클러스터링 기법을 활용한 관광지 대표문장 추출

김다희*, **, 이강우**, 임지원**, 홍순구**, ***
*동아대학교 컴퓨터공학과
**동아대학교 스마트거버넌스 연구센터
***동아대학교 경영정보학과
kdahee0598@gmail.com, kangwooster@gmail.com, ji1e@naver.com, shong@dau.ac.kr

Extracting Representative Sentences about Tourist Sites Using a Clustering Method

DaHee Kim*, **, KangWoo Lee**, JiWon Lim**, Soon-Goo Hong**, ***
*Dept. of Computer Engineering, Dong-A University
**Smart Governance Research Center, Dong-A University
**Dept. of Management Information System, Dong-A University

요 약

'파리의 더러운 지하철', '런던의 비싼 물가' 등 관광지에 대한 몇 마디 말은 관광지를 직관적으로 이해하는데 도움을 준다. 관광지에 대한 직관적 평가를 파악하기 위해서 클러스터링 기법을 사용하였다. '주차', '경치', '시설'과 같은 다양한 라벨을 부여하여 클러스터링을 비교한 결과 '주차', '경치' 등 비슷한 문맥의 리뷰가 같은 클러스터로 묶인 것을 확인할 수 있었고, 각 분야의 문맥을 파악하기 위해 대표문장을 추출하였다. 각 분야의 대표문장은 해당 분야의 평가를 잘 파악할 수 있었고, 해당 분야의 만족도뿐만 아니라 불편사항 등을 이해하는데 도움을 준다.

1. 서론

관광 분야에서 온라인 리뷰는 관광지에 대한 정보를 제공한다. 이러한 리뷰 데이터는 의사결정과정에서 필수적인 역할을 한다. 하지만 관광지에 대한 이러한 온라인 리뷰 데이터를 클러스터링하고 대표문장을 추출하는 분석과정을 거친다면 관광지 방문자들의 관광지 평가를 파악하기 쉬울 것이다.

따라서 본 연구에서는 관광지 리뷰 데이터를 수집하고 관광지의 세부적인 평가를 파악하고자 한다. 복합적인 정보를 담고 있는 리뷰 데이터를 세부적으로 클러스터링하여 분야별로 분류하고, 해당 분야의 평가 정도를 한눈에 파악할 수 있다. 대표문장을 추출함으로써 각 분야의 문맥을 파악해 보고자 한다. 이를 위한 구체적인 연구문제는 다음과 같다.

연구문제 1. 국내 관광지 리뷰의 클러스터링 결과는 어떠한가?
연구문제 2. 국내 관광지 리뷰의 클러스터의 중심점은 각 클러스터를 대표할 수 있는가?

관광지 리뷰 데이터를 분석하여 다양한 리뷰 데이터를 분야별로 한눈에 파악하고, 각 분야의 대표적인 문맥을 파악함으로써 해당 관광지의 시설, 경치 등 다양한 분야의 불편사항을 인지할 수 있다. 이를 통해 관리자는 관광지의 세부적인 평가를 파악하는 것에 있어서 도움이 되는 기초 자료를 수립할 수 있으며, 방문객 또한 관광지의 평가를 한눈에 파악하여, 한층 좋은 관광을 할 수 있는 자료를 제공할 수 있으리라 기대한다.

2. 관련연구

(1) BERT

자연어 문장을 기계가 이해하기 위해선 문장 분석을 수행해야 한다. BERT는 이러한 문장 분석을 수행하기 위해 wiki 등에서 발췌한 unlabeled data를 대규모로 학습하고, 특정 task에서 역할을 수행하기 위해 task-specific한 fine-tuning을 진행한다. 기존 연구는 단방향의 분석을 수행하여 문맥을 파악하는 것에 반해 BERT는 양방향의 분석을 수행하여서 문장 벡터화를 보다 효율적으로 수행한다. 이러한 BERT 모델은 실제 영어, 한국어 데이터에서 효과적으로 작동한다[1].

(2) SKT KoBERT

BERT는 영어뿐 아니라 다양한 언어를 학습시킨 외국어 모델인 BERT base multilingual cased

model이 존재한다. SKTBrain팀에서 연구한 KoBERT는 기존 BERT base multilingual cased model의 정확도인 0.875보다 높은 0.901을 기록했다. 한국어 위키 문장 500만 개, 단어 5400만 개를 사용했고, 한국어 뉴스 문장 2000만 개, 2억 7000만 단어를 사용해 학습을 진행하였다. 이 모델의 경우 BERT 알고리즘을 활용하여 데이터와 파라미터를 한국어 문장에 맞게 설정해 정확도를 높인 모델이다[1][2].

(3) Kakaobrain KorNLUDataset [3]

- NLI, STS

NLI는 전제와 가설 두 문장을 입력받아 그 관계를 entailment, contradiction, and neutral로 분류한다.

STS는 두 문장 사의의 의미적 유사성을 평가하고, 모델이 의미상 두 문장의 친밀도를 얼마나 잘 파악하는지 또는 문장의 의미적 표현을 얼마나 잘 구현하는지 평가하기 위해 사용된다.

- KorNLUDataset

NLI(Natural language inference)와 STS(semantic textual similarity)는 NLU(natural language understanding)의 주요 과제이다. 하지만 한국어 NLI와 STS 데이터 셋은 존재하지 않기 때문에 Kakaobrain팀에서 KorNLI와 KorSTS를 개발하였다. KorNLUDataset은 한국어 NLU를 위한 새로운 데이터 셋이다.

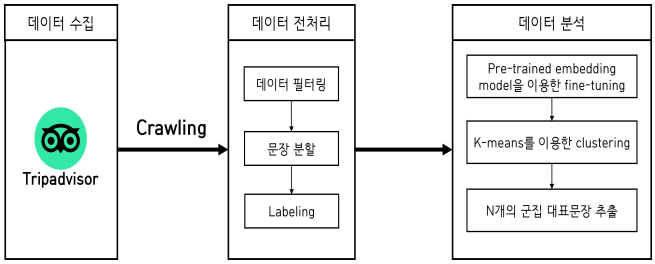
(4) Ko-Sentence-BERT-SKTBERT

Ko-Sentence-BERT-SKTBERT는 Siamese BERT-Networks와 SKT의 KoBERT, kakaobrain의 KorNLUDataset을 통해 학습시킨 모델이다[4].

3. 관광지 리뷰 분석을 위한 클러스터링

본 논문에서는 국내 관광지 방문객의 리뷰를 분야별로 분류하여 해당 분야에서 관광객의 평가가 어떠하였는지 알아보기 위해 관광지 리뷰를 클러스터링하고자 하였다. 그리고 각 클러스터의 중심점이 해당 클러스터를 대표할 수 있는지 알아보고자 한다.

본 연구는 리뷰 데이터를 분석하여 나타나는 특징을 파악하기 위해 (그림 1)과 같이 연구를 설계하였으며 데이터 수집, 전처리, 분석으로 구성된다.



(그림 1) 대표문장 추출을 위한 알고리즘 구조

데이터 수집

데이터 수집 단계에서는 TripAdvisor의 2007년 8월부터 2021년 5월까지의 국내 부산 지역 관광지 리뷰 데이터 27,752개를 수집하였다[5]. 수집한 데이터는 관광지명, 언어, 날짜, 리뷰, 평점 등의 정보를 포함한다.

데이터 전처리

데이터 수집 후 분석을 하기에 알맞은 데이터 형식으로 가공하기 위한 과정이다.

(1) 데이터 필터링

TripAdvisor에는 중국어, 일본어, 영어 등 다양한 언어 리뷰 데이터가 존재한다. 한국어를 제외한 다른 언어는 제거하고 한국어 리뷰 데이터만을 추출하였다.

(2) 문장 분할

관광지 리뷰 안에는 관광지의 경관이나 시설, 교통 등의 다양한 분야에 대한 평가가 복합적으로 작성되어있다. 이러한 리뷰 데이터를 구체적으로 평가하기 위해 여러개의 문장으로 작성된 리뷰를 한 문장씩 분리하였다. 본 논문에서는 파이썬 한국어 문장 분리기인 'Korean Sentence Splitter(KSS)'을 활용하여 문장 분리를 수행하였다[6].

(3) Labeling

데이터에 대한 분석 성능을 평가하기 위해 분석할 데이터를 추출하여 Labeling 작업 해주었다. 앞서 언급했듯이, 리뷰 데이터는 해당 관광지의 음식, 시설, 경치, 주차와 같은 다양한 평가가 포함되어 있다. 그 중 '주차/교통', '경치/풍경', '음식', '시설', '인파', '가격' 6개의 라벨로 분류하였다. 임의로 500개의 리뷰 데이터를 추출하여서 6개의 라벨을 부여하였다.

데이터 분석

(1) Loading pre-trained BERT Model

전처리가 끝난 데이터는 클러스터링을 하기 위해

서 임베딩한다. 한국어 데이터를 바탕으로 사전 학습된 모델인 'Ko-Sentence-BERT-SKTBERT'를 사용하였다. 임베딩은 'Ko-Sentence-BERT-SKTBERT'에서 제공하는 Pre-Training된 모델 중 STS 데이터로만 학습한 'training_sts' 모델을 사용하여 임베딩하였다.

(2) Clustering

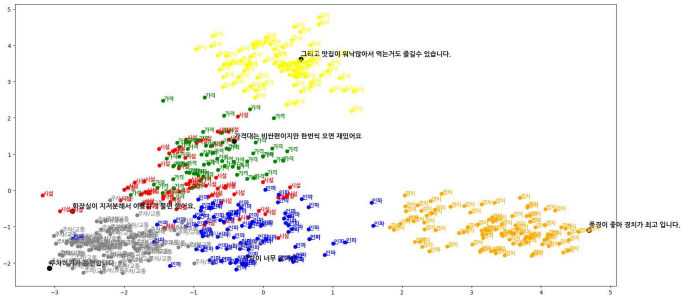
'Ko-Sentence-BERT-SKTBERT'를 사용하여 임베딩된 데이터를 클러스터링한다. N개의 그룹으로 클러스터링을 해주는 파이썬 Scikit-learn의 'K-means'를 사용하여서 임베딩된 데이터를 바탕으로 클러스터링을 수행하였다[7]. 클러스터의 수는 Labeling 과정에서 6개의 라벨로 분류하였기 때문에 6개의 그룹으로 클러스터링을 진행하였다. 클러스터링된 결과를 principal component analysis (PCA)를 사용하여 2차원으로 나타낸 뒤 클러스터링의 결과를 라벨과 비교한다.

(3) 대표 문장 추출

'K-means'를 통해 클러스터링된 결과의 문맥을 알아보기 위하여 각 클러스터의 대표 문장을 추출한다. 대표 문장을 추출하기 위해서는 'K-means'에서 제공하는 'cluster_centers_' 파라미터를 사용하여 각 클러스터의 중심점을 알아낸다. 해당 중심점과 가장 유사한 리뷰 데이터를 대표 문장으로 추출한다. 각 클러스터의 중심점과 리뷰 데이터를 코사인 유사도를 통하여 가장 유사도가 높은 대표 문장을 추출한다.

결과 분석

본 절에서는 1절, 2절, 3절에서 제시한 연구 방법을 기초하여 연구 결과를 분석하였다.



(그림 2) K-means를 이용한 임베딩 리뷰 문장 클러스터링

사람이 직접 부여한 라벨은 그래프의 점 위에 텍스트로 출력하였다. K-means에서 클러스터링한 결과는 색상으로 나타내었다. 같은 클러스터로 분류할 경우 같은 색상을 띠게 하였다. <그림 2>의 결과를 살펴보면 두 개의 데이터를 제외하고는 모든 데이터가 사람이 부여한 라벨과 색상이 일치하는 것으로 나타난다.

먼저, 명확하게 클러스터링이 된 '경치' 분야의 데이터와 클러스터링 결과를 비교해보자면, A 열이 사람이 직접 남긴 리뷰 데이터이고, B 열이 직접 부여한 라벨, C 열이 K-means가 클러스터링한 결과이다. '경치'와 관련된 데이터가 가장 많았고 이외에도 '야경', '풍경'과 같이 동일한 분야에 대한 리뷰 데이터가 같은 클러스터로 분류되었다. '경치' 클러스터의 대표문장은 '풍경이 좋아 경치가 최고입니다.'로 추출되었다. <그림3>의 리뷰 데이터와 대표문장을 비교해보면 같은 맥락을 말하고 있는 것으로 보아 클러스터의 중심점과 유사한 데이터는 '경치' 클러스터를 대표하는 문장으로 볼 수 있다.

	A	B	C
103	경치가 최고입니다.	경치	2
104	야경이 예쁩니다.	경치	2
105	경치가 완전 좋습니다.	경치	2
106	공기도 맑으며 일단 경관이 너무나 뛰어납니다.	경치	2
107	아름다운 경치 경치가 아주 아름답습니다.	경치	2
108	경치 좋아요	경치	2
109	경치가 멋있습니다.	경치	2
110	경치가 넘 멋졌습니다.	경치	2
111	관리가 매우 잘 돼서 경치가 다 예뻐요.	경치	2
112	경치가 좋아요	경치	2
113	경치가 좋아요	경치	2
114	경치가 좋아요	경치	2
115	좋은 경치를 볼 수있습니다.	경치	2
116	경치가 아주 멋져요!!	경치	2
117	뭉경도좋아요	경치	2
118	예뻐요~ 멀리서 보는 것도 예쁘지만~ 광안대교를 건너면서 주변을 볼때도 경치가 너무 좋아요~~ 야경도 정말 예쁩니다	경치	2
119	너무 예쁜 경치! 경치가 너무 예쁜곳 입니다.	경치	2
120	날씨가 좋아서 그런지 사진도 잘 나오고 경치가 엄청 좋았어요.	경치	2
121	우선 경치가 너무 좋았어요.	경치	2
122	날씨가 좋아서 경치가 정말 멋졌어요.	경치	2
123	너무 이뻐요 ㅠㅠ~~ 경치가 죄입니다	경치	2
124	경치가 아름다운 곳 방문했을때 날씨가 아주 화창했는데요.	경치	2
125	풍경 멋져요.	경치	2

(그림 3) '경치' 데이터

다음으로는 뭉쳐있는 클러스터인 '인파', '시설' 클러스터를 같이 살펴본다.

	A	B	C
43	화장실 냄새 너무나 심했음	시설	4
44	화장실관리 화장실에서 너무냄새가나고.. 관리가 어려우면 화장실을 멀리 옮기든가 외국분들도 많이 방문	시설	4
45	화장실관리가 개판이고 냄새가 너무많이나서 외국인 분들께 너무창피스러웠음.	시설	4
46	너무 냄새가 심합니다.	시설	4
47	근처에 화장실 깨끗하고요.	시설	4
48	그리고, 하수구 냄새가 너무 심합니다.	시설	4
49	화장실이 지저분해서 이용하기 불편 했어요.	시설	4
50	하지만 냄새나 물에 젖은 바닥 등이 구경하기 불편한 부분도 있더라구요	시설	4
51	중간에 화장실도 사용자가 적어 깨끗하게 관리되고 있습니다.	시설	4
52	화장실은 다소 적고 청결하진 않지만 그래도 각 매점에는 핫한 아이템들이 많으니 조개속의 진주를 발견하	시설	4
53	역사 내부는 깨끗하고 쾌장히 큽니다.	시설	4
54	화장실도 크게 잘 관리되어 있어요.	시설	4
55	화장실도 깨끗하고 칸이 많아서 사람이 많아도 빨리 사용할 수 있었던 것 같습니다	시설	4
56	예전에는 화장실문제가 좀 심었는데 나름 관리되고 있는 화장실도 괜찮았다	시설	4
57	다만 유명한 관광지임에도..화장실 상태가....	시설	4
58	이곳저곳 널어져있는 쓰레기와 그에 동반된 악취, 그리고 해변가까이의 수많은 흡연가는 당신의 기분을 나쁘	시설	4
59	단 쓰레기 뒷정리가 잘안되고 냄새 날때가 있어요	시설	4
60	정말로 세련된 카페였습니다만 입구부터 카페로 올라가는 계단의 문을 여니 엄청난 하수구 냄새가...	시설	4
61	화장실이 좀 멀어서 불편했습니다.	시설	4
62	생선냄새가 너무 많이났고 벌레가 많이 날아다녔습니다	시설	4
63	관광객들이많아 좀 여수선하지만 목욕탕은 생각보다 깨끗했습니다.	시설	4
64	찜질방 오랜만에 방문했는데 관리가안되는지 청소상태도엉망이고 사람도 별로없네요	시설	4
65	하지만 사람들이 많아지면서 쓰레기 문제등이 겹치면서 점점 질이 저하되고 있는 곳입니다.	시설	4

(그림 4) '시설' 데이터

	A	B	C
100	사람도 많았고요.	인파	0
101	여긴 사람이 항상 많네요.	인파	0
102	사람도 참 많구요	인파	0
103	생각보다 엄청 크고 사람들도 많네요.	인파	0
104	하지만 사람이 너무 많아서 시끄러웠어요	인파	0
105	사람이 너무 많아서 서서보긴 했지만 너무 좋네요.	인파	0
106	사람 많음	인파	0
107	주말엔 사람이 너무 많아요	인파	0
108	다 좋은데 사람 너무 많음	인파	0
109	여름엔 사람이 너무 많은데다가 잡상인도 너무 많다.	인파	0
110	사람도 많았구요.	인파	0
111	생각보다 사람도 적어서 좋았어요	인파	0
112	경치는 멋있으나 사람이 너무 많아요	인파	2
113	방송에 소개된 덕에 더 사람이 많은듯합니다.	인파	0
114	주말 성수기에는 사람이 너무 많아요.	인파	0
115	사람들이 넘쳐나는 곳 유동인구가 워낙 많아서 사람사는곳 같다고 느끼는 반면 너무 복잡했던 기억이네요	인파	0
116	여름은 역시 사람이 너무 많군요.	인파	0
117	사람도 많지 않았어요	인파	0
118	한번씩 가기 좋았는데 요즘은 사람이 너무너무 많아져서 갈 엄두가 안납니다.	인파	0
119	사람이 너무 많다는 불편함이 있습니다	인파	0
120	이 지역에 정말 많아요.	인파	0
121	매번갈때마다 사람이 엄청많습니다	인파	0
122	옆에 해운대는 사람이 굉장히 많았는데 이곳은 왠일인지 사람이 별로 없더라구요.	인파	0

(그림 5) '인파' 데이터

<그림 4>는 '시설' 라벨의 데이터는 관광지에 대한 청결도나 불편함 등의 시설에 대한 평가가 있는 데이터이다. <그림 5>는 '인파' 라벨의 리뷰 데이터이다. '인파' 라벨은 사람들의 방문도에 대한 평가가 있는 데이터이다. '인파', '시설'의 데이터들이 <그림 2>의 그래프를 보시면 상당 부분이 겹치는 것으로 보인다. 다른 클러스터임에도 비슷하게 그래프에 나타나는 것으로 보아서 클러스터링의 결과가 좋지 않다고 분석할 수 있다. 하지만 '시설'의 리뷰 데이터를 보면 '사람이 많아지면서 쓰레기 문제 등이 겹치면서 점점 질이 저하되고 있는 곳입니다' 등과 같이 사람들의 방문도에 대한 평가와 시설에 대한 평가가 합쳐져 있는 리뷰 데이터임을 알 수 있다. 이와 같은 '시설' 리뷰의 경우, '인파' 분야의 리뷰 데이터와 유사하다 분석할 수 있으며, 클러스터링의 결과 또한 정확도가 높다고 분석할 수 있다.

4. 결론

부여된 라벨과 K-means로 클러스터링한 결과를 비교해본 결과 클러스터링의 정확성이 높다고 볼 수 있었다. 또한, 각 클러스터의 문맥을 파악하기 위해 대표문장을 추출하여 올바른 문맥을 나타내는지 비교해보았다. 클러스터의 대표문장을 추출하기 위해서 클러스터의 중심점의 값과 가장 유사한 값을 가지는 리뷰 데이터를 추출하여 해당 클러스터의 대표문장으로 선택했다. 각 클러스터의 대표문장과 클러스터 안의 리뷰 데이터를 직접 비교해보았을 때, 중심점과 유사한 데이터는 해당 클러스터를 대표하는 문장으로 볼 수 있었다. 이러한 대표문장은 특정 관광지 혹은 특정 키워드를 담고 있는 데이터를 대표하는 문장을 함축하는 것으로 볼 수 있다. 도출된 결과는 관광 분야의 기초 자료로 활용이 가능하다. 관광지 이용자들이 시설에 대해 어떻게 평가하였는지 또, 관리자들은 해당 시설이 불편사항을 인지하여 개선한다면 관광지의 만족도가 향상될 수 있을 것이다.

Acknowledgement

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018S1A3A2075240)

참고문헌

[1] 이한범, 구자환, 김응모, "BERT를 활용한 문장 감정 분석 연구", 2020 온라인 추계학술발표대회, 제27권, 제2호 pp. 909-911, 2020

[2] "SKTBrain/KoBERT", https://github.com/SKTBrain/KoBERT, 2019

[3] J. Ham, Y. Choe, K. Park, I. Choi, H. Soh, "KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding", arXiv:2004.03289, 2020

[4] "Ko-Sentence-BERT-SKTBERT", Github, https://github.com/BM-K/KoSentenceBERT_SKT, 2020

[5] TripAdvisor, https://www.tripadvisor.co.kr/

[6] "KSS : Korean Sentence Splitter", Github, https://github.com/likejazz/korean-sentence-splitter, 2020

[7] Scikit-learn, https://scikit-learn.org/stable/