Tourism Information QA Datasets for Smart Tourism Chatbot

Tae-Seung Ko[1] Myeong-Cheol Jwa[2] and Jeong-Woo Jwa[3*]

[1,2,3]*Jeju National University, Jeju, Korea.*
E-mail: lcr02@jejunu.ac.kr

Abstracts: Smart tourism uses artificial intelligence (AI) technology to provide easy and convenient travel services to tourists. The task-oriented chatbot system is a way to provide tourists more efficiently with travel services that were previously provided on the web or apps. In this paper, we develop the question answering (QA) dataset for an AI-based tourism information chatbot system. The tourism information QA dataset is developed in JSON format of KLUE MRC based on the tourism information database and tourism knowledge base built for smart tourism apps and rule-based chatbot services, respectively. To apply the QA model along with the DST and NER models to the smart tourism chatbot system, we develop the QA dataset by considering the previously developed the tourism information NER dataset and the smart tourism DST dataset. We evaluate the tourism information QA datasets with the koBigBird model, which can handle sequences of 4,096 tokens, and the EM (Exact Match) and F1 score are 96.85 and 98.84, respectively.

Keywords: Tourism Information QA, Machine Reading Comprehension, Smart Tourism Chatbot, Pre-trained language model, Mobile App.

1. INTRODUCTION

Smart tourism provides personalized travel services to tourists using the Internet of Things (IoT), communication infrastructure, big data, artificial intelligence (AI), AR/VR/MR/Metabus. Smart tourism services can be composed of travel planner and tour guide services. The travel planner service provides tourist information, recommended travel products, etc. so that tourists can create a personalized itinerary before travel. The tour guide service provides tourists with easy and convenient travel services at the destination according to their itinerary prepared before the trip. After the trip, the tourist can register the itinerary modified during the trip as a recommended travel product on the smart tourism platform. The smart tourism platform enables easy and convenient travel by providing AI-based travel planner and tour guide services to tourists.

We developed a smart tourism Android and React apps (Hyun-Ji Cho, 2022) that provide tourism information to tourists. The smart tourism app provides a travel planner service that allows tourists to create personalized travel products using recommended travel products before travel. We developed a voice guide service using a TTS server based on the location of the tourist to provide a smart tour guide service according to the personalized travel product for the tourist on the trip (KiBeom Kang, 2017).

The smart tourism chatbot service can easily and conveniently provide travel services to tourists using a task-oriented AI-based chatbot system. The task-oriented AI-based chatbot system uses QA algorithms as well as dialogue state tracking (DST) and named entity recognition (NER) algorithms. We developed the tourism information DST and NER datasets (Myeong-Cheol Jwa, 2022; Myeong-Cheol Jwa & Jeong-Woo Jwa, 2022) using pre-learning language models (PLMs) to develop the AI-based tourism information chatbot system. The tourism information DST algorithm defines domain, slot, and value to understand the intention of the tourist's question. The tourism information DST algorithm can improve the performance of the previously developed rule-based chatbot system (Dong-Hyun Kim, 2021). The rule-based chatbot system uses the Khaiii morpheme analyzer (Kakao khaiii) and the predefined rules to understand the tourist's question. The tourism information knowledge base was developed using Neo4J graph database to provide tourism information chatbot service. The NER algorithm is used to accurately distinguish proper nouns such as tourist destination names and administrative district names in tourist questions. The tourism information chatbot system can accurately understand user questions by combining DST and NER algorithms. In this paper, we develop the tourism information QA dataset to provide tourism information provided by smart tourism apps as a smart tourism chatbot service. The proposed tourism information QA dataset is performed for questions that cannot identify the intention of the tourist question with the domain, slot, and value of the DST algorithm. We perform training on the developed the Korean tourism information QA dataset using the KoBigBird model (Manzil Zaheer, 2020).

2. RELATED LITERATURE

QA models are deep learning models that can retrieve answers to user questions from a given context or knowledgebase. This is useful when searching for answers to user questions in the documentation. In this paper, we develop the tourism information QA datasets to provide tourism information to users through a smart tourism chatbot service. The tourism information QA datasets are developed using web surfing data based on the tourism information database of the smart tourism information system and the tourism information knowledge base of the rule-based chatbot system.

2.1. Textual Question Answering (QA) Datasets

Various datasets for evaluating QA models have been proposed (Zhen Wang, 2022) and QA datasets can be classified into textual QA, image QA, and video QA. In the image QA and video QA datasets, questions and answers are usually text, and the contexts are images and video clips, respectively. The tourism information QA dataset is textual QA, and the dataset consists of questions, answers, and sentences containing answers. Datasets for evaluating QA models include SQuAD, TriviaQA, NewsQA, and WikiQA. The Stanford Question Answering Dataset (SQuAD) is a popular benchmark dataset for evaluating QA models created by crowd workers using Wikipedia articles (Pranav Rajpurkar, 2016; Yuanjun Li & Yuzhu Zhang). SQuAD 2.0 contains the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions. TriviaQA contains 650K question-answer-evidence triples and 95K question-answer pairs (Mandar Joshi, 2017). NewsQA consists of over 100,000 question-answer pairs from over 10,000 CNN news articles (Adam Trischler, 2017). WikiQA consists of 3,047 questions and 29,258 sentences sampled from Bing query logs (Yi Yang, 2015).

The Korean Question Answering Dataset (KorQuAD) is a Korean machine reading comprehension (MRC) dataset released by LG CNS (Seungyoung Lim, 2019; Kim, Youngmin, 2019; Wang, & Wu, 2022). The KorQuAD 2.1 dataset consists of 102,960 pairs for 47,957 Wikipedia articles. Training set consists of 83,486 pairs and test set consists of 10,165 pairs. KorQuAD 2.1 also contains tables and lists to understand the document structure through HTML tags. Korean language understanding evaluation (KLUE) is a benchmark dataset for Korean natural language processing (NLP) that can evaluate 8 tasks (Park, Sungjoon, 2021). The KLUE MRC dataset adds news_category and source items to the KorQuAD dataset. The sources of the KLUE MRC dataset are Wikipedia, Acrofan, and Korea Economic Daily. QA models are typically evaluated on metrics like Exact Match (EM) and F1 score.

2.2. Question Answering (QA) Models

QA models can be broadly divided into two types based on how answers are generated: (1) extractive QA model (2) generative QA model. Extractive QA model predicts the span within the context with a start and end position which indicates the answer to the question. In the SQuAD dataset used for extractive QA model evaluation, answer_start and answer text fields are provided in JSON format to find the position of the answer in context. Representative extractive QA models are Bidirectional Encoder Representations from Transformers (BERT) (Jacob Devlin, 2018) and Korean BERT (KoBERT) models. The BERT-based QA models can be trained through fine-tuning BERT with context up to 512 tokens. Various studies have been conducted to train datasets with longer contexts than 512 tokens. BigBird and KoBigBird QA models can train contexts up to 4,096 tokens by applying a sparse attention mechanism to the BERT model. In this paper, we verify the performance of the tourism information QA dataset through fine-tuning KoBigBird. Chat Generative Pre-trained Transformer (ChatGPT) is a representative generative QA model. GPT models use the decoder layer of Transformer (Vaswani A., 2017). ChatGPT uses supervised learning, Reinforcement Learning from Human Feedback (RLHF) (Yuntao Bai, 2022; Coetser, & Nisa, 2023) and Proximal Policy Optimization (PPO) algorithms (John Schulman, 2017).

3. TOURISM INFORMATION QA SERVICES

We have been developed the rule-based chatbot services as well as smart tourism apps, Instagram, and YouTube to efficiently provide tourist information that tourists want. Apps and Instagram are common ways to provide tourism information to tourists, and chatbot services are a way to provide them more conveniently. The

tourism information chatbot service provides tourism information QA service by using the tourism information database collected through web surfing as well as the tourism information provided by smart tourism apps.

3.1. Smart tourism service platform

The smart tourism service platform provides smart tourism apps, Instagram, YouTube, and chatbot services. Figure 1 shows the database for smart tourism service and chatbot service in the smart tourism service platform.

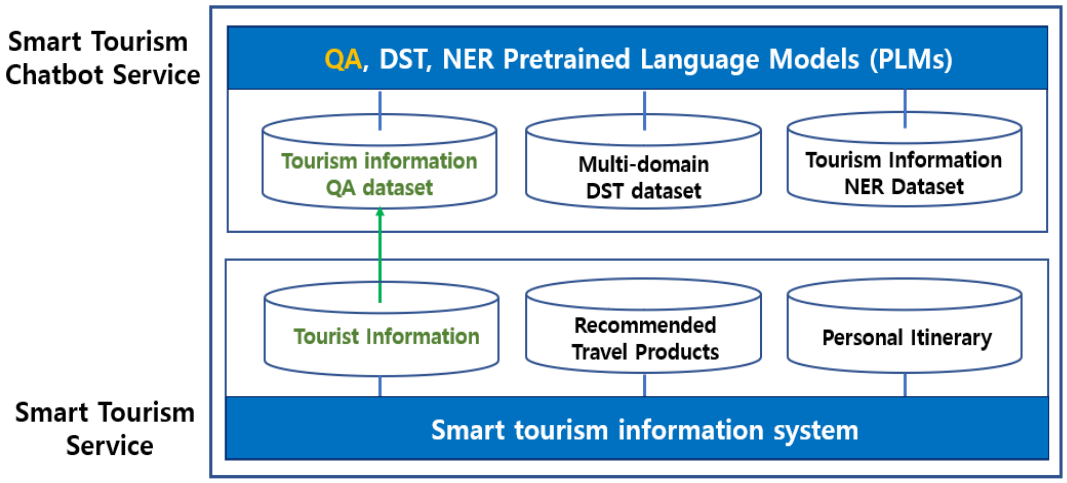

Figure 1. Smart tourism service platform.

The smart tourism service platform implements tourism information, recommended travel products, and my itinerary data for smart tourism services as a MySQL database. The smart tourism service platform includes datasets for developing QA, DST, and NER models for smart tourism chatbot services.

The tourism information database is used in the smart tourism app and includes the name of the tourist destination, classification code, address, phone number, homepage, operating hours, admission fee, as well as description and storytelling content. The recommended travel product database is for a travel planner service and enables a user to create a personalized travel itinerary using travel products recommended by travel experts. The personal itinerary is a personalized itinerary based on tourist information and recommended travel products and is used for tour guide services in travel. The tourism information DST dataset is for evaluating the DST model to understand the intent of user questions to advance the rule-based chatbot system into an AI-based chatbot system. The tourism information NER dataset is for the evaluation of the NER model to accurately understand the intention of the user's query along with the DST model. The smart tourism chatbot system requires a tourism knowledge base, which is also made based on a database of tourism information and recommended travel products.

3.2. Tourism information QA dataset

We are developing a task-based smart tourism chatbot system using the DST model and the NER model. The smart tourism chatbot system can understand the intention of the user's question with the DST model and provide answers to the user through search in the tourism knowledge base. The smart tourism chatbot system will have to use the QA model to provide answers when the DST model does not understand the intention of the user's question or when the answer to the user's question cannot be found in the tourism knowledge base. In this paper, we develop the tourism information question answering (QA) datasets to provide tourist information easily and conveniently as a smart tourism chatbot service in personalized travel planner and tour guide services.

We develop a tourism information QA dataset in the JSON format of the KLUE MRC dataset. Context in the QA dataset should be written to provide accurate answers to tourist questions. We create the context of the QA dataset using the contents acquired through web surfing based on the description of tourist destinations and storytelling

data in the tourism information database of the smart tourism service platform. If we use the content acquired through web surfing as context without refining it, we can provide wrong information to the user according to the original data. We develop a tourism information QA dataset for more than 1,000 tourist destinations in Jeju Island, Korea's representative tourist destination. The tourism information QA dataset consists of 1,025 contexts and more than 10,000 questions.

We are building recommended travel products for personalized travel planner service and tourism information data for smart tourism apps with MySQL database in the smart tourism information system. In addition, we are building a tourism knowledge base for rule-based smart tourism chatbot service with Neo4J graph database. The tourism information QA dataset is developed by adding tourism knowledge base of the smart tourism information system. We reflected the rules of tourism information defined in the rule-based chatbot system and the 5 domains and 22 slots defined when creating the smart tourism DST data set when creating 'question' of the tourism information QA dataset.

Figure 2 shows an example of a tourism information QA dataset for evaluating a tourism information QA model. The QA dataset contains answer_start to find the answer 9.6km in context for the question "How long is the Seongpanak trail?". The questions in the QA dataset include tourism information provided by the smart tourism app. The questions and answers of the currently developed QA dataset are configured to provide users with tourism information provided by Android and smart tourism apps as a QA model.


```
{
  "version": "smartTour-mrc-v1",
  "data": [
    {
     "title": "Introduction to Hallasan Seongpanak Trail",
     "paragraphs": [
       {
         "context": "Seongpanak Trail, the eastern course of Hallasan Mountain, is a trail
                     along with Gwaneumsa Trail that allows you to climb Baeknokdam,
                     the summit of Hallasan Mountain. The length of Seongpanak Trail is
                     the longest among Hallasan trails at 9.6km, and ....",
         "qas": [
           {
             "question": "How long is the Seongpanak trail?",
             "answers": [
               {
                 "text": "9.6km",
                 "answer_start": 87
               }
             ],
             "question_type": 1,
             "is_impossible": false,
             "guid": "klue-mrc-v1_train_20001"
           },
    ...
     ],
     "news_category": "Mt. Hallasan",
     "source": "Hallasan National Park"
    }
}
```


Figure 2. Tourism information QA dataset for smart tourism chatbot service.

The QA dataset can be used to evaluate BERT-like extractive QA models. The KLUE BERT base model, a pre-trained Korean BERT model, was developed to evaluate the Korean MRC dataset. The KLUE BERT base model can process 512 token sequences as input. QA models capable of handling input sequences larger than 512 tokens

have been proposed to process the context of long sentences in the extractive QA model. The KoBigBird model is a sparse-attention model that can handle a sequence of 4,096 tokens rather than BERT-like QA models. In this paper, we evaluate the performance of the proposed tourism information QA dataset using the KoBigBird model. As a result of performance evaluation of the KoBigBird model using the tourism information QA dataset, the EM and F1 scores were 96.85 and 98.84, respectively. We correct errors in the QA dataset while performing QA model evaluation.

4. CONCLUSIONS AND FURTHER STUDY

Smart tourism service can provide tourists with personalized travel itinerary and provide tour guide service according to the travel itinerary. The task-bed smart tourism chatbot service is an easy and convenient way to provide smart tourism services. Tourist information, recommended travel products, and tour guide services according to the itinerary can be provided by the smart tour chatbot system, but a QA system that finds answers in context is additionally required. In this paper, we develop a tourism information QA dataset and analyze the performance of the dataset using the KoBigBird QA model. In the KoBigBird model evaluation process, errors in the tourism information QA dataset are corrected to develop a dataset with excellent performance in EM and F1 score.

We plan to develop the smart tourism chatbot system that applies DST and NER models to a rule-based chatbot system. The developing smart tourism chatbot system understands the intention of the user's question with the DST and NER models and searches and provides answers from the tourism knowledge base. The smart tourism chatbot system can use the QA model to find answers in context and provide them to users for questions that cannot be answered in the tourism knowledge base. The smart tourism chatbot system using DST, NER, and QA models based on pre-trained language models provides smart tourism services to tourists.

5. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2021R1A2C1093283).

REFERENCES

[1] Hyun-Ji Cho, Jin-Yi Lee, Tae-Rang Park, & Jeong Woo Jwa, "React Native and Android Mobile Apps for Smart Tourism Information Service to FITs", The International Journal of Internet, Broadcasting and Communication, vol.14, no.2, pp.63-69, 2022.

[2] Coetser, L. C., & Nisa, N. (2023). Exploring the challenges faced by project managers regarding project losses caused by poor workmanship on construction sites in South Africa. Journal of Advances in Humanities and Social Sciences, 9(1)

[3] KiBeom Kang, JeongWoo Jwa, & SangDon Earl Park, “Smart Audio Tour Guide System using TTS”, International Journal of Applied Engineering Research, pp.9846-9852, 2017.

[4] Myeong-Cheol Jwa, Tae-Seung Ko, Byeong-Joo Kim, & Jeong-Woo Jwa, "Tourism Information Multi-domain Dialogue State Tracking Datasets for Smart Tourism Chatbot", International Journal of Intelligent Systems and Applications in Engineering, vol.10, no.1S(2022), pp. 192-196, 2022.

[5] Myeong-Cheol Jwa, & Jeong-Woo Jwa, “Development of Tourism Information Named Entity Recognition Datasets for the Fine-tune KoBERT-CRF Model”, International Journal of Internet, Broadcasting and Communication, Vol.14, No.2, pp 55-62, 2022.

[6] Dong-Hyun Kim, Hyeon-Su Im, Jong-Heon Hyeon, & Jeong-Woo Jwa, “Development of the Rule-based Smart Tourism Chatbot using Neo4J graph database”, International Journal of Internet, Broadcasting and Communication, Vol.13, No.2, pp 179-186, 2021.

[7] Kakao khaiii(Kakao Hangul Analyzer III),
https://tech.kakao.com/2018/12/13/khaiii/

[8] Neo4j graph database, https://neo4j.com/

[9] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, & Amr Ahmed, "Big Bird: Transformers for Longer Sequences", 34th Conference on Neural Information Processing Systems (NeurIPS 2020), pp. 1-42, 2020.

[10] https://github.com/monologg/KoBigBird

[11] Zhen Wang, "Modern Question Answering Datasets and Benchmarks: A Survey", https://doi.org/10.48550/arXiv.2206.15030, 2022.

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, & Percy Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text", Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392, 2016.

[13] Yuanjun Li, & Yuzhu Zhang, "Question Answering on SQuAD 2.0 Dataset", Stanford University.

[14] Mandar Joshi, Eunsol Choi, Daniel S. Weld, & Luke Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension", Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1601–1611, 2017.

[15] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, & Kaheer Suleman, "NewsQA: A Machine

Comprehension Dataset", Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 191–200, 2017.

[16] Yi Yang, Wen-tau Yih, & Christopher Meek, "WikiQA: A Challenge Dataset for Open-Domain Question Answering", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2013–2018, 2015.

[17] Seungyoung Lim, Myungji Kim, & Jooyoul Lee, "KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension", https://doi.org/10.48550/arXiv.1909.07005, 2019.

[18] Kim, Youngmin, Lim, Seungyoung, Lee, Hyunjeong, Park, Soyoon, & Kim, Myungji, "KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension", Annual Conference on Human and Language Technology, pp. 97-102, 2019.

[19] Park, Sungjoon, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, & Chisung Song et al., "KLUE: Korean Language Understanding Evaluation.", arXiv, 2021.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint arXiv:2005.08314, 2018.

[21] SKT KoBERT, https://github.com/SKTBrain/KoBERT

[22] ChatGPT: Optimizing Language Models for Dialogue, https://openai.com/blog/chatgpt/

[23] Vaswani A., Shazeer N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & and Polosukhin,I., "Attention is all you need", Advances in neural information processing systems, pp. 5998–6008, 2017.

[24] Yuntao Bai, et al, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback", arXiv:1909.07005, 2022.

[25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, & Oleg Klimov, "Proximal Policy Optimization Algorithms", arXiv:1707.06347, 2017.

https://huggingface.co/klue/bert-base

[26] Wang, C. H., & Wu, K. C. (2022). Interdisciplinary Collaborative Learning with Modular Programming and Information Visualization of Urban Smart Spaces. Journal of Advances in Technology and Engineering Research, 8(1)

DOI: https://doi.org/10.15379/ijmst.v10i1.1451