리뷰 데이터 기반의 사용자 친화적인 호텔 추천 챗봇 시스템

박정현, 이경민
한성대학교
{shshjhjh4455, gyeongmin}@hansung.ac.kr

User-friendly Hotel Recommendation Chatbot System based on Review Data

Joenghyeon Park, Gyeongmin Lee
Hansung University

요 약

본 연구는 최근 변화하고 있는 숙박 서비스 산업 내에서 AI 기술을 활용한 맞춤형 호텔 추천 시스템을 제안한다. 본 시스템은 Chat-GPT API를 사용해 사용자의 상세 요구 사항과 취향을 파악하고, KoBERT와 SBERT 모델을 사용하여 리뷰 데이터 기반의 호텔 추천 시스템을 구성하였다. 연구의 주요 목적은 사용자와의 유연한 상호작용을 통해 개인화된 호텔 추천을 제공하는 것이며, 이를 위해 자연스러운 대화형 인터페이스를 제공하는 호텔 추천 챗봇 시스템을 구축한다. 감성 분석과 문장 임베딩 기법을 적용하여 사용자의 리뷰에 대한 감성적 이해와 사용자의 요구 사항과 리뷰 간의 의미적 유사성을 분석하고, 호텔에 대한 종합적인 추천 스코어를 계산하여 사용자에게 최적의 호텔을 추천한다.

1. 서 론

최근 몇 년 동안 숙박 서비스 산업은 크게 변화하고 있다. 새로운 투자와 기존 시설의 업그레이드를 통해 여가 시설을 제공하고 혁신적인 기술 도입 등이 이루어지고 있다. 특히, 지역 숙박의 등장은 전통적인 관광 숙박에 대한 보완적 대안으로 작용하며, 이러한 변화들은 비즈니스 모델의 가치를 높이고, 소비자 수요에 맞춘 경쟁력 있는 제안을 만들어내고 있다 [1].

특히 인공지능(AI) 기술의 발전이 사용자 맞춤형 서비스와 추천 시스템 분야에서 혁신을 주도하고 있다. 추천 시스템은 정보 과부하와 개인화된 요구사항을 해결할 수 있는 큰 잠재력을 가지고 있다 [2]. 숙박 서비스 산업에서도 개인화된 서비스 제공의 중요성이 강조되고 있으며, AI 기반의 추천 시스템이 고객 만족도를 높이는 핵심 요소로 자리매김하고 있다. 이에 본 연구는 사용자와의 유연한 상호 작용이 가능한 호텔 추천 챗봇 시스템을 제안한다.

본 시스템은 Chat-GPT API를 활용하여 자연스러운 대화형 인터페이스를 제공하고 있으며, 대규모 호텔 리뷰 데이터셋을 활용하여 사용자의 상세 요구 사항과 취향을 분석한다. KoBERT (Korean BERT)와 SBERT (Sentence-BERT)를 사용하여 사용자의 세부적인 요구 사항을 파악하고 맞춤형 호텔 추천을 제공하고 있다. 이 접근법은 사용자에게 수많은 호텔 옵션 중에서 직접 선택할 필요 없이, 사용자의 취향과 요구 사항에 기반해 가장 적절한 호텔을 추천함으로써 사용자 경험을 극대화할 수 있다.

2. 관련 연구

SBERT는 전통적인 BERT 모델이 갖는 여러 한계를 극복하고자 개발된 모델로, BERT의 구조를 기반으로 하여 문장 또는 단락 단위의 임베딩을 최적화하여 생성하는 모델이다. 이 모델은 표준 BERT에 비해 더 높은 효율성과 정확도를 제공한다. 특히, 의미적으로 유사한 문장들 사이의 미묘한 차이를 정밀하게 포착하는 능력이 뛰어나다 [3].

BERT는 단어의 문맥을 파악하는 데 강점을 가지지만 전체 문장이나 단락의 의미를 이해하는 데는 한계가 있다 [4]. 이에 반해 SBERT는 문장 전체의 의미를 임베딩하는 데 특화되어 있으며, 이는 호텔 리뷰와 같은 사용자 생성 텍스트에서 더욱 정밀한 의미 분석을 가능하게 한다. 또한, SBERT는 처리 속도와 효율성 면에서도 기존 BERT 모델에 비해 우수한 성능을 보인다. 이는 대량의 리뷰 데이터를 신속하고 정확하게 처리할 수 있음을 의미한다.

SKT Brain에서 개발한 KoBERT 모델 [5]은 한국어의 복잡한 어휘 구조와 문법적 특성을 효과적으로 처리할 수 있는 능력을 가지고 있어 한국어의 자유로운 어순과 다양한 조사 및 어미 사용과 같은 언어적 특성을 고려할 때, 영어 중심의 언어 모델로는 도달하기 어려운 정확한 의미 파악과 감정 분석이 가능하다.

3. 연구 방법

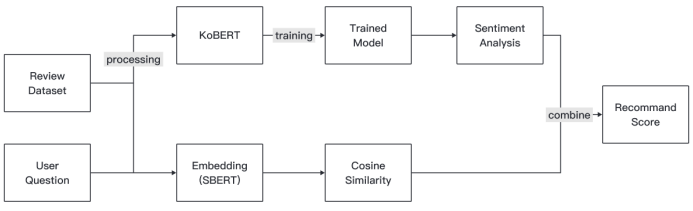

그림 1: 호텔 추천 모델 구조도

3.1 호텔과 리뷰 데이터셋

본 연구에서는 호텔 관련 데이터와 호텔별 리뷰 데이터를 중심으로 데이터셋을 구축하였다. 데이터를 수집하기 위해 인기 숙박 업체인 '야놀자' 홈페이지에서 서울 지역의 모든 호텔 정보를 파이썬의 selenium 라이브러리를 활용하여 크롤링하였으며, HTML 파싱을 통해 필요한 정보만을 추출하였다.

연구 데이터의 효율적인 관리와 분석을 위해, 테이블 스키마를 설계하고 MySQL 데이터베이스를 구축하여 수집한 정보를 저장하였다. 데이터베이스는 두 개의 테이블로 구성되어 있다. 먼저 호텔 테이블은 호텔의 ID, 이름, 위치, 평균 별점 등 각 호텔의 기본적인 정보를 담고 있다. 다음으로 리뷰 테이블은 리뷰의 ID, 호텔의 ID, 별점, 리뷰 텍스트로 구성되어 있다.

3.2 KoBERT를 활용한 감성분석

한국어로 작성된 호텔 리뷰 데이터에 대한 감성 분석을 위해, 한국어 텍스트 처리에 최적화된 BERT 기반의 KoBERT 모델을 파인튜닝하여 사용하였다. 파인튜닝은 모델이 특정 데이터셋, 즉 호텔 리뷰 데이터에 더 잘 적응하고, 감성 분석에서 더욱 높은 정확도를 달성할 수 있도록 조정하는 과정이다. 이를 통해 모델은 호텔 리뷰 데이터의 특성과 감성적 뉘앙스를 더욱 잘 이해하고, 정교한 분류를 수행할 수 있다. 평점이 4 이상인 리뷰는 긍정적 감성을 나타낸다고 가정하고, 그 이하인 경우에는 부정적 감성으로 분류하였다. 이렇게 레이블링된 데이터는 KoBERT 모델에 입력되며, 모델은 이를 바탕으로 각 리뷰의 감성 레이블을 예측하는 방법을 학습한다.

$$\text{sentiment}_i = \text{KoBERT}(R_i) \tag{1}$$

파인튜닝된 KoBERT 모델을 이용해 리뷰 데이터의 감성 분석을 진행하였다. 수식 (1)에서 sentiment_i는 호텔 리뷰 R_i가 긍정적인 리뷰일 확률을 나타낸다.

3.3 SBERT를 활용한 임베딩 벡터 생성

문장 임베딩은 의미적 유사성과 같은 복잡한 NLP (Natural Language Processing)작업을 훨씬 간단한 벡터 비교 작업으로 변환할 수 있고, 의미론적 유사성 비교 외에도 의미론적 검색을 통한 클러스터링이 가능하다 [6].

$$\mathbf{v}_i = \text{SBERT}(R_i) \tag{2}$$

$$\mathbf{v}_s = \text{SBERT}(s) \tag{3}$$

본 연구에서는 SBERT를 사용하여 수집된 호텔 리뷰 데이터와 사용자의 요구 사항을 임베딩 벡터로 변환하였다. 수식 (2)과 (3)에서는 호텔 H의 리뷰 R_i에 대해, SBERT를 사용하여 각 리뷰를 임베딩 벡터 $\mathbf{v}_i$로 변환하고, 사용자가 원하는 호텔에 대해 챗봇에게 전달한 문장 s를 임베딩 벡터 $\mathbf{v}_s$로 변환한다.

3.4 임베딩 벡터 간의 유사도 계산

두 문서 간의 유사성은 두 문서의 용어 벡터 사이의 코사인 값을 계산함으로써 도출할 수 있다 [7]. 코사인 유사도는 벡터 간의 각도를 기반으로 한 유사도 측정 방법으로, 벡터의 크기에 영향을 받지 않고 방향성에 초점을 맞추어 의미적 유사성을 측정한다. 이 방법은 길이가 다양한 텍스트 간의 비교에 효과적이며, 텍스트 데이터의 의미적 유사성을 정확하게 평가할 수 있다 [8]. 두 문장의 코사인 유사도가 1에 가까울 수록 두 문장은 의미론적으로 유사하고, -1에 가까울수록 정반대를 의미한다.

$$\text{similarity}(\mathbf{v}_i, \mathbf{v}_s) = \frac{\mathbf{v}_i \cdot \mathbf{v}_s}{\|\mathbf{v}_i\| \|\mathbf{v}_s\|} \tag{4}$$

수식 (4)는 각 호텔 H에 대해, 모든 리뷰 임베딩 $\mathbf{v}_i$와 입력 임베딩 $\mathbf{v}_s$ 간의 코사인 유사도를 계산한다.

3.5 추천 스코어를 통한 호텔 추천

3.2절과 3.4절에서 구한 각 리뷰의 감성 분석 결과와 사용자가 입력한 문장과 리뷰 문장 간의 유사도를 결합하여 사용자의 요구와 가장 유사한 호텔을 추천한다.

$$\text{score}_H = \frac{1}{N} \sum_{i=1}^{N} (\text{sentiment}_i \times \text{similarity}(\mathbf{v}_i, \mathbf{v}_s)) \tag{5}$$

수식 (5)에서 N은 호텔 H에 대한 리뷰의 개수를 나타낸다. 각 리뷰가 긍정일 확률과 사용자가 입력한 문장이 리뷰와 유사한 정도를 곱하고, 그 값의 평균을 호텔별 추천 스코어로 사용한다. 이렇게 구한 추천 스코어 값이 가장 높은 호텔을 사용자에게 추천한다.

3.6 챗봇 시스템 구축

그림 2는 챗봇의 실제 사용 예시를 나타낸다. 챗봇 시스템은 GPT 4 API를 사용하고, 프롬프트 엔지니어링을 통해 원하는 결과를 생성하였다. 또한 사용자가 편리하게 챗봇을 사용할 수 있도록 Flask를 사용해 웹앱을 구축하였다. 챗봇은 사용자와 대화를 통해 사용자의 호텔 취향을 찾게 된다. 사용자의 요구 사항이 명확하지 않은 경우, 챗봇은 추가적인 질의를 할 수 있다. 사용자의 요구 데이터를 취합한 후, 추천 시스템에게 전달하여 추천된 호텔을 사용자에게 전달한다.

4. 실험

4.1 두 호텔의 추천 스코어 비교

본 절에서는 전체적으로 호텔이 깨끗하고 친절하다는 평이 많은 호텔 A (그림 3 참조)와, 더럽고 냄새가 난다는 평이 많은 호텔

그림 2: 챗봇 시스템 사용 예시

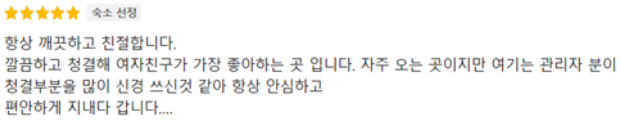

그림 3: 호텔 A 리뷰 예시

B (그림 4 참조)에 대한 추천 스코어를 비교한다. 사용자가 "깨끗하고 친절한 호텔을 원해" 라고 입력한 경우를 가정하고 실험을 진행하였다.

표 1: 두 호텔 간의 결과 비교

	호텔 A	**호텔 B**
리뷰 감성 분석 결과의 평균	0.9058	0.5189
입력 문장과 리뷰 문장 간의 유사도의 평균	0.3746	0.1020
추천 스코어	0.3394	0.0529

표 1에 따르면, 호텔 A와 호텔 B 리뷰에 대해 감성분석을 진행한 겨로가 호텔 A의 경우 0.9058로 0.5189가 나온 호텔 B보다 대체로 긍정적인 리뷰라는 것을 알 수 있다. 유사도의 평균값은 호텔 A가 0.3746으로 0.1010인 호텔 B에 비해 약 3.7배 높다는 것을 볼 수 있다. 이는 사용자가 원한 깨끗하고 친절한 호텔은 호텔 B보단 호텔 A에 가깝다는 것을 의미한다. 추천 스코어는 0.3394, 0.0529로 호텔 A가 호텔 B보다 약 6.4배 높다는 것을 확인할 수 있다. 이 결과는 우리가 설계한 추천 스코어가 유의미한 추천을 이끌어낼 수 있다는 것을 알려준다.

5. 결론

본 연구에서는 호텔 추천 스코어 계산을 통해 리뷰 데이터 기반의 호텔 추천이 가능한 모델을 설계하였다. 본 모델은 현재 SOTA

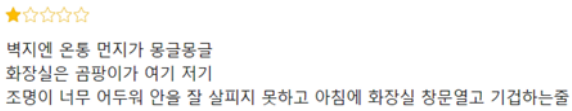

그림 4: 호텔 B 리뷰 예시

를 달성한 자연어 처리 모델을 활용하여 사용자 맞춤형 서비스를 제공하는 데 기여한다. 이 모델은 특히 숙박 서비스 산업에서 고객의 세부적인 요구 사항과 선호도를 분석하고, 이에 기반하여 최적의 호텔을 추천함으로써 사용자 경험을 향상시킨다. KoBERT와 SBERT를 사용한 감성 분석과 문장 임베딩은 리뷰 데이터를 통해 개인화된 호텔 추천이 가능해진다.

생성형 모델 기반의 챗봇 시스템과 BERT 모델 기반의 추천 알고리즘의 결합은 다양한 산업 분야에서의 응용을 위한 밑거름이 될 수 있다. 예를 들어, 유사한 방식으로 음식점, 여행지, 쇼핑 아이템 등을 추천하는 시스템 개발에도 활용될 수 있다.

하지만 이 연구의 한계점은 현재 사용자가 호텔 추천을 요청했을 때 완전한 답변을 받기까지 많은 시간이 소요된다는 것이다. 추후 연구에서는 계산 과정을 최적화하여 챗봇 속도를 높일 예정이다. 또한, 보다 더 정확한 추천 스코어 계산을 위해 다양한 방법을 연구할 계획이다.

참고 문헌

[1] J. Marques and R. P. Marques, "Trends in the hospitality industry: A global perspective," in *Digital Transformation of the Hotel Industry: Theories, Practices, and Global Challenges*, pp. 1–16, Springer, 2023.

[2] M. Aamir and M. Bhusry, "Recommendation system: state of the art approach," *International Journal of Computer Applications*, vol. 120, no. 12, 2015.

[3] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] S. Telecom, "Kobert: Skt open-source." https://sktelecom.github.io/en/project/kobert, 2021.

[6] S. Zoupanos, S. Kolovos, A. Kanavos, O. Papadimitriou, and M. Maragoudakis, "Efficient comparison of sentence embeddings," in *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, pp. 1–6, 2022.

[7] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[8] A. Singhal *et al.*, “Modern information retrieval: A brief overview,” *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.