문장 생성 모델 학습 및 관광지 리뷰 데이터를 활용한 관광지 분류 기법

문준형 [1], 조인휘 [2]
[1] 한양대학교 컴퓨터소프트웨어학과 석사생
[2] 한양대학교 컴퓨터소프트웨어학과 교수

sadalsuud287@hanyang.ac.kr, iwjoe@hanyang.ac.kr

Tourist Attraction Classification using Sentence Generation Model and Review Data

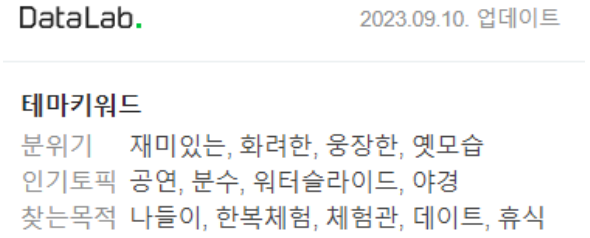
Jun-Hyeong Moon[1], In-Whee Joe[2]
[1]Dept. of Computer Science, Han-Yang University
[2]Dept. of Computer Science, Han-Yang University

요 약

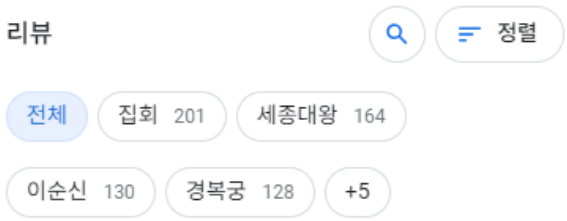
여러 분야에서 인공지능 모델을 활용한 추천 방법들이 많이 사용되고 있다. 본 논문에서는 관광지의 대중적이고 정확한 추천을 위해 GPT-3 와 같은 생성 모델로 생성한 가상의 리뷰 문장을 통해 KoBERT 모델을 학습했다. 생성한 데이터를 통한 KoBERT 의 학습 정확도는 0.98, 테스트 정확도는 0.81 이고 실제 관광지별 리뷰 데이터를 활용해 관광지를 분류했다.

1. 서론

코로나 시대가 끝나면서 많은 사람들이 다시 관광지를 찾기 시작했다. 대부분의 사람들은 관광지를 선택할 때 자신과 비슷한 성향의 관광지를 가고 싶어한다. 조용한 성격의 사람은 시끄러운 장소를 좋아하지 않고 시끌벅적한 분위기를 좋아하는 사람은 조용한 장소를 잘 찾지 않는다. 각각의 성향에 맞는 관광지를 추천해 주기 위해서는 관광지와 관광객 사이의 유사성을 찾아야 하는데 사람의 경우 스스로 자신의 성향을 선택할 수 있는 반면 관광지의 경우 제 3 자가 직접 관광지의 성향을 결정해 주어야 한다.



(그림 1) 네이버 지도에서의 광화문광장에 대한 키워드



(그림 2) 구글 지도에서의 광화문광장에 대한 키워드

하지만 사람마다 특정 관광지에서 느꼈던 감정이나 경험이 다를 수 있기 때문에 관광지를 분류하는 사람마다 해석이 달라질 수 있고 (그림 1), (그림 2) 와 같이 사이트마다 관광지에 대한 키워드가 달라서 키워드의 통합이 필요하다.

본 논문에서는 이를 해소하기 위해 GPT-3[1]를 기반으로 생성된 리뷰 데이터로 학습한 KoBERT 모델로 각 관광지의 실제 리뷰 데이터를 통해 관광지를 각 특성에 맞춰 분류하고자 한다.

2. 제안 방법

리뷰데이터를 활용하여 감성분석을 하는 시도는 특정 장소, 미디어 등 여러 분야에서 사용되어왔다.[2]

본 논문에서는 관광지를 대중적인 성격으로 분류하기 위해 네이버 블로그의 리뷰글을 크롤링하여 사용한다. 크롤링 된 리뷰데이터를 KoBERT 모델로 분류하기 위해서 분류하고싶은 키워드로 라벨링된 데이터셋이 필요하다.

최근 생성모델의 성능이 향상됨에 따라 여러 연구에서 목적성에 맞는 데이터를 구하기 힘든 경우 생성모델을 활용해 학습을 진행하는 경우[3]가 있다. 본 논문에서는 학습데이터를 구축하기 위해 자연어 생성모델인 GPT-3 를 활용하여 분류하고자 하는 키워드의 느낌이 담긴 가상의 리뷰문장을 생성했다. 생성된 가상의 리뷰문장으로 KoBERT 모델을 학습하고 크롤링된 리뷰데이터를 학습된 KoBERT 모델을 통해 리뷰문장들을 분류한다. 전체 과정은 다음과 같다.

1) 키워드 선정 및 가상의 리뷰 데이터 생성

'조용한', '시끌벅적한', '걷기 좋은' 등 관광지를 분류하고자 하는 키워드를 선정한 후 GPT-3 와 같은 생성 모델로 선정한 키워드의 분위기를 담고 있는 가상의 리뷰 문장을 생성한다. 생성된 문장들은 특정 키워드를 토대로 생성한 문장이기 때문에 각 키워드를 각 문장의 라벨로 사용할 수 있다.

2) KoBERT 분류 모델 학습

앞서 생성한 문장들로 이루어진 데이터셋을 사용해 KoBERT 분류 모델을 학습한다.

3) 관광지 리뷰 데이터 분류

학습된 KoBERT 분류 모델로 크롤링 한 관광지의 리뷰 데이터를 분류한다.

3. 실험

실험을 위해 키워드는 '조용한', '시끌벅적한', '걷기 좋은', '활동적인', '드라이브하기 좋은', '경치가 좋은', '저렴한', '책 읽기 좋은', '음식이 맛있는', '예술적인'등 100 가지를 선정하고 키워드의 분위기가 담긴 가상의 리뷰 문장을 GPT-3 모델을 사용해 아래 (그림 3)과 같이 생성했다.

이 장소는 조용한 안식처 같았어요.	조용한
주변이 조용해서 집중하기에 딱 좋았습니다.	조용한
북적이는 소음을 벗어나서 찾아온 곳이에요.	조용한
조용한 분위기 속에서 차 한 잔의 여유를 즐겼어요.	조용한
이 곳은 마치 소음의 오아시스 같았습니다.	조용한
친구들과 조용한 대화를 나눌 수 있는 곳이에요.	조용한
이곳에서는 정말로 마음을 찾을 수 있었습니다.	조용한
독서하기 딱 좋은 조용한 도서관 분위기였어요.	조용한
이곳은 시간을 천천히 보내기에 딱 좋은 공간이에요.	조용한
고요한 자연 속에서 마음을 정화했습니다.	조용한
이 카페는 조용한 산적 속에서 좋은 탈출구예요.	조용한
여기는 정말로 평화로운 분위기가 흐르고 있어요.	조용한

(그림 3) GPT-3 모델로 생성한 리뷰 문장

각각의 키워드마다 1000 문장씩 총 3000 문장을 생성했고 이를 활용해 KoBERT 로 문장 분류 모델을 학습했다. KoBERT 모델은 SKTBrain 의 사전학습된 KoBERT 를 fine tuning 하여 문장 분류 모델로 사용했다. 생성한 데이터로 학습한 모델의 학습 정확도는 0.98, 테스트 정확도는 0.81 이다.

실제 테스트를 위해 '고려대아이스링크', '남산공원', '어린이대공원' 등 894 가지의 장소를 선정했고 각각의 리뷰정보를 네이버 지도 및 구글 지도에서 크롤링 (그림 4) 하여 사용했다.

0	목동실내아이스링크	["'가격이 합리적이에요"', '"안전하게 관리해요"', '"규모가 커요"', '"...
1	인서울27골프클럽	["'친절해요"', '"캐디의 진행이 매끄러워요"', '"필드 상태가 좋아요"', ...
2	태릉CC	["'친절해요"', '"가격이 합리적이에요"', '"필드 상태가 좋아요"', '"뷰...
3	고려대아이스링크	["'친절해요"', '"안전하게 관리해요"', '"주차하기 편해요"', '"편의시설...
4	어린이회관눈썰매장	["'즐길 거리가 많아요"', '"주차하기 편해요"', '"친절해요"', '"가격이...
...	...	...
889	현대백화점 미아점	["'친절해요"', '"시설이 깔끔해요"', '"품질이 좋아요"', '"주차하기 편...
890	현대백화점 압구정본점	["'친절해요"', '"종류가 다양해요"', '"시설이 깔끔해요"', '"품질이 좋...
891	현대백화점 유플렉스 신촌점	["'종류가 다양해요"', '"친절해요"', '"시설이 깔끔해요"', '"품질이 좋...
892	현대백화점 천호점	["'친절해요"', '"종류가 다양해요"', '"품질이 좋아요"', '"시설이 깔끔...
893	화랑대철도공원	["'볼거리가 많아요"', '"산책로가 잘 되어있어요"', '"뷰가 좋아요"', '...

894 rows × 2 columns

(그림 4) 네이버 지도 및 구글지도에서 크롤링한 리뷰정보

각각의 리뷰정보를 크롤링 하여 문장단위로 분리했고 문장별로 분류 모델을 통해 문장 분류를 시행했다. 가장 많이 분류된 키워드 2 개를 선택하여 나열한 결과 아래(그림 5) 와 같은 결과가 나왔다.

0	목동실내아이스링크	[웨이팅이 적은, 연인이랑]
1	인서울27골프클럽	[많이 걷는, 차타고가기 좋은]
2	태릉CC	[웨이팅이 적은, 촌캉스]
3	고려대아이스링크	[술 마시기 좋은, 농촌여행]
4	어린이회관눈썰매장	[걷기 좋은/산책하기 좋은, 많이 걷지 않는]
...	...	...
889	현대백화점 미아점	[많이 걷지 않는, 촌캉스]
890	현대백화점 압구정본점	[제로웨이스트, 소리 지르기 좋은]
891	현대백화점 유플렉스 신촌점	[자전거 타기 좋은, 웨이팅이 적은]
892	현대백화점 천호점	[차타고가기 좋은, 사진 찍기 좋은]
893	화랑대철도공원	[반려동물과 함께하는, 인적이 드문]

894 rows × 2 columns

(그림 5) 관광지 분류 결과

4. 결론

추천 모델이 가장 활발히 사용되는 분야인 음악이나 동영상의 경우 '발라드', '트로트', '드라마', '공포'등과같이 각각 콘텐츠별 카테고리가 정해져 있다. 사람에 따라 발라드를 트로트라고 이야기하는 경우는 잘 없지만 관광지의 경우 같은 장소임에도 각각 겪었던 기억, 각 사람들의 환경 등에 따라 좋은 기억이 있는 장소일 수도 있고 안 좋은 기억이 있는 장소일 수도 있다. 따라서 관광지를 추천하기 위해서는 객관적이고 대중적인 라벨링이 필요했고 이를 위해 여러 사람

들의 리뷰 데이터를 사용하였다. 그로 인해 한 사람이 아닌 여러 사람의 의견이 반영된 분류 모델을 만들 수 있었다.

현재는 문장 분류를 위해 가상의 생성 데이터를 사용한 지도 학습 모델로 리뷰 문장을 분류했지만 향후에는 지도학습이 아닌 비지도 학습 기반의 Sent2Vec, FastText 등의 임베딩 방법을 활용하여 리뷰 데이터를 직접 분류하지 않고 관광지들을 더욱 정확하게 분류하고 추천할 수 있는 방법에 대해 연구할 계획이다.

본 논문은 2023 년도 한국콘텐츠진흥원의 지원을 받아 수행된 연구임 [R2022020116, AI 기반 관광객 상황 인식 및 관광정보 큐레이션을 통한 맞춤형 관광 여정 (Itinerary) 추천 플랫폼 기술개발]

참고문헌

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jered Kaplan “Language Models are Few-Shot Learners” Advances in Neural Information Processing Systems 33 (2020): 1877-1901.

[2] 임영희, 김홍범 “호텔 온라인 리뷰 빅데이터를 활용한 감성분석에 관한 연구” 호텔경영학연구 28.7 (2019): 105-123.

[3] Varun Kumar, Ashutosh Choudhary, Eunah Cho “Data Augmentation Using Pre-trained Transformer Models” arXiv preprint arXiv:2003.02245 (2020).