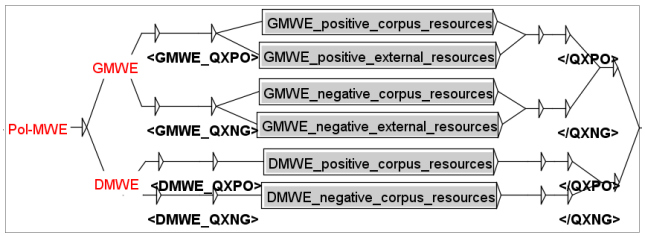

그림 3. DECO-MWE 내 Pol-MWE 구성

그림 3의 LGG 사전을 앞서 예문 (2)와 (3)에 적용하면 '<GMWE_QXPO> ABC DEF </QXPO>'와 같은 형태로 마크업된다. 즉 (4)와 (5)처럼 주석된다.

(4)ㄱ. <GMWE_QXPO>**마음에 드는**</QXPO> 제품이에요.
ㄴ. <GMWE_QXPO>**끝내 주는**</QXPO> …
ㄷ. <GMWE_QXPO>**인생 아이템**</QXPO>이 됐어요.
(5)ㄱ. 자꾸 <DMWE_QXNG>**잔여물이 남아요**</QXNG>
ㄴ. …<DMWE_QXNG>**뭉치기 쉬워요**</QXNG>.
ㄷ. <DMWE_QXNG>**가루 날림**</QXNG>이 있으니 …

4.2 화제 관련 MWE(Topic MWE; Top-MWE)

화장품 후기글에 나타나는 화제(Topic) 관련 MWE은 일반 명사의 결합으로 실현된다. 이를 위해 우선 코퍼스에서 나타나는 명사 연쇄 구성을 검토하였다. 개체명의 경우, 브랜드명과 제품 종류이 연관되어 실현됨을 확인하였다. 이를 추출하기 위한 LGG를 구축하여 이를 바탕으로 추출되는 명사 연속체를 1차적인 MWE 후보로 선정하였다. 자질 명사의 경우 또한 1차적인 자질 명사 목록을 확보한 후, 코퍼스에 기반하여 이를 확장하고 명사들의 공기 관계를 토대로 MWE 후보를 구성하였다. 개체명 및 자질명 관련 MWE의 예를 들면 다음과 같다.

3) Pol-MWE의 결합유형은 '체언 + 체언'{__NN.grf}, '체언+용언'{__NP.grf}, '용언+용언' 결합형{__PP.grf}으로 분류했으며, 해당 유형 외 굳어진 형태의 MWE 표현을 {__Frozen.grf}로 구성하였다.

(6)ㄱ. **겔랑 라이트 파우더**는 워낙 유명해서 …
ㄴ. **로레알 글로스 틴트**의 색상이 …
ㄷ. 티트리 성분을 가진 **아로마티카 오일**
(7)ㄱ. 이 크림은 **컬러 밝기**가 너무 좋아요.
ㄴ. 나인포인트 웜코튼 퍼퓸의 **향 농도**가 …
ㄷ. 그린티 씨앗의 **성분 함량**이 …

먼저, 개체명과 관련된 MWE(즉 EMWE)를 기술하기 위하여, (6ㄱ)과 같은 개체명의 일반적 패턴을 분석하였다. 이는 그림 4의 맨 위의 경로로 기술되었다. 즉 브랜드명과 제품종류명을 축으로 일련의 명사가 삽입되는 형태로, 이를 인식하고 추출하기 위해 XML 방식으로 정규화하였다. (6ㄴ)에서와 같은 개체명은 '색깔/성분' 등과 같은 자질 어휘를 선행하는 위치 정보를 통해 추론될 수 있다고 판단된 유형으로, 이를 토대로 그림 4의 두 번째 경로가 기술되었다. 즉 속격 조사(<GEN>)를 통해 개체명과 자질 명사의 의미관계가 구성된 경우이다. 자질 어휘는 FMWE의 하위 모듈인 {FMWE_Standard.grf}를 활용하였고, 이러한 방법을 통해 MWE 구축을 위한 이중 증식을 수행하였다. 마지막으로 (6ㄷ)에 대한 분석을 통해 '자질 어휘 + 목적격 조사 + (부사) + 가진/갖춘/띄는(소유동사의 관형형)' 뒤에 나타나는 명사 연쇄를 개체명으로 추론하는 것이 가능하다. 이는 그림 4의 맨 하단의 경로로 기술되었다.

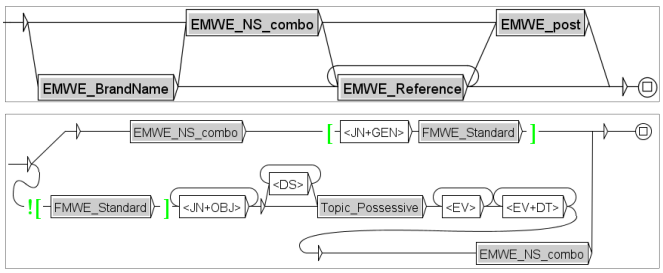

그림 4. 개체명 추론 EMWE 처리를 위한 전체 경로

이와 같은 방식으로 구축된 EMWE 그래프사전을 코퍼스에 적용하면 예문 (6)의 개체명 MWE들은 (8)과 같은 처리 양상을 보이게 된다.

(8)ㄱ. <겔랑-파우더_XXPR>
겔랑 라이트 파우더</XXPR>는 …
ㄴ. <로레알-틴트_XXPR>
로레알 글로스 틴트</XXPR>의 색상이…
ㄷ. 티트리 성분을 가진
<아로마티카-오일_XXPR>**아로마티카 오일**</XXPR>

같은 방법으로, 이번에는 (7)과 같은 자질명 MWE를 기술하는 과정을 진행한다. (7ㄱ)에 나타난 '컬러 밝기'와 같은 표현 뿐만 아니라 성분/향기/모양 등에 대한 다양한 변이의 MWE들을 정규화하는 것이 필요하다. 그림 5는 이를 COLOR/INGREDIENT/SCENT와 같은 형태로 정규화시키며, 각각의 자질을 XML 형식으로 마크업하는 그래프를 보인다.

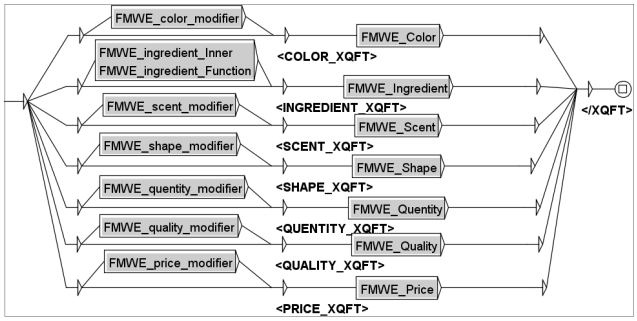

그림 5. FMWE 처리를 위한 경로

또한 (7ㄴ)과 같이 개체명에 후치되는 명사 연쇄를 '개체-자질'의 의미 관계로 설정하여, 후치 명사의 '향 농도'를 SCENT 자질로 정규화하는 것이 가능하다. 즉 앞서 EMWE에서와 같이 속격 조사 및 소유동사 그래프를 이용하는 방법으로, 이는 그림 6의 상단 그래프의 경로를 통해 드러난다.

유사한 맥락으로, (7ㄷ)과 같이 개체명 뒤에 주격/장소격/출처격 조사<JN+SUB/LOC/SOU>가 실현되는 경우, 해당 명사 뒤에 나타나는 서술어의 의미 선택 범주에 따라 자질 추정 명사를 각각 COLOR/SCENT 등으로 정규화할 수 있다. 즉 이 경우는 동사의 의미 선택제약에 따라 자질 추정 어휘를 정규화하는 것이 가능하다. 이는 그림 6의 하단 경로를 통해 표상된다.

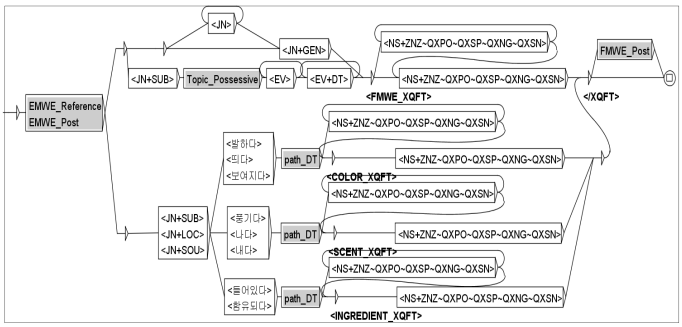

그림 6. 의미제약 관계 기반 추론을 통한 FMWE 처리 경로

이상과 같은 방식으로 FMWE 자원이 구축되면 예문 (7)에서 나타나는 MWE를 (9)와 같이 정규화할 수 있다.

(9)ㄱ.이 크림의 코랄 빛
<COLOR_XFQT>**컬러 밝기**</XFQT>가 …
ㄴ. 나인포인트 웜코튼 퍼퓸의
<SCENT_XFQT>**향 농도**</XFQT>가 …
ㄷ. 그린티 씨앗의
<INGREDIENT_XFQT>**성분 함량**</XFQT>이 …

5. 성능 평가

이상에서 구축된 DECO-MWE를 평가를 위해 별도로 구축한 300개의 화장품 리뷰 문장에 적용한 결과를 보이면 표 3과 같다.

Evaluation	GMWE	DMWE	EMWE	FMWE
Precision	93.5	92.1	69.0	86.3
Recall	80.5	74.6	65.7	76.8
F-measure	86.5	82.4	67.3	81.2
Evaluation	**Pol-MWE**		**Top-MWE**	
F-measure	84.4		74.2	

표 3. DECO-MWE 모듈의 정보 검색 결과

극성 분석 성능에 직결되는 Pol-MWE의 경우 84.4%의 준수한 조화평균(F-measure)을 보이고 있다. 반면 Top-MWE 중에서 EMWE의 성능은 67.3%의 상대적으로 저조한 점수를 보이는데, 이는 평균적으로 7~15 음절로 이루어지는 화장품 상품명의 복합적인 결합관계에 의한 것으로, 개체명 MWE 자원의 범위 및 정확도를 더 확장시킬 필요성을 시사해 준다.

6. 결론 및 향후 연구

자질기반 감성분석(FbSA)에 필요한 MWE 언어 자원의 일환으로서, 본 연구에서는 한국어의 체언, 용언 및 조사, 어미 결합에 대한 방대한 정보를 제공하는 DECO 사전을 토대로 DECO-MWE 그래프사전을 구축하였다. 화장품 후기글 코퍼스를 분석하여 여기 나타나는 고빈도 체언과 용언들의 결합 패턴을 분석함으로써 MWE를 네 가지 하위 유형으로 자원화하였다.

다음은 [10]에서 소개되었던 DecoLGGbasedFSA 모듈을 사용하여 본 연구에서 구축된 언어자원을 화장품 후기글 코퍼스에 적용한 FbSA 결과의 시각화 예를 보인다.

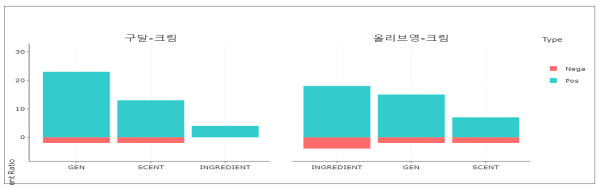

그림 7. '브랜드-크림' EMWE의 자질별 감성 분석 시각화

그림 7은 DECO 사전에 내장되어 있는 단일어 감성어휘 표제어(PolCalss)와 본 연구에서 구축한 DECO-MWE 그래프사전을 적용하여 획득한 감성분석 결과를 보인다. 즉 화장품 후기글 사이트 '파우더룸'에서 크롤링된 코퍼스로부터 '브랜드명&크림' 유형의 평가대상에 대해 분석한 결과 중 상위 두 개에 대한 각 자질별 감성분류 결과를 시각화한 것이다. 위의 그래프는 왼쪽부터 오른쪽 순으로 '브랜드명&크림'의 오름차순 결과를 시각화하고 있는데, 여기서 'GEN'은 문장 내 자질을 지칭하는 단어 없이 극성만 표현된 문장을 일괄 처리한 값이다. 위의 예시는 본 연구에서 구축한 DECO-MWE가 향후 FbSA 연구에 효과적으로 사용될 수 있음을 암시한다.

본 연구에서 제안한 MWE 언어자원 구축 방법론은 향후 다른 도메인의 코퍼스에도 동일한 방식으로 확장되어 적용될 수 있다. 향후 범용성을 가지는 극성 표현의 확장과 도메인 의존적인 극성 표현에 대한 체계적인 자원 구축이 수행된다면 보다 신뢰할만한 FbSA 연구를 기대할 수 있을 것이다.

참고문헌

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

[2] J, Nam, "Study on automatic recognition of Korean negation markers shifting opinion polarity", Language and Linguistics, 57, 61-94, 2012.

[3] B. Erman and B. Warren, "The idiom principle and the open choice principle", Text, 20(1), 29-62, 2000.

[4] S. Piao, P. Rayson, D. Archer, A. Wilson and T. McEnery, "Extracting multiword expressions with a semantic tagger", In Proceedings of the ACL Workshop on Multiword Expressions, 49-56, 2003.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis", Computational Linguistics, 37 (2), 267-307, 2011.

[6] T. Baldwin and S. Kim, Handbook of natural language processing, CRC Press, Boca Raton, USA, 2nd edition, 2010.

[7] K. Lee, J. Kim and B. Yun, "Extracting Multiword Sentiment Expressions by Using a Domain-Specific Corpus and a Seed Lexicon", Computational Linguistics, 37(2):267-307, 2011.

[8] K. Kim, "(An) analysis on international competitiveness of Korean cosmetics industry by diamond model", Graduate School, Seongkyungwan University, Thesis of Master degree, 2017.

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

[10] 유광훈, 남지순, "DECO 감성사전과 LGG 문법에 기반한 관광 숙박 온라인 후기글의 감성 분석 연구," 한국사전학, 30, 100-154, 2017.

[11] S. Paumier, "Unitex 3.1 Users' Manual", 2003.

[12] 최석재, 정연주, 정경미, 홍종선, "구조에서 나타나는 감정 표현 관용구의 의미", 한국어 의미학, 35, 311-333, 2011.