Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 1, pp. 056-067, February 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.1.056

<Application Research>

비지도 학습 기반의 한국어 속성 추출에 적합한 전처리 방법 연구 : 육아용품 상품평을 대상으로

박명현 · 최희련 · 이홍철†

고려대학교 산업경영공학과

Study on Preprocessing Method Suitable for Korean Aspect Extraction based on Unsupervised Learning : For Childcare Products Reviews

Myung-Hyeon Park · Hoeryeon Choi · Hong-Chul Lee

Department of Industrial Management Engineering, Korea University

In aspect-based sentiment analysis, aspect extraction is a key task and various methods are used. However most of the aspect extraction studies target English reviews, and studies on Korean reviews are insufficient. So in this study, we proposed five preprocessing methods that fit the characteristics of Korean reviews and used Word2Vec and Attention models to derive meaningful results of aspect extraction based on unsupervised learning. As a result of the evaluation of aspect extraction performance, the method of spacing correction, morpheme analysis, and removing unnecessary POS by domain were superior to other methods. In addition, we have confirmed that Word2Vec is the most suitable embedding method for Korean aspect extraction using Attention‐based Aspect Extraction compared to GloVe and FastText. We expect that the proposed preprocessing and embedding method will be useful for research on the aspect extraction of Korean reviews in the future.

Keywords: Preprocessing, Aspect Extraction, Attention, Word2Vec, Unsupervised learning

1. 서 론

온라인상에서 소비자가 상품평을 통해 제품과 서비스에 대해 평가하게 되면서 방대한 양의 텍스트 데이터가 생성되고 있다. 온라인 상품평은 소비자와 생산자 모두에게 영향을 미치는데, 상품평을 작성한 소비자들은 잠재적인 소비자에게 제품에 대한 경험을 공유하면서 제품 소비에 직·간접적인 영향을 줄 수 있고, 생산자들은 상품평을 피드백으로 활용하여 자신들의 제품 및 서비스 품질 개선에 반영한다. 상품평에 나타나는 사용자의 감성 극성을 긍정 또는 부정으로만 분류하는 초기 연구 방식은 문서에 담겨있는 전체 의견을 반영하지 못하여 일부의 감성 극성만 도출하는 문제가 발생하였다(Hu and Liu, 2004). 따라서 제품의 속성에 대한 감성을 분석하는 속성 기반 감성 분석이 등장하였고, 속성 기반 감성 분석의 가장 많은 연구가 속성 추출(aspect extraction)에서 이루어지고 있다(Rana and Cheah, 2016). 속성 추출은 (1)속성 용어 추출, (2)추출된 용어가 어떤 범주(category)에 속하는지를 범주화하는 2단계로 이루어진다(He *et al*., 2017).

기존의 속성 추출 연구들은 대부분 영어를 다루고 있으며, 문장 분석의 최소 단위로 띄어쓰기 기준의 단어(word)를 사용한다. 그러나 한국어는 조사와 어미가 발달하여 단어의 형태 변화가 심한 교착어의 특성을 갖고 있어 형태소(morpheme)가 문장의 구조에서 중요한 역할을 한다. 즉, 단어(word) 기준의 방식을 한국어에 적용하면 같은 어간의 단어도 조사와 어미에

† 연락저자 : 이홍철 교수, 02841 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3389, Fax : 02-929-5388, E-mail : hclee@korea.ac.kr
2020년 8월 11일 접수; 2020년 10월 22일 수정본 접수; 2020년 10월 27일 게재 확정.

따라 다른 어휘로 학습되어 비효율적 학습이 이루어질 우려가 있으므로 형태소 단위로 나누어 분석하는 방법이 합리적이다(Park *et al*., 2018). 예를 들어 '가격도 싸고 흡수성도 좋아요'라는 문장을 어절 단위로 나누면 '가격도', '싸고', '흡수성도', '좋아요'의 4개 단어로 벡터화 되는데, 다른 문장에서 '가격이', '가격은'과 같이 다른 조사를 사용하는 경우 '가격도', '가격이', '가격은'의 세 단어는 서로 다른 벡터값을 가지게 되어 학습 효율이 저하된다. 특히 상품평은 다른 텍스트와 비교하여 띄어쓰기나 오타와 같은 맞춤법 오류, 신조어 사용이 빈번한 특징을 가진 데이터로 그에 맞는 전처리가 필요하다.

따라서 본 연구에서는 한국어 상품평의 특성에 맞는 형태소 단위의 전처리 방식과 그에 적합한 임베딩 방법을 제안하고, 이를 비지도 학습(unsupervised learning)인 Attention 기반의 속성 추출 모델(Attention-based Aspect Extraction, ABAE)에 적용하였다(Mnih *et al*., 2014; He *et al*., 2017).

본 연구는 다음과 같은 의의를 갖는다. 첫째, 온라인 상품평 속성 추출 문제에 맞는 한국어 상품평의 전처리 방식 및 임베딩 방법을 제안했다. 둘째, 한국어에 대해서는 연구되지 않았던 딥러닝 기반 비지도 방식의 속성 추출 모델을 한국어 상품평에 적용하여 유의미한 결과를 도출하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 한국어 상품평 전처리와 속성 추출의 관련 연구를 소개한다. 제 3장에서는 연구에서 사용한 단어 임베딩 방법론과 ABAE 모델에 대해 서술하고, 제 4장에서는 적용한 데이터 셋과 실험 방법에 대해 설명한다. 제 5장에서는 실험 결과를 분석하며 제 6장에서는 결론 및 향후 연구에 대해 서술한다.

2. 관련 연구

상품평을 분석하여 속성을 추출하는 가장 큰 목적은 소비자의 구매 활동에 적합한 정보를 제공하기 위함이다. 소비자는 제품에 대한 경험이나 지식 없이 구매해야 하는 첫 구매의 상황일 때 어려움을 겪기 쉬우며, 특히 육아관련 제품은 대상자인 영유아 또한 고려해야 하므로 소비자들은 구매에 더욱 신중할 수밖에 없고 관련 정보를 더욱 필요로 하게 된다(Jung, 2018). 이러한 맥락에서 소비자의 의사결정을 지원하기 위해 본 논문에서는 육아용품 도메인을 실험 대상으로 선정하였다.

2.1 한국어 상품평 전처리

한국어 상품평을 대상으로 감성 분석과 같은 다양한 연구가 이루어지고 있으며 수행하는 과제에 적합한 전처리 방식에 대한 연구도 진행되고 있다. Oh *et al*.(2019)은 한국어 영화 상품평의 감성 분석을 위해 단어를 단어, 음절, 음소 단위로 나누는 방법을 제시하였고 Bidirectional Long Short Term Memory 구조를 이용하여, 음절 단위 전처리 방식이 가장 효과적임을 확인하였다. Park *et al*.(2018)은 한국어 화장품 상품평을 대상으로 형태소 단위 분리와 품사 태깅의 전처리 방식을 적용하여 CNN으로 감성 극성을 분류하였다.

이처럼 한국어 상품평의 전처리를 위해 단어를 소리 단위나 형태소 단위로 나누고 품사 태깅을 수행하는 등 다양한 방법이 사용되고 있으며, 연구의 목적에 따라 각기 다른 전처리 방식이 우수한 성능을 나타내고 있다.

2.2 속성 추출(Aspect Extraction)

속성 추출은 Hu와 Liu(2004)가 속성 기반 감성 분석에서 제품의 특징(features)을 뽑아내는 연구를 수행하면서 하나의 과제로 등장하였다. 속성 추출은 규칙 기반 방식(rule-based), 지도 학습 방식(supervised learning), 비지도 학습 방식(unsupervised learning) 세 가지로 나눠진다.

Hu와 Liu는 명사의 빈도와 명사구의 규칙을 통해 제품의 특징을 추출하는 빈도 기반 방식을 제안하였고(Hu and Liu, 2004), 이후 문장에서 빈번하게 나타나는 패턴을 이용하여 제품의 특징을 뽑는 연구가 진행되었다(Qiu *et al*., 2011). 패턴을 이용한 연구들은 속성 용어들을 명사로 제한할 때만 잘 작동하여 사전에 정의된 규칙에 상당한 의존성을 갖고 있어 추출한 속성 용어를 범주화(categorization)하기에는 부족한 것으로 나타났다.

지도 학습 방식은 일반적으로 속성 추출 문제를 시퀀스 라벨링(sequence labeling) 문제로 모델링한다. 최근에는 신경망 모델의 이용이 확대되면서 Recurrent Neural Network, Conditional Random Fields, Attention, Convolution Neural Network 등의 방법론이 속성 추출에 사용되고 있다(Yin *et al*., 2016; Xu *et al*., 2018). 이 같은 지도 학습 방식은 학습을 위해 다량의 레이블이 있는 데이터가 필요하다는 특성이 있다.

비지도 학습 방식은 레이블 없이 학습이 진행되는 방법으로 토픽 모델링이 대표적이며, 여러 연구들이 LDA(Latent Dirichlet allocation)(Blei *et al*., 2003)을 기반으로 확장된 속성 추출 방법을 제안하였다(Wan *et al*., 2020). 인공 신경망을 이용하면서 Wang *et al*.(2017b)은 여러 개의 Attention 층을 연결하여 속성 용어와 감성 용어를 동시에 추출하는 모델을 제안하였다. 다중 Attention 레이어는 속성용어와 감성용어의 정보를 상호 학습하면서 간접적 관계까지도 추출하였으나 구조적인 복잡성이 크다는 특징이 있다. He *et al*.(2017)은 단어 임베딩과 Attention 모델을 함께 사용하는 ABAE 모델을 제안하였다. 이전 토픽 모델 방식은 속성의 일관성이 부족하다는 한계점이 있는 반면에 ABAE 모델은 임베딩된 단어의 동시 발생 분포를 이용하여 일관성(coherence)을 유지할 수 있게 하였다. 또한, Attention은 속성과 관계없는 단어들의 가중치는 낮게 학습하는 구조로 속성 추출의 성능을 높였다. 선행연구들은 대부분 영어 텍스트를 대상으로 이루어졌으며 특히 신경망을 이용한 연구는 SemEval 등의 데이터 셋을 이용하여 노트북이나 식당 상품평 등과 같은 일부 도메인으로 제품군의 범위가 한정되어 있다는 특징이 있다.

한국어 상품평을 대상으로 하는 속성 기반 감성 분석은 활발하게 이루어지고 있으나 감성의 대상인 속성을 추출하는 연구는 부족하며 한국어 데이터에 딥러닝 기반의 속성 추출을 적용한 사례는 아직 없다. 한국어 속성 추출과 유사한 연구는 다음과 같다.

Park and On(2017)은 토픽 모델링 방법인 LDA를 이용하여 문장의 토픽을 추출하고 핵심 토픽 단어를 대상으로 감성 사전을 구축하여 속성의 개념과 유사하게 활용하였다. 핵심 토픽 단어에 대한 긍·부정 점수를 산출하여 차량 성능을 대상으로 상품평을 요약하는 방식을 제안하였다. Jeong *et al*.(2018)은 한국어 휴대폰 상품평을 대상으로 단어 임베딩과 네트워크를 분석하여 문서의 다중 범주 가중치를 산출하는 비지도 방식의 방법론을 제안하였다. 임베딩된 단어 간 거리에 따른 가중치 행렬과 동시 발생 행렬의 요소 곱을 통해 네트워크를 구축하고 중심성 지표가 높은 중요 키워드의 가중치를 문서에 적용하여 가장 높은 가중치 점수를 갖는 기능이 문서의 범주가 된다. 휴대폰 기능에 대한 핵심 범주는 휴대폰이 갖는 속성과 매우 유사하며 하나의 상품평에 대한 단일 및 다중 범주 분류에 좋은 성능을 보였다.

따라서 본 논문에서는 영어에 한정적이었던 속성 추출을 한국어 육아용품 상품평에 적용하여 한국어 상품평의 속성을 추출하는 것을 목적으로 하며, 정답 레이블이 없는 데이터를 사용하기 때문에 비지도 학습 방식의 Attention 기반 ABAE 모델을 사용하였다.

3. 방법론

3.1 Word Embedding

텍스트 데이터를 입력값으로 사용하기 위해서는 연산이 가능한 형태로 변환해야 한다. 대표적인 방법으로 각 단어를 연속된 공간상의 실수(real number)로 변환하는 분산표상(distributed representation)이 있다. 분산 표상 방식은 단어의 의미적(semantic), 구문적(syntactic) 정보를 반영한다는 장점이 있다. 본 연구에서는 신경망을 이용한 Word2Vec(Mikolov *et al*., 2013a; Mikolov *et al*., 2013b)의 Continuous Bag-of-Words(CBOW) 모델을 사용하였으며 속성 추출 성능 비교를 위해 GloVe(Pennington *et al*., 2014)와 FastText(Bojanowski *et al*., 2017)를 이용하였다.

Word2Vec은 간단한 인공 신경망 구조를 이용하며 대용량의 데이터를 효율적으로 계산하여 벡터를 얻을 수 있고, 표현된 벡터가 단어 간의 관계 정보를 포함하기 때문에 감성 분석(sentiment analysis) 등의 자연어 처리 분야에서 다양하게 쓰이고 있다. Word2Vec은 CBOW와 Skip-gram의 두 가지 방식으로 나뉘며 <Figure 1>는 윈도우 크기(window size)가 c일 때 두 모델의 구조를 나타낸 것이다.

CBOW 모델은 문장이나 문서에서 윈도우 크기만큼의 주변 단어가 주어질 때 중심 단어를 잘 예측하는 단어 표상을 찾도록 학습한다. 주변 단어를 one-hot vector의 형태로 입력하면 가중치 행렬 W를 통해 원하는 차원으로 사영되고 가중치 행렬 W'을 통해 중심 단어가 출력된다. 모델은 중심 단어가 나올 확률을 최대로 하는 가중치 행렬 W와 W'을 학습하게 되며 최종적으로 W가 임베딩 행렬이 된다. Skip-gram 모델은 중심 단어가 주어질 때 주변 단어를 예측하는 방법으로 입력과 출력이 CBOW 방식과 반대로 작동한다.

GloVe는 Pennington *et al*.(2014)이 제안한 분산 표상 방식의 하나이다. Word2Vec 방식이 사용자가 정한 윈도우 크기에 따라 부분적으로 문맥을 학습하는 반면, GloVe는 말뭉치 전체를 대상으로 두 단어 간 동시 발생 수를 학습에 이용한다. GloVe는 목적 함수가 임베딩된 단어 벡터의 내적(inner product)이 전체 말뭉치에서 단어의 동시 발생 확률이 되도록 학습하므로 단어 유사도 추정 등에서 좋은 성능을 보인다.

FastText는 단어를 글자의 n-gram으로 표현하는 Skip-gram 기반의 분산표상 방식이다(Bojanowski *et al*., 2017). FastText를 통해 표현되는 단어 벡터는 n-gram 벡터의 합으로 표현되며 학습방식은 Word2Vec과 유사하다.

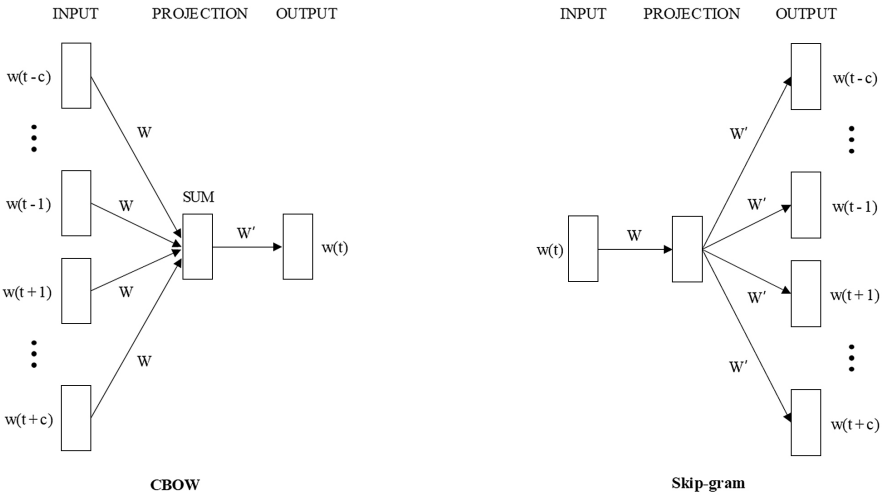

Figure 1. The Architecture of CBOW and Skip-Gram(Mikolov *et al*., 2013)

3.2 ABAE(Attention-based Aspect Extraction)

본 연구에서는 He *et al.*(2017)이 제안한 Attention 기반의 비지도 속성 추출 모델을 사용하였다. Attention 모델은 입력한 정보 중에서도 해결하려는 문제와 가장 적합한 정보에 집중하는 특성이 있다. 이 점을 이용하여 ABAE는 임베딩 공간에서 각 속성을 나타내는 대표 단어(representative word)를 통해 해석이 가능한 속성 임베딩(aspect embedding)을 학습하는 것을 목표로 한다. <Figure 2>는 한 문장이 3개의 단어로 이루어져 있을 때 ABAE의 구조를 시각화 한 것이다.

$$d_i = e_{w_i}^{\mathrm{T}} \cdot \mathrm{M} \cdot y_s \tag{1}$$

$$y_s = \frac{1}{n}\sum_{i=1}^{n} e_{w_i} \tag{2}$$

$$a_i = \frac{\exp(d_i)}{\sum_{j=1}^{n}\exp(d_j)} \tag{3}$$

$$z_s = \sum_{i=1}^{n} a_i e_{w_i} \tag{4}$$

$$p_t = \mathrm{softmax}(\mathrm{W} \cdot z_s + b) \tag{5}$$

$$r_s = T^{\mathrm{T}} \cdot p_t \tag{6}$$

문장의 단어인 w는 Word2Vec 모델로 학습된 임베딩 행렬 E (V×d)를 통해 d차원의 벡터 e_w로 변환되어 모델에 입력된다. 여기서 V는 어휘의 크기를 의미한다. 식 (1)과 식 (2)를 이용하여 변환 행렬 M을 통해 비속성(non-aspect) 단어를 걸러내고, 문장 전체의 문맥 정보를 포함하는 y_s와 내적하여 문장과 걸러진 단어와의 관계 d_i를 얻는다. 다음으로 식 (3)의 Attention 가중치를 계산하고 식 (4)의 문장 임베딩 z_s를 도출한다. Attention 가중치는 각 단어가 문장의 핵심 주제를 알아내는데 얼마나 주목할 만한지를 확률로 나타낸다. 식 (5)를 통해 입력 문장이 어떤 속성에 속하는지를 나타내는 확률값 p_t를 계산하고, 속성 임베딩 행렬 T를 이용하여 속성 임베딩의 선형 결합으로 식 (6)의 문장 임베딩 r_s를 재구성(reconstruction)한다. 차원 축소 및 재구성을 통해 ABAE 모델은 가능한 왜곡을 최소화하도록 z_s를 r_s로 변환하면서 K개의 임베딩 된 속성에서 속성 단어의 정보를 최대한 보존한다.

$$J(\theta) = \sum_{s \in D}\sum_{i=1}^{m} \max(0, 1 - r_s z_s + r_s n_i) \tag{7}$$

$$U(\theta) = \| \mathrm{T}_n \cdot \mathrm{T}_n^{\mathrm{T}} - \mathrm{I} \| \tag{8}$$

$$L(\theta) = J(\theta) + \lambda U(\theta) \tag{9}$$

모델은 재구성 오류를 최소화하기 위해 부정표본(negative sample) n_i를 무작위로 추출하고, 재구성된 r_s가 z_s와는 유사하되 n_i와는 다르도록 식 (7)을 목적함수로 두고 학습한다. 최종적으로 속성 임베딩 행렬의 중복 문제 해결을 위해 식 (8)의 정규화 과정을 거쳐 식 (9)의 비용 함수가 정의되며, 학습을 통해 각 문장의 속성에 해당되는 대표 단어를 얻게 된다.

4. 실험 설계

실험은 <Figure 3>에 따라 진행되었다. 먼저 온라인 쇼핑몰에서 상품평을 수집하고 특수문자 제거, 띄어쓰기 교정과 형태소 분석 등 입력 문장을 다섯 가지 방식으로 전처리하였다. 전처리된 문장은 Word2Vec, GloVe와 FastText를 통해 단어 수준으로 임베딩 되고, 전처리 방식과 임베딩 방법에 따라 ABAE 모델에 입력된다. 각 문장은 Attention 구조와 속성 임베딩 행렬을 거치면서 속성별 대표 단어를 추출하고, 이것을 기준으로 속성 키워드와 기준 속성으로 매핑된다. 최종적으로 가장 높은 가중치를 가진 속성이 상품평의 속성으로 할당된다. 정답 레이블이 없는 비지도 방식의 학습 결과를 평가하기 위하여 설문 조사를 활용하였으며 설문 조사 결과를 레이블로 하여 전처리 방식 및 임베딩 방식의 성능을 평가하였다. <Figure 4>는 속성 추출의 기술적 과정을 설명한다.

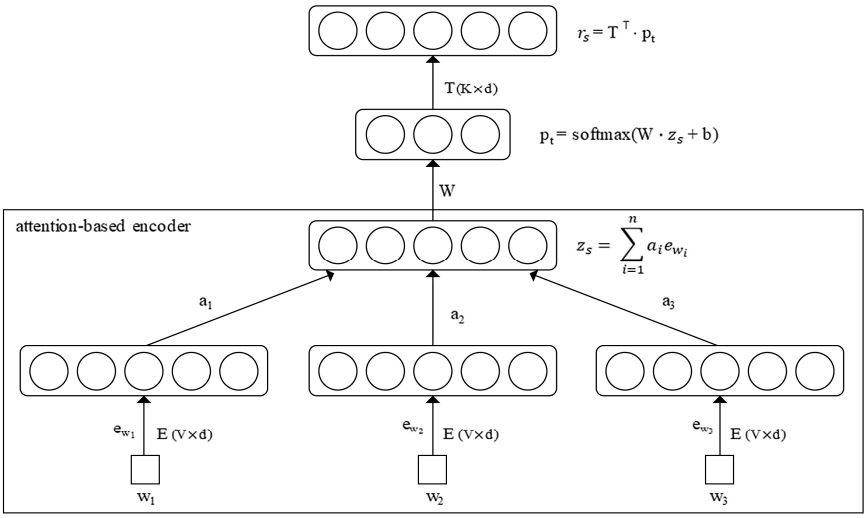

Figure 2. An Example of ABAE Structure(He *et al.*, 2017)

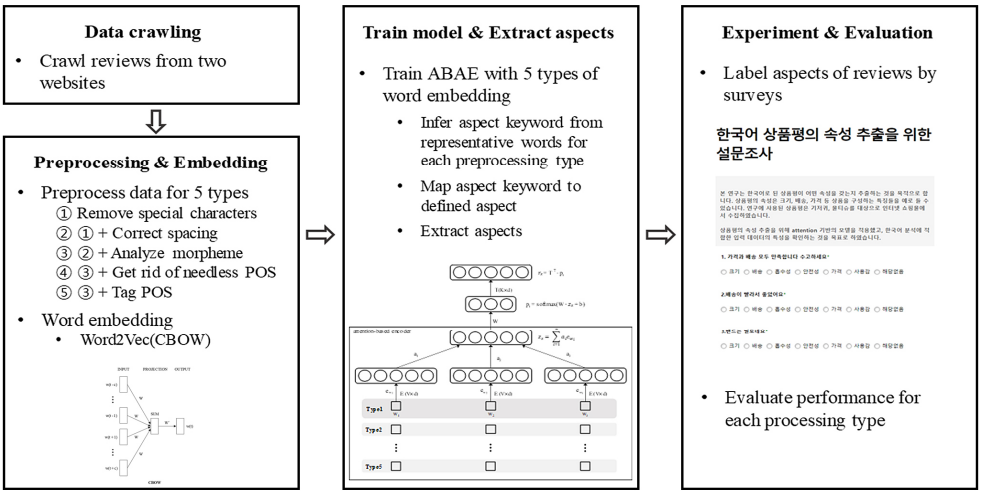

Figure 3. Framework of Research

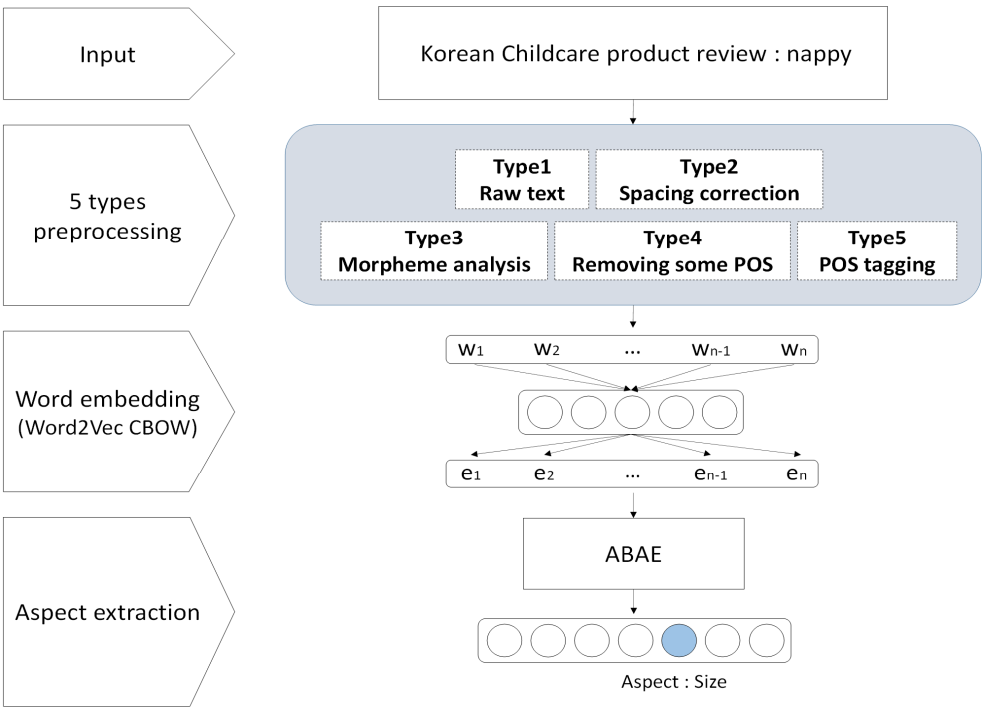

Figure 4. Aspect Extraction Procedure

4.1 데이터 수집 및 전처리

본 연구의 수행을 위해 온라인 쇼핑몰에서 기저귀와 물티슈의 상품평 데이터를 215,585개 수집하였다. 실험에서 하나의 상품평이 하나의 문장으로 사용되었으며 데이터에 대한 설명은 <Table 1>과 같다.

한국어로 된 여러 종류의 글에서 정해진 표준 문장 길이는 없지만, 명확한 목적을 가진 기사문, 요약문, 방송문 등에서는 독자의 이해를 돕기 위해 50자 정도를 기준으로 삼는다(Jang, 2002). 그러나 상품평의 경우에는 사용자의 경험과 감정 서술에 목적을 두고 있어 정해진 양식이 없으며 구어체로 작성된다. 또한 홑문장보다는 수 개의 문장이 합쳐진 겹문장으로 작성하는 경우가 대부분인 특징을 가진다. 수집한 기저귀 도메인

Table 1. Dataset Description

Domain	Reviews	Preprocessed reviews	Average number of characters
기저귀	85,066	57,397	88.1
물티슈	130,519	96,389	61.7

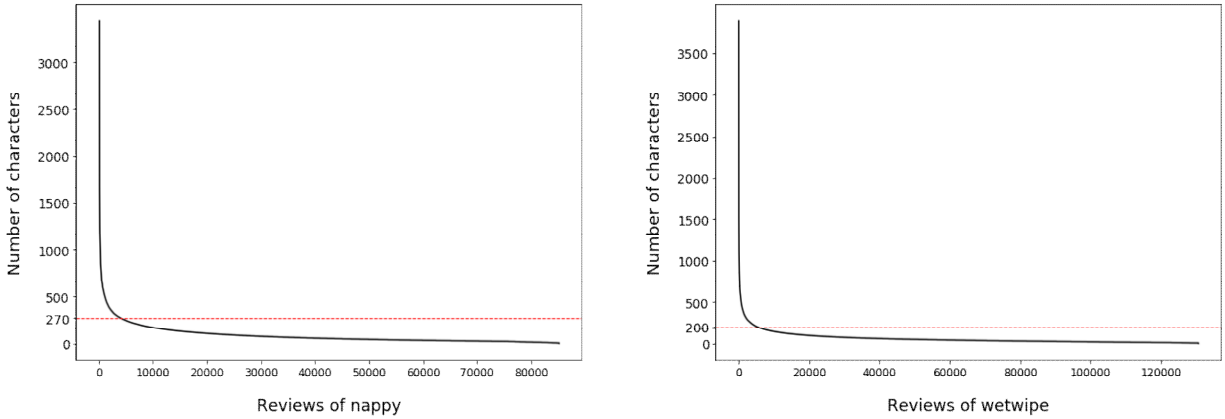

Figure 5. Distribution of Reviews in Nappy and Wetwipe Domains

상품평의 평균 글자 수는 88.1자이고, 물티슈 상품평의 평균 글자 수는 61.7자이다. <Figure 5>의 상품평 분포를 보면 기저귀 상품평은 270자 이하에서, 물티슈 상품평은 200자에서 변곡점이 나타난다. 각 상품평이 하나 이상의 속성을 포함하도록 주어와 서술어를 고려하고, 정제된 글의 문장 길이, 상품평의 특징, 수집한 데이터의 분포를 고려하여 최대 150자의 상품평을 선별하였다. 기본 전처리로 특수문자를 제거한 결과 기저귀와 물티슈 제품군에서 각각 57,397개와 96,389개의 상품평을 도출하였다.

본 연구에서는 한국어 상품평을 5개 방식으로 전처리하여 Attention 기반의 속성 추출에 적용하였고, 이를 통해 한국어 상품평의 속성 추출에 적합한 전처리 방식을 제안하였다. 5개 방식의 전처리는 ① 특수문자 제거, ② 띄어쓰기 교정, ③ 형태소 분석, ④ 형태소 분석+불필요 품사 제거, ⑤ 형태소 분석+품사 태깅으로 진행하였다.

영어 상품평의 경우 일반적인 전처리 과정은 특수문자 제거, 대소문자 변환, 관사 및 부정어 등의 불용어 제거로 진행된다. 본 연구에서는 영어 전처리 방식과의 비교를 위해 특수문자 제거만을 수행하는 방식(Type 1)과 띄어쓰기를 교정(Type 2)하는 전처리를 포함하였고 이 두 가지 전처리는 다른 세 가지 방식에 공통으로 적용되었다.

Type 3은 형태소 분석, Type 4는 형태소 분석 후 의미상의 속성이나 속성에 대한 감정을 나타낼 수 있는 명사/형용사/동사/부사/영어/숫자 외의 품사를 제거하였다. 마지막 Type 5는 형태소에 품사 태깅을 적용하여 총 5개 형태의 텍스트 데이터로 정제하였고 형태소 분석과 품사 선택 및 품사태그 부착에는 한국어 자연어처리를 위한 Python 패키지인 konlpy의 Okt 형태소 분석기를 사용하였다. Okt는 비표준어나 이모티콘 사용이 많은 온라인 상품평 분석에 적절한 것으로 알려져 있다. <Table 2>는 본 논문에서 제시하는 5개 전처리 방식을 적용한 상품평의 분석 결과를 그대로 제시한 형태별 예시이다.

4.2 단어 벡터

단어 벡터는 Gensim 라이브러리의 Word2Vec, FastText와 glove_python 라이브러리의 GloVe를 사용하여 도출하였다. Word2Vec의 경우 CBOW 모델을 적용하여 200차원의 공간에 임베딩 하였으며 윈도우 사이즈는 4, 최소 빈도수는 10, 네거티브 샘플링 크기(negative sampling size)는 5로 설정하였다. 신경망 반복 횟수는 100으로 충분히 학습하여 수렴할 수 있도록 설정하였다. GloVe와 FastText도 임베딩 차원과 윈도우 사이즈 등 파라미터값을 Word2Vec과 동일하게 적용하여 학습하였다. FastText의 특성상 품사 태깅한 데이터의 임베딩 결과가 ABAE 적용에 부적절하여 Type 5의 FastText 임베딩은 속성 추출 실험에서 제외되었다.

Table 2. Examples of Preprocessing Types. An Example of Type 5 Result Includes the wrong POS Tag, but this Table Shows the Actual Analysis Result as it is

Preprocessing types		Examples of review
Type 1	기본 전처리	확실히 달라요 흡수력좋네요 아기도 편해하는것같아요
Type 2	띄어쓰기 교정	확실히 달라요 흡수력 좋네요 아기도 편해하는 것 같아요
Type 3	형태소 분석	확실히 달라 요 흡수 력 좋네요 아기 도 편해하는 것 같아요
Type 4	형태소 분석+불필요 품사 제거	확실히 달라 흡수 좋네요 아기 편해하는 것 같아요
Type 5	형태소 분석 + 품사 태깅	(‘확실히’, ‘Adjective’)(‘달라’, ‘Noun’)(‘요’, ‘Josa’)(‘흡수’, ‘Noun’) (‘력’, ‘Suffix’)(‘좋네요’, ‘Adjective’)(‘아기’, ‘Noun’)(‘도’, ‘Josa’) (‘편해하는’, ‘Adjective’)(‘것’, ‘Noun’)(‘같아요’, ‘Adjective’)

4.3 속성 추출 모델 학습

3.2절에서 소개하였듯 입력된 단어 벡터들은 Attention 층을 거쳐 마지막 단계에서 속성 임베딩 행렬 T를 통해 차원 축소되고, 문장과 가장 연관성이 큰 속성을 확률값으로 나타내는 재구성된 문장 임베딩으로 출력된다. 입력단계에서는 5개 방식으로 전처리된 데이터가 한 문장 단위로 입력되며, 문장은 각각 Word2Vec, GloVe와 FastText를 통해 학습된 임베딩 행렬을 거쳐 단어 벡터로 변환되어 Attention 층으로 입력된다. 최종적으로 문장을 각 속성별 확률로 나타내주는 속성 임베딩 행렬 T는 단어 임베딩 공간에서 k-평균 군집화(k-means clustering)를 시행하여 중심(centroid) 값으로 초기화하였다. 학습 모델의 하이퍼파라미터인 속성의 개수 K는 14로 설정하였다. 예비 실험에서 K를 5, 10, 15, 20으로 설정하여 실험한 결과 K가 15일 때 해석이 가장 용이하였고 손실 값이 14일 때와 큰 차이를 보이지 않아 선행연구의 설정값과 동일하게 14로 실험을 진행하였다 (He *et al*., 2017). 모델의 학습 파라미터는 <Table 3>과 같으며, 배치 크기는 50, 최적화 방식으로는 아담(Adam)(Kingma and Ba, 2014)을 사용하였고 학습률은 0.001을 적용하였다. 신경망 반복 횟수(epoch)는 선행연구에서 제안한 15로 학습을 한 경우 모델의 손실 값이 수렴하지 못하여 충분한 학습을 위해 200으로 설정하였다.

Table 3. ABAE Model Hyperparameters

Category	Value
K	14
Regularization(λ)	1.0
Batch Size	50
Number of Epochs	200
Learning rate	0.001

Table 4. List of Manually Defined Aspects for Nappy and Wetwipe Reviews, which were Preprocessed by Type 3 Method, with Top Representative Words for Each Aspect

(a) representative words, aspect keyword and defined aspect of nappy domain

Representative words	Aspect keyword	Defined aspect
좋아요, 좋네요, 좋습니다, 좋아여, 좋고, 좋으네요, 조아요	만족감	해당없음
쓰는, 쓰던, 써오던, 하던, 썼던, 사는, 써오는, 있는, 쓰든	일반성	
합니다, 해요, 하였는데, 했습니다, 하고, 하였습니다, 했고	일반성	
쓰다가, 쓰다, 다시, 샀다가, 바꾸면서, 썼는데, 썼다가	일반성	
저렴하게, 싸게, 저렴히, 저렴한, 싼, 나은, 저렴해서, 싸서	가격	가격
가격, 금액, 품질, 값, 퀄리티, 질도, 비싼, 질, 성능, 혜택	가성비	
저렴히, 저렴하게, 싸게, 대에, 할인, 싼값, 가격, 경제	가격	
몸무게, 센티, 치고는, 키로, kg, 개월수, 체격, 체중, 통통	크기	크기
조금, 살짝, 약간, 좁아서, 짧고, 감싸는, 좁은, 짧은, 짧아서	크기	
배송, 정확하게, 배달, 급했는데, 보장, 받았네요, 감사합니당	배송	배송
쓰고, 키우고, 입히고, 경향, 떼고, 채워주고, 붙어, 입고	착용감	사용감
아기, 애가, 굵은, 하체, 맞지, 부들, 예민한, 민감한	착용, 촉감	
밤, 밤중, 저녁, 낮, 밤잠, 새벽, 자는데, 낮잠, 외출, 밤샘	밤새사용	흡수성
발진, 트러블, 두드러기, 습진, 땀띠, 알러지, 난적, 짓무름	피부이상	안전성

(b) representative words, aspect keyword and defined aspect of wetwipe domain

Representative words	Aspect keyword	Defined aspect
했어요, 했습니다, 했어용, 하였습니다, 합니다, 했네요	일반성	해당없음
쓰고, 샀어요, 샀네요, 쓰구, 썼는데, 썼어요, 쓰려구요	일반성	
믿고, 감사합니다, 만족하면서, 고맙습니다, 감사히, 만족	만족감	
써봤는데, 써보다, 쓰다가, 해봤는데, 해봤어요, 사봤는데	일반성	
쓰던, 사던, 쓰는, 썼던, 하던, 써오던, 시키는, 나은, 사는	일반성	
좋아요, 좋네요, 좋습니다, 좋아여, 괜찮아여, 좋더라구요	만족감	
하기, 하기에는, 하기에, 다니기, 닦기, 차안, 집안, 평상시	편의성	편리성
닦기, 닦고, 베란다, 요긴, 용도, 이유, 닦기에, 늘어납니다,	편의, 용도	
쓰고, 키우고, 쓸거에요, 경향, 닦아주고, 불편할, 종종	편의성	
빼서, 장이, 뽑혀, 씩, 한 장, 뽑아져, 줄줄이, 우르르, 연달	뽑힘성	
비싼, 훌륭한, 만족하는, 최대한, 알뜰하게, 저렴히, 구성	가성비	가격
배송, 배달, 친절하시고, 와요, 정확하게, 발송, 오전, 택배	배송	배송
작습니다, 작음, 아쉬워요, 작구요, 남네요, 크지만, 얇았어요	크기, 두께	크기/두께
감도, 촉촉함도, 질도, 향기도, 적당함, 은은하니, 톡톡하니	촉감	촉감

ABAE 모델의 학습이 끝나면 임베딩 공간상에서 코사인 유사도를 이용하여 각 상품평에 해당되는 속성과 가장 가까운 대표 단어를 추출한다. 본 연구에서는 도메인별로 전처리 형태에 따른 5개의 학습 결과를 도출하였고, 14개의 속성마다 추출된 30개의 대표 단어를 기준으로 각 속성을 나타내는 키워드를 선정하였다. 속성 추출 연구에 사용되는 영어 데이터 셋의 경우 gold-standard 속성 레이블이 정의되어 있으나, 한국어 속성 추출 데이터 셋은 존재하지 않기 때문에 본 연구에 사용된 기저귀와 물티슈 도메인에도 정의된 속성 레이블이 없다. 따라서 기저귀는 한국소비자원에서 발표한 기저귀 제품에 대한 성능 비교 결과(2014)를 참고하여 7개의 레이블을, 물티슈의 경우 온라인 쇼핑몰 상품평의 범주를 참고하여 6개의 기준 속성 레이블을 정의하였다. 도메인별 14개 속성 키워드는 키워드 종류에 따라 기저귀 7개(해당없음, 가격, 크기, 배송, 사용감, 흡수성, 안전성), 물티슈 6개(해당없음, 편리성, 가격, 배송, 크기/두께, 촉감)의 기준 속성 레이블에 수동으로 맵핑하였다. 배송 속성은 제품 자체의 특징과는 관계가 없으나 판매자 측면에서 반드시 고려해야 할 요소이며 상품평에서 많이 언급되는 항목이기에 포함하여 실험을 진행하였고 속성별 대표 단어와 속성 키워드, 기준 속성 레이블은 <Table 4>와 같다.

4.4 설문 조사 및 평가 지표

본 실험은 비지도 학습 기반으로 수행되어 각 상품평의 정답 레이블이 존재하지 않으므로 설문 조사를 통해 각 전처리 형태별 속성 추출 결과의 성능을 평가하였다. 기저귀 도메인에 해당되는 7개의 속성과 물티슈 도메인의 6개 속성을 대상으로 각 80개의 상품평을 추출하였다. 5개의 전처리 방식에 따라 ABAE를 학습하여 각각 속성 추출을 진행하였고, 5개 방식 중 3개 이상의 전처리 방식이 같은 속성으로 예측한 상품평을 우선적으로 추출하여 설문 문항으로 채택하였다. 설문 대상자들은 현재 육아를 하고 있으며 기저귀와 물티슈를 모두 사용해본 경험이 있는 사람으로 제한하여 선정하였다. 하나의 상품평에 대하여 기저귀 상품평은 7개 항목, 물티슈 상품평은 6개의 항목 중에서 하나의 속성을 선택할 수 있도록 문항을 작성하였다. 한 명의 참여자는 속성별로 10개씩 총 130개 상품평의 속성을 설문하며, 다섯 명이 동일한 상품평을 설문하도록 구성하였다. 설문은 총 40명을 대상으로 1,040문항에 대하여 진행하였고, 다섯 명 중에서 가장 많이 선택한 속성을 정답 레이블로 두고 5개 전처리 방식의 속성 추출 결과와 비교하였다. 다만 다중 속성을 포함한 상품평을 고려하여 top-2 정확도를 산출할 때는 차상위 속성까지 정답 레이블로 고려하였다.

각 상품평을 올바른 속성으로 분류하는 분류 성능을 평가하기 위해 <Table 5>의 혼동행렬(confusion matrix)을 사용하였다. 혼동행렬의 행은 사용자 관점에서 분류한 속성이고 열은 속성 추출 모델이 분류한 속성이며 a는 속성 클래스를 의미한다. 다중 클래스 분류의 평가 척도로 Accuracy와 Micro Average를 사용하였고 결과는 식 (10)~식 (13)으로 계산하였다(Ahn *et al*., 2019).

Table 5. Confusion Matrix of Multiple Classification

		Predicted aspect				
		a_1	a_2	…	a_K	Total
User assignment	a_1	n^{11}	n^{12}		n^{1K}	$\sum_{i=1}^{K} n^{1i}$
	a_2	n^{21}	n^{22}		n^{2K}	$\sum_{i=1}^{K} n^{2i}$
	…	…	…	…	…	…
	a_K	n^{K1}	n^{K2}		n^{KK}	$\sum_{i=1}^{K} n^{Ki}$
	Total	$\sum_{i=1}^{K} n^{i1}$	$\sum_{i=1}^{K} n^{i2}$	…	$\sum_{i=1}^{K} n^{iK}$	n

$$Accuracy = \frac{1}{K}\sum_{i=1}^{K} n^{ii} \tag{10}$$

$$MiA\,of\,Recall_{a_j} = \frac{1}{K}\sum_{j=1}^{K}\frac{n^{jj}}{\sum_{i=1}^{K} n^{ji}} \tag{11}$$

$$MiA\,of\,\mathrm{Precision}_{a_j} = \frac{1}{K}\sum_{j=1}^{K}\frac{n^{jj}}{\sum_{i=1}^{K} n^{ij}} \tag{12}$$

$$MiA\,of\,F1-score = \frac{1}{K}\sum_{j=1}^{K}\frac{2\times(Recall^{j}\times \mathrm{Precision}^{j})}{Recall^{j}+\mathrm{Precision}^{j}} \tag{13}$$

5. 실험 결과 및 성능 평가

본 연구에서 제안하는 한국어 전처리 방식을 적용한 Word2Vec+ABAE 모델의 성능 비교를 위해 비지도 방식의 속성 추출에 이용되는 LDA 모델과, GloVe+ABAE, FastText+ABAE 모델을 사용하였다. GloVe+ABAE, FastText+ABAE의 실험 환경은 Word2Vec+ABAE와 동일하게 설정하였고 LDA의 토픽 개수도 다른 두 모델과 동일하게 14로 설정하여 실험을 진행하였다. 성능 비교 결과는 <Table 6>, <Table 7>과 같다.

학습 결과, 식 (5)의 확률을 계산하여 가중치가 가장 높은 속성을 상품평의 속성으로 할당하여 속성을 추출하였고, 설문 조사 결과를 토대로 각 상품평에 하나의 속성을 정답 레이블로 두고 5개 전처리 방식의 속성 추출 성능을 평가하였다. <Table 6>을 보면 기저귀 도메인에서 가장 좋은 성능을 보인 전처리 방식은 형태소 분석을 수행한 Type 3이며, 이 방식은 네 가지 속성 추출 모델 전부에서 다른 방식보다 높은 성능 점수를 나타냈다. 자연어 처리에서 일반적으로 좋은 성능을 보이는 Type 4 방식보다 Type 3 방식이 좋은 성능을 보인 이유는, 여러 속성에 대한 서술이 다수 나열되어 있는 기저귀 도메인 상품평의 특성 때문인 것으로 추정된다. 네 가지 모델 중에서 Word2Vec+ABAE 모델이 품사태그를 부착한 Type 5 방식을 제외하고 LDA나 GloVe, FastText를 사용한 모델보다 성능이 좋았다. ABAE를 사용한 모델이 LDA보다 좋은 성능을 보인 것은 Attention 구조가 속성과 연관된 단어에 더 집중하고 비속성 단어는 강조하지 않는 특성이 학습에 반영된 것으로 판단된다.

Table 6. Evaluation of Nappy Domain

Preprocessing types	Method	Micro average			Top-1 Accuracy	Top-2 Accuracy
		Precision	Recall	F1-score		
Type 1	Word2Vec+ABAE	**0.328**	**0.339**	**0.318**	**0.305**	**0.402**
	GloVe+ABAE	0.245	0.250	0.232	0.250	0.363
	FastText+ABAE	0.227	0.221	0.220	0.221	0.302
	LDA	0.149	0.098	0.088	0.098	0.159
Type 2	Word2Vec+ABAE	**0.387**	**0.393**	**0.379**	**0.393**	**0.494**
	GloVe+ABAE	0.240	0.246	0.207	0.246	0.325
	FastText+ABAE	0.164	0.182	0.168	0.182	0.300
	LDA	0.112	0.114	0.101	0.114	0.212
Type 3	Word2Vec+ABAE	**0.513**	**0.518**	**0.510**	**0.525**	**0.664**
	GloVe+ABAE	0.288	0.270	0.274	0.270	0.393
	FastText+ABAE	0.354	0.348	0.340	0.348	0.427
	LDA	0.175	0.168	0.146	0.168	0.272
Type 4	Word2Vec+ABAE	**0.407**	**0.330**	**0.321**	**0.321**	**0.423**
	GloVe+ABAE	0.288	0.295	0.264	0.295	0.418
	FastText+ABAE	0.261	0.245	0.228	0.245	0.327
	LDA	0.090	0.102	0.084	0.102	0.236
Type 5	Word2Vec+ABAE	0.201	0.229	0.205	0.227	0.396
	GloVe+ABAE	**0.336**	**0.305**	**0.250**	**0.305**	**0.398**
	FastText+ABAE	-	-	-	-	-
	LDA	0.122	0.125	0.108	0.125	0.200

Table 7. Evaluation of Wetwipe Domain

Preprocessing types	Method	Micro average			Top-1 Accuracy	Top-2 Accuracy
		Precision	Recall	F1-score		
Type 1	Word2Vec+ABAE	**0.484**	**0.429**	**0.427**	**0.429**	**0.550**
	GloVe+ABAE	0.412	0.356	0.352	0.356	0.454
	FastText+ABAE	0.281	0.252	0.244	0.252	0.369
	LDA	0.184	0.175	0.161	0.175	0.281
Type 2	Word2Vec+ABAE	**0.477**	**0.463**	**0.465**	**0.463**	**0.565**
	GloVe+ABAE	0.217	0.185	0.163	0.185	0.310
	FastText+ABAE	0.275	0.252	0.250	0.252	0.344
	LDA	0.156	0.173	0.158	0.173	0.269
Type 3	Word2Vec+ABAE	**0.398**	**0.392**	**0.382**	**0.392**	**0.532**
	GloVe+ABAE	0.322	0.304	0.307	0.304	0.458
	FastText+ABAE	0.398	0.385	0.385	0.385	0.508
	LDA	0.151	0.110	0.109	0.110	0.220
Type 4	Word2Vec+ABAE	**0.588**	**0.471**	**0.469**	**0.471**	**0.613**
	GloVe+ABAE	0.324	0.279	0.283	0.279	0.410
	FastText+ABAE	0.322	0.292	0.290	0.292	0.413
	LDA	0.215	0.156	0.156	0.156	0.248
Type 5	Word2Vec+ABAE	0.301	0.273	0.278	0.273	0.394
	GloVe+ABAE	**0.382**	**0.340**	**0.328**	**0.340**	**0.496**
	FastText+ABAE	-	-	-	-	-
	LDA	0.083	0.094	0.080	0.094	0.223

<Table 7>을 보면 물티슈 도메인에서 가장 좋은 성능을 보인 전처리 방식은 형태소 분석 및 불필요 품사 제거를 수행한 Type 4이며 네 가지 모델에서 모두 다른 전처리 방식보다 높은 점수를 보였다. 네 가지 속성 추출 모델 중에서는 기저귀 도메인과 마찬가지로 Type 5 방식을 제외하고 Word2Vec+ABAE 모델에 사용한 방식의 성능이 가장 높은 지표를 나타냈다. 기저귀와 완전히 다른 물티슈 도메인에서도 속성과 연관된 단어에 초점을 맞춰서 학습하는 Attention 구조의 이점을 확인하였다.

예외적으로, 품사 태깅을 한 Type 5 방식의 경우 GloVe+ABAE 모델에 적용했을 때 기저귀와 물티슈 도메인 모두에서 가장 좋은 성능을 보였다. 이것은 GloVe가 전체 말뭉치에서 단어와 품사가 조합된 단위로 동시 발생 확률을 학습하는 방식이 속성과 관련된 단어 학습에 도움을 주었기 때문인 것으로 추정된다.

그러나 전반적으로 두 도메인에 대해서 영어 데이터의 속성 추출 결과와 비교했을 때 성능이 낮은 이유는 데이터 셋의 차이이다. 영어 상품평을 대상으로 진행한 선행연구(He *et al*., 2017)에서는 평가의 모호성을 피하기 위해 속성 레이블이 하나인 문장만을 사용하였다. 또한 단어 사용 등에서 명확한 유형을 보이지 못하는 다른 속성 레이블은 평가에서 제외하여 Food, Staff, Ambience 세 가지 주요 속성만 평가에 사용하였다. 본 연구에서 사용한 데이터 특성을 살펴보면 기저귀의 경우 구매자는 본인의 경험과 실제 수혜자인 영유아의 반응에 민감하게 반응하여 상품평을 기술하였다. 따라서 기저귀를 다각도로 평가하여 하나의 상품평에 수 개의 속성을 포함하는 경우가 많았다. 이것은 물티슈 도메인에도 동일하게 적용되는 부분이며 추가적으로 구매자 자신이 사용하는 경우에도 상품의 여러 측면을 고려하여 작성하는 경우가 많았다. <Table 8>은 Type 3, Type 4 방식으로 전처리한 Word2Vec+ABAE 모델이 할당한 속성과 다중 속성을 가진 상품평의 정답 레이블을 나타낸 예시이다.

<Table 8>의 첫 번째 상품평을 보면 “같은사이즈라도 이건 길이가 길어서 좋아요 재질도 부드럽구요”라는 문장을 대상으로 설문자들은 사용감 속성을 부여하였다. 그러나 사이즈, 길이라는 단어는 크기 속성을 나타내고 있어 Type 3과 Type 4 방식은 이 상품평에 크기 속성을 할당하였다. 즉 하나의 상품평이 다중 속성을 포함하는 경우가 다수 발생하여 단일 속성을 기준으로 하는 평가 성능이 낮게 나타났다.

최종적으로 성능 지표를 고려했을 때 기저귀 도메인에서는 Type 3 방식이, 물티슈 도메인에서는 Type 4 방식이 속성 추출에 가장 적합한 전처리였다. 두 도메인 모두 하나의 상품평이 두 개 이상의 속성을 포함하는 경우가 많은데 특히 기저귀 도메인이 물티슈 도메인보다 더 많은 속성을 포함하고 여러 가지 속성이 하나의 상품평에 표현되는 경우가 많았다. 이러한 데이터의 특성 차이 때문에 가장 좋은 성능을 보인 전처리 방식이 서로 다르게 나타난 것으로 생각된다.

Table 8. Examples of Multi-Aspects Reviews

No.	Multi-aspects reviews			
1	같은사이즈라도 이건 길이가 길어서 좋아요 재질도 부드럽구요			
	Top-1 label	Top-2 label	Type 3	Type 4
	사용감	크기	크기	크기
2	밤에 자꾸 새는거 같지만 밴드도 되고 팬티도 되는게 편해여			
	Top-1 label	Top-2 label	Type 3	Type 4
	사용감	흡수성	흡수성	흡수성
3	9키로부터 5단계 썼어요 4단계는 좀 작은 느낌이 있어요 부드러운 편이고 알갱이 안 나와서 좋아요			
	Top-1 label	Top-2 label	Type 3	Type 4
	사용감	크기	크기	크기
4	이물티슈가젤좋아요 크기 두께 수분량 적당하고요 가격도 100장캡인데 만원도 안되고요 낱장으로 잘뽑혀요			
	Top-1 label	Top-2 label	Type 3	Type 4
	촉감	편리성	크기/두께	편리성
5	생각보다 얇은데 청소용이라 상관없어요 티슈가 줄줄이 나와서 좀 성가셔요			
	Top-1 label	Top-2 label	Type 3	Type 4
	편리성	크기/두께	크기/두께	편리성
6	처음엔 상품후기 좋아서 구내해서 썼지만 이젠 품질에 반해서 믿고 씁니다 가격도 착한데 두툼하고 장수많고 성분도 좋다고하니 저희집은 반려견때문에 물티슈 많이써요 요거로 양치도 해주고 사료먹고나면 입도 닦아주고			
	Top-1 label	Top-2 label	Type 3	Type 4
	촉감	편리성	가격	편리성

본 연구에서 제안한 Type 3과 Type 4 두 가지 방식의 전처리 조합이 적합한 첫 번째 이유는 띄어쓰기 교정이 형태소 분석의 성능에 영향을 주기 때문이다. 형태소는 단어를 구성하는 기본 단위인데 한국어는 형태소의 조합으로 발음의 기본 단위인 어절을 생성한다. 어절은 띄어쓰기의 단위와 대부분 일치하는데 띄어쓰기가 지켜지지 않으면 여러 개의 품사가 붙어서 하나의 어절을 형성하게 되고 형태소 분석 과정에서 아예 다른 형태소의 조합으로 분리할 수 있다. 따라서 띄어쓰기를 지키지 않는 상품평이 많을수록 띄어쓰기를 교정하지 않고 형태소 분석할 경우 분석의 오류가 커질 수밖에 없다. 두 번째 이유는 형태소 분석을 적용하여 형태소 단위로 임베딩이 학습되기 때문이다. 학습에서 사용한 Word2Vec의 CBOW 방식은 윈도우 크기만큼의 주변단어를 통해 중심단어를 예측한다. 예를 들어 '아버지가 방에 들어가신다'를 Word2Vec으로 학습하면 '아버지가'와 '들어가신다'를 통해 '방에'를 예측한다. 이 경우 '-가', '-에'와 같은 조사의 형태에 따라 다른 단어로 취급하여 어간이 같더라도 다른 벡터값을 가지게 된다. 그러나 형태소 분석을 수행하고 임베딩 하면 같은 어간일 경우 같은 벡터값을 가질 수 있다. 따라서 형태소 분석 후 임베딩을 수행한 것이 학습에 더 적합한 벡터를 얻을 수 있었다. 마지막 이유는 불필요한 품사를 제거하여 정보를 가진 형태소만 사용했기 때문이다. 명사/형용사/동사/부사/영어/숫자 외의 품사를 제거하고 임베딩을 수행하면 조사, 어미, 'ㅋㅋ'와 같은 한국어 자모음 등은 학습에서 제외되고 제품의 특성과 성질, 그에 대한 사용자의 평가와 관련된 정보만 임베딩 되어 학습에 사용되므로 효율적인 학습을 수행할 수 있었다.

세 가지 임베딩 방식 중에서는 ABAE 모델에 Word2Vec 임베딩을 적용한 방식이 가장 효과적이었다. FastText와 GloVe가 다양한 자연어처리 과제에서 좋은 성능을 보이나, 본 연구에서 사용한 ABAE 모델에는 적합하지 않았다. ABAE 모델을 통해 속성별 대표 단어를 추출한 결과, Word2Vec을 이용한 경우에는 속성과 관련된 유사 의미의 단어들을 추출했으나 FastText와 GloVe는 속성과 관련되지 않은 단어들을 더 많이 추출하였다. FastText는 단어의 형태적 부분에 집중하고, GloVe는 단어의 동시 등장에 중점을 두어 임베딩을 학습하는 반면에, Word2Vec은 문맥적인 부분에 집중하여 학습하는 특성을 가진다. 따라서 세 가지 임베딩을 ABAE 모델에 적용했을 때 Word2Vec이 더 효과적인 결과를 보인 것이라 추정된다.

6. 결론 및 향후 연구 방향

본 연구는 한국어로 작성된 온라인 상품평을 대상으로 비표준어가 많은 한국어 상품평의 속성 추출에 적합한 전처리 방식과 임베딩 방법을 제안하였다. 또한 한국어 속성 추출을 위해 비지도 방식의 ABAE 모델을 적용하여 의미있는 결과를 도출하였다. 실험을 위해 온라인 쇼핑몰 사이트에서 기저귀, 물티슈 제품에 대한 상품평을 수집하여 각 상품평을 본 논문에서 제안한 5가지 방식으로 전처리하였다. 전처리 된 상품평은 Word2Vec 모델을 이용하여 200차원으로 임베딩 되어 속성 추출 모델에 입력되고, Attention 구조를 거친 각 문장은 속성별 확률값을 나타내는 벡터로 재구성된다. 학습 결과, 확률의 산술평균값이 가장 높은 속성을 상품평의 속성으로 할당하였고, 설문 조사를 통해 얻은 사용자의 인지적 평가를 토대로 전처리 형태에 따른 속성 추출 성능을 측정하였다. 기저귀 도메인에서는 특수문자 제거, 띄어쓰기 교정과 형태 소 분석을 수행한 방식이, 물티슈 도메인에서는 특수문자 제거, 띄어쓰기 교정, 형태소 분석, 불필요 품사 제거를 수행한 방식이 효과적이었다. 본 연구를 통해 한국어 상품평에는 특수문자 제거, 띄어쓰기 교정과 형태소 분석을 수행하고, 도메인에 따라 불필요 품사를 제거한 후에 Word2Vec 방식으로 임베딩 하는 것이 속성 추출에 효과적이라는 결론을 얻을 수 있었다. 또한 연구가 미비한 한국어 상품평 속성 추출에 Attention 기반 비지도 학습을 적용하여 의미 있는 수준을 달성할 수 있었다.

연구를 진행하면서 다수 사용자가 여러 가지 속성을 하나의 상품평에 작성한다는 특성과, 다중 속성을 가진 상품평의 경우에 하나의 정답 속성과 차이가 발생하여 오차가 커지는 결과를 확인하였다. 이 점을 활용하여 온라인 쇼핑몰을 운영하는 생산자들이 제품의 세부 속성별로 상품평을 작성하도록 양식을 제시하고 상품평을 받는다면 제품과 서비스의 개선을 위한 구체적인 데이터로 사용할 수 있을 것으로 기대된다. 또한 향후 연구에서는 정제되지 않은 상품평의 특성을 고려하여 n-gram, BPE 등의 전처리 방식을 형태소 분석과 함께 활용한다면, 한국어 상품평 속성 추출의 성능을 발전시킬 수 있을 것으로 예상된다.

참고문헌

Ahn, G., Yoo, J. H., Lee, S. H., and Kim, S. B. (2019), Explainable Convolutional Neural Networks for Multi-Sensor Data, *Journal of the Korean Institute of Industrial Engineers*, **45**(2), 146-153.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), Latent Dirichlet Allocation, *Journal of Machine Learning Research*, **3**, 993-1022.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017), Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, **5**, 135-146.

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2017), An unsupervised Neural Attention Model for Aspect Extraction, In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, **1**, 388-397.

Hu, M. and Liu, B. (2004), Mining and Summarizing Customer Reviews, *In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168-177.

Jung, S. (2018), A Study on the Difficulties and Coping Strategies in the First Purchase of Consumers-Focusing on the Purchase of Baby care product, Doctorial dissertation, *Graduate School of Seoul National University*.

Jeong, J., Mo, K. H., Seo, S., Kim, C. Y., Kim, H., & Kang, P. (2018),

Unsupervised Document Multi-Category Weight Extraction based on Word Embedding and Word Network Analysis : A Case Study on Mobile Phone Reviews, *Journal of the Korean Institute of Industrial Engineers*, **44**(6), 442-451.

Jang, J. (2002), Sentence Expression Dictionary, Moonjangyeongusa, Seoul, Korea.

Kingma, D. P. and Ba, J. (2014), Adam : A method for stochastic optimization, *arXiv preprint arXiv : 1412.6980*.

Korea Consumer Agency, (2014), Disposable Panty-Type Diaper (for Infants) Quality Comparison Test Results Report.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013), Efficient Estimation of Word Representations in Vector Space, *arXiv preprint arXiv: 1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013), Distributed Representations of Words and Phrases and their Compositionality, *In Advances in Neural Information Processing Systems*, 3111-3119.

Mnih, V., Heess, N., and Graves, A. (2014), Recurrent Models of Visual Attention, *In Advances in Neural Information Processing Systems*, 2204-2212.

Oh, Y., Kim, M., and Kim, W. (2019), Korean Movie-Review Sentiment Analysis Using Parallel Stacked Bidirectional LSTM Model, *Journal of KIISE*, **46**(1), 45-49.

Park, H., Song, M., and Shin, K. (2018), Sentiment Analysis of Korean Reviews Using CNN-Focusing on Morpheme Embedding, *Korea Intelligent Information System Society*, **24**(2), 59-83.

Park, S. and On, B. (2017), Latent Topics-based Product Reputation Mining, *Korea Intelligent Information System Society*, **23**(2), 39-70.

Pennington, J., Socher, R., and Manning, C. D. (2014), Glove : Global Vectors for Word Representation, *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011), Opinion Word Expansion and Target Extraction through Double Propagation, *Computational Linguistics*, **37**(1), 9-27.

Rana, T. A. and Cheah, Y. (2016), Aspect Extraction in Sentiment Analysis : Comparative Analysis and Survey, *Artif Intell Rev*, **46**, 459-483.

Wan, C., Peng, Y., Xiao, K., Liu, X., Jiang, T., and Liu, D. (2020), An Association-Constrained LDA Model for Joint Extraction of Product Aspects and Opinions, *Information Sciences*, **519**, 243-259.

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2017), Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms, *31st AAAI Conference on Artificial Intelligence*, 3316-3322.

Xu, H., Liu, B., Shu, L., and Yu, P. S. (2018), Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction, *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, **2**, 592-598.

Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., and Zhou, M. (2016), Unsupervised Word and Dependency Path Embeddings for Aspect Term Extraction, *ArXiv Preprint ArXiv:1605.07843*.

저자소개

박명현 : 육군사관학교에서 무기시스템공학과에서 학사학위를 취득하고 고려대학교 산업공학과 석사과정에 재학중이다. 연구분야는 데이터 마이닝, 자연어처리이다.

최희련 : 단국대학교 산업공학과에서 1993년 학사학위를 취득하고 고려대학교 산업공학과에서 석사 및 박사수료를 하였다. 2007년부터 고려대학교 산업경영공학부 강사로 재직하고 있으며, 연구분야는 생산시스템, 인공지능 및 Blockchain이다.

이홍철 : 고려대학교 산업공학부에서 1983년 학사, Texas Arlington 대학교 산업공학과에서 1988년 석사학위를 취득하고 Texas A&M대학교에서 산업공학 박사학위를 취득하였다. 1996년부터 고려대학교 산업경영공학부 교수로 재직하고 있다. 연구분야는 AI, 생산공학시스템, 시뮬레이션이다.