Hindi Tourism QA System: Low-Resource Question Answering using MT5-small

Asha Hegde[1], Sharal Coelho[1], Amrithkala M Shetty[2] and Mohammed Zaher Taljeh[1]

[1]*Department of Computer Science, Mangalore University, India*

[2]*Department of Computer Applications, Nitte Institute of Professional Education, Nitte (Deemed to be University), Karnataka, India*

Abstract

In natural language processing, Question Answering (QA) systems are built to reply automatically to user queries based on a database or a collection of documents, giving correct answers in natural language [1]. Open-domain QA systems such as Siri and Cortana are popular but fail when it comes to solving domain-specific requirements like tourism. As the tourism business is developing rapidly and many online travel websites are emerging, specialized QA systems are in greater demand. A tourism oriented QA system can retrieve and process a large amount of travel information, provide the right and timely answers to enrich the user experience and improve business competitiveness in the travel industry [2]. To encourage QA system in tourism domain specifically in Indian language like Hindi, "VATIKA: Varanasi Tourism in Question Answer System Indian language" shared task is organized at FIRE 2025. To address the challenges provided by the shared task organizers, we have fine-tuned MT5-based QA system and the proposed model obtained BLEU score of 26.

Keywords

Question-Answering, Transfer learning, Text generation, Low-resource language

1. Introduction

In recent years, Natural Language Processing (NLP) has witnessed significant advancements in tasks such as machine translation [3], text classification [4] [5], and Question Answering (QA) systems. QA have demonstrated the ability to process natural language queries and return concise, relevant answers. A QA system is a form of information retrieval system [6]. The natural language problem is transformed into a query statement that can be recognized by a machine. The answer is then returned in the form of natural language by querying a knowledge graph [7]. The general question answering systems include three parts: question analysis, information retrieval, and answer generation.

With the rise of tourism in recent years, travel websites have become increasingly popular. However, finding the right information on these sites often requires users to navigate through details about destinations, regions, and specific travel needs [8]. The process of gathering travel information is often complex and time-intensive, requiring users to navigate multiple websites to find relevant details. Frequently, the results obtained are irrelevant or misaligned with their specific needs. Implementing a tourism-focused quality assurance system would significantly enhance user convenience by streamlining access to accurate and pertinent information.

In the past, travelers depended on manual inquiries at tourist sites, which frequently led to inconsistent or unreliable information. Implementing a structured quality assurance system would address these issues, offering a more efficient and dependable way to access critical travel details. The development of QA systems powered by NLP technology has significantly reduced the time spent gathering information at scenic spots [9]. By utilizing NLP, these systems can accurately interpret users' intentions and provide personalized information about destinations. Compared to traditional search engine-based methods, QA systems provide more accurate and straightforward responses, enhancing the efficiency of information retrieval. However, most current QA systems are designed for general use or focus heavily on English,

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

hegdekasha@gmail.com (A. Hegde); sharalmucs@gmail.com (S. Coelho); amrithkalas@gmail.com (A. M. Shetty); mhd.zaher.taljeh.88@gmail.com (M. Z. Taljeh)

often lacking support for Indian languages and tourism-specific needs. Despite the growing number of tourists, access to reliable, multilingual, and context-relevant information is still limited.

The objective of the VATIKA shared task is to encourage the development of robust, domain-specific QA models capable of delivering authentic information to tourists. Such systems aim to enhance the overall visitor experience, streamline access to services, and support sustainable tourism development.

The paper is organized as follows: Section 2 contains Related Work. Section 3 contains the Task description and dataset. While Section 4 describes the Methodology, Section 5 gives a description of the Experiments, Results, and Observations followed by Conclusion in Section 6.

2. Related work

Recently, the tourism QA systems use Knowledge Graphs (KGs) [10], deep learning models [11], and semantic techniques [12] to provide precise responses to user queries about destinations, recommendations, and corresponding information. A study by Do et al.[13] on a Vietnamese tourism QA system proposed the BERT+vnKG model, combining a fine-tuned multilingual BERT for QA and a Vietnamese tourism KG (vnKG) built from over 4600 entity relationships extracted via VnCoreNLP dependency parsing. Their dataset comprised 300 manually crafted QA pairs from Vietnamese travel websites, split into 240 for training and 60 for testing. Experiments showed BERT+vnKG outperforming LSTM (F1: 0.90, MCC: 0.59) and standalone BERT (F1: 0.43, MCC: 0.25) with an F1 score of 0.94, MCC of 0.63, and reduced processing time (3 hours vs. 4-18 hours for 300 pairs), achieving higher accuracy in entity-specific predictions due to the KG's search space optimization.

In contrast, Contractor et al. [9] worked on large-scale QA using tourism data introduced the Cluster-Select-Rerank (CSR QA) architecture, featuring Universal Sentence Encoder (USE) for clustering review sentences, Duet neural retrieval for candidate selection, and a Siamese network with bi-directional GRU and attention layers for reranking. The dataset included 47,124 real user QA pairs from travel forums across 50 cities, with 216,033 entity reviews. Results demonstrated CSR QA's superiority over baselines like BM25 (Accuracy@3: 6.72%) and ablations, achieving Accuracy@3 of 21.44%, Accuracy@5 of 28.20%.

Li et al.'s [6] framework for a Chinese tourism QA system integrated BERT (base model with L=12, H=768) for intention recognition and answer generation, HanLP for NER, and Neo4j with Cypher for KG querying. A research by Sui's [14] on a tourism KG-based QA system employed a Bert-BiLSTM-CRF model for Named Entity Recognition (NER), an improved Naive Bayes for query classification and template matching, and Cypher queries on a Neo4j KG. The NER dataset, sourced from tourism sites like Horse Beehive and Baidu Travel, totaled 32,786 entries. The system achieved state-of-the-art NER performance with precision of 90.23% and F1-score of 91.10% via 4-fold cross-validation, surpassing models like HMM and BiLSTM+CRF.

Earlier, Kongthon et al.[15] semantic-based QA system for Thai tourism mapped natural language to SPARQL queries using pattern analysis on the OnTour Ontology, with an inference engine for recommendations and lexicons for domain-specific terms. Datasets contains crawled tourism data from Thai websites and 300 accommodation requests from Pantip.com forums. Experiments yielded 89% recall and 95% precision in request identification, with SPARQL conversions for top patterns, though descriptive conditions posed challenges, suggesting future rule-based enhancements.

Collectively, these studies highlight the efficacy of hybrid KG-deep learning approaches in overcoming scalability, linguistic, and domain-specific challenges in tourism QA, with performance gains in accuracy and efficiency, informing the development of robust, multilingual systems for enhanced user experiences.

3. Task Description

VATIKA[1], a Hindi-language QA dataset specifically designed to support Machine Reading Comprehension (MRC) and QA applications in the tourism domain. Centered on the culturally rich city of Varanasi,

[1] https://sites.google.com/view/vatika-2025/

the dataset reflects realistic queries that travelers and pilgrims might ask regarding locations, logistics, services, and spiritual landmarks.

VATIKA is unique in that it spans 10 tourism-relevant domains, including Ganga Aarti, Cruise, Food Court, Public Toilet, Kund, Museum, General, Ashram, Temple and Travel. Each domain includes detailed paragraph-level Hindi contexts followed by multiple question-answer pairs, simulating real-world information-seeking behavior in natural language. The questions range from factual to navigational and experiential, enhancing coverage across diverse tourist concerns. The VATIKA dataset is written entirely in Hindi, using the Devanagari script, and serves as a valuable language resource for building and evaluating QA systems. It supports both open-domain and contextual MRC-style question answering.

3.1. Dataset

The dataset is provided in a structured JSON format, where information is grouped first by domain, then by the related context, and finally by individual question–answer pairs. Each pair contains a unique ID, a Hindi question, and its corresponding answer. The following Table 1 contains statistics of provided train and validation data.

Table 1
Statistics of the Dataset

Dataset	Contexts	QA Pairs
Train set	5,358	13,408
Validation set	1,158	2,963

4. Methodology

The primary goal of this work is to develop an effective Hindi Question Answering (QA) system targeted towards the tourism sector, in this case, framework of the proposed method is shown in Figure 1. The system will generate automatically answers to user queries based on contextual information present within the dataset. Hindi being a morphologically complex [16], the development of the system addresses the lack of tourism-based digital solutions for the native population. In addition, using mT5-small, a multilingual transformer model, this study attempts to strike a balance between accuracy and computational cost. The long-term vision is to offer a strong QA system that can be applied across different domains and local languages.

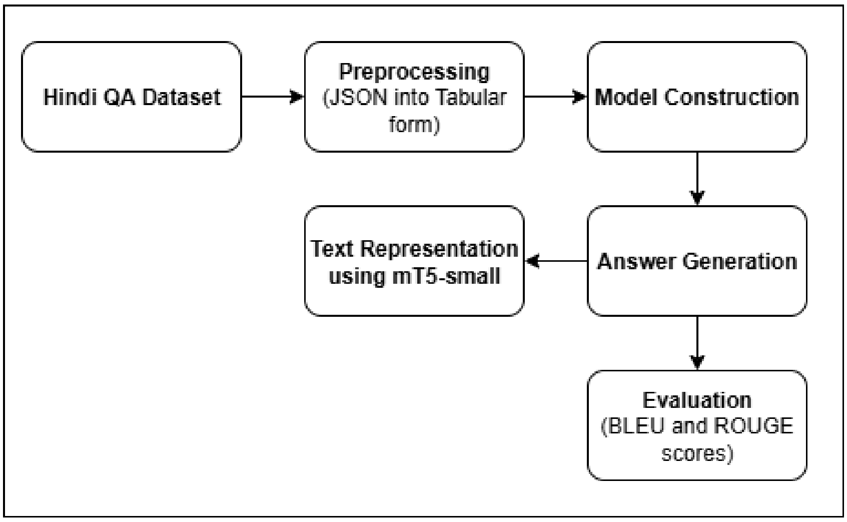
Hindi QA Dataset
Preprocessing (JSON into Tabular form)
Model Construction
Answer Generation
Text Representation using mT5-small
Evaluation (BLEU and ROUGE scores)

Figure 1: Framework of the proposed method

4.1. Preprocessing

The JSON dataset is initially converted into a structured tabular form with the question, context, and answer fields. Empty or missing answers are substituted with a particular placeholder phrase to deal with unanswerable questions. Text normalization is done by stripping leading and trailing spaces and making the dataset uniform. In order to accelerate training and minimize computational overhead, a train subset of 2000 samples and a validation subset of 500 samples is employed. The mT5 tokenizer is then used to perform tokenization such that input sequences (question + context) and target sequences (answers) are truncated or padded to fixed lengths.

4.2. Text representation

For text encoding, this research employs the mT5-small encoder-decoder architecture and tokenizer. For each question-background pair, these are combined into a single string in the format so that the model gets to see both the query and its respective background information. The tokenizer tokenizes these strings into subword tokens by making use of a SentencePiece model, which effectively deals with the intricacies of Hindi script. Padding is used to ensure consistent input lengths, and target answers are tokenized independently with padding tokens having their masking done during loss calculation. This allows the model to learn how to project an input sequence to its answer representation within the target space.

The mT5 (Multilingual T5) [17] model is a transformer sequence-to-sequence model trained on a huge multilingual corpus of over 100 languages. The "small" version is a light version of the model with fewer parameters for quick training within resource-restricted environments such as Kaggle notebooks. Although of small size, mT5-small has the fundamental strengths of T5, such as generalizability between languages through subword vocabulary sharing and transfer learning. Its encoder-decoder architecture allows efficient processing of tasks such as QA, translation, and summarization in Hindi and other low-resource languages.

The mT5-small is especially helpful since it is capable of extracting both syntactic and semantic structures in Hindi sentences and still be computationally lightweight. The model uses attention mechanisms to pay attention to the appropriate components of the input context when it generates answers. Being pre-trained from a massive multilingual dataset, it possesses an in-built capability to cater to Hindi-specific issues without needing huge domain-specific data. This turns mT5-small into a perfect fit for developing a scalable QA system within the tourism field.

4.3. Model Construction

The proposed QA system uses the Hugging Face Transformers library with Seq2SeqTrainer[2] for training and testing. The model accepts a tokenized question-context pair as input and produces the respective answer sequence as output. Training is conducted through cross-entropy loss with padding tokens within the target sequence being masked with -100 to prevent their contribution to the loss computation. For efficiency, the model is trained at a batch size of 2, gradient accumulation, and mixed-precision (FP16) GPU training. The training environment involves early model checkpointing, validation BLEU scoring, and best-model saving based on evaluation loss. At inference time, answers are produced by the model's generate() method with beam search disabled for quicker decoding. The predictions are incorporated into the test JSON format so that the system is prepared for actual tourism-based QA applications.

5. Experimental Results

To address the challenges of QA, we fine-tuned a multilingual sequence-to-sequence model based on mT5. The model is trained on the dataset provided by the shared task organizers [18]. The preprocessing

[2] https://huggingface.co/docs/transformers/main_classes/trainer

Table 2
Performance of the proposed model on Test Data-II

Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	QA-F1
Results	26	12.2	6	3.5	0.0128	0.0020	0.0128	0.2808

steps includes cleaning question–answer pairs, normalizing Hindi text, and tokenizing using the SentencePiece tokenizer from mT5. Hyperparameters such as learning rate, batch size, and training epochs are optimized empirically to prevent overfitting on the relatively small corpus.

The performance of proposed model on Test Data-II was evaluated using F1 score, BLEU (1–4) and ROUGE-L metrics. The detailed scores are presented in Table 2. The model obtained BLEU-1: 26, BLEU-2: 12.2, and QA-F1: 0.2808. As shown in result Table 2, while the model shows reasonable unigram precision (BLEU-1), the lower higher-order BLEU and ROUGE scores indicate challenges in capturing longer semantic dependencies and producing contextually coherent answers. These results shows that domain-adapted multilingual transformers can successfully comprehend Hindi questions, retrieve relevant information, and generate accurate answers. The relatively higher BLEU-1 and BLEU-2 scores suggest the model captures surface-level lexical overlap effectively, while the lower ROUGE and F1 values reveal opportunities for improving semantic accuracy and deeper answer reasoning.

6. Conclusion and Future Work

This study presented our system for the VATIKA: Varanasi Tourism in Question Answer System in Indian Language (Hindi) shared task at FIRE 2025. To address the inherent challenges of developing QA systems for low-resource languages, we fine-tuned a multilingual mT5-based model, achieving a BLEU score of 26 on the official evaluation set. These findings underscore the efficacy of using multilingual pre-trained language models to improve the quality of answer generation in domain-specific quality assurance tasks. Future research will focus on incorporating domain adaptation strategies, curating larger and more diverse datasets, and investigating hybrid retrieval-generation approaches to further improve robustness and generalizability.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat GPT-4 in order to:Grammar and spelling check. Paraphrasing was handled via QuillBot. With this tool, the author(s) reviewed and revised the content as required, while assuming full responsibility for the publication's integrity.

References

[1] E. Mutabazi, J. Ni, G. Tang, W. Cao, A review on medical textual question answering systems based on deep learning approaches, Applied Sciences 11 (2021) 5456.

[2] Y. Li, Enhancing the accuracy of question-answering systems using machine learning: A case study in the tourism domain, Applied and Computational Engineering 160 (2025) 113–119.

[3] A. Hegde, H. L. Shashirekha, Kt2: Kannada-tulu parallel corpus construction for neural machine translation, in: Proceedings of the 20th International Conference on Natural Language Processing (ICON), 2023, pp. 743–753.

[4] S. Coelho, A. Hegde, P. Lamani, H. L. Shashirekha, et al., Mucsd@ dravidianlangtech2023: Predicting sentiment in social media text using machine learning techniques, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2023, pp. 282–287.

[5] A. Hegde, S. Coelho, H. Shashirekha, Mucs@ dravidianlangtech@ acl2022: ensemble of logistic regression penalties to identify emotions in tamil text, in: Proceedings of the second workshop on speech and language technologies for Dravidian languages, 2022, pp. 145–150.

[6] J. Li, Z. Luo, H. Huang, Z. Ding, Towards knowledge-based tourism chinese question answering system, Mathematics 10 (2022) 664.

[7] H. Azarbonyad, Z. L. Zhu, G. Cheirmpos, Z. Afzal, V. Yadav, G. Tsatsaronis, Question-answer extraction from scientific articles using knowledge graphs and large language models, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 1120–1129.

[8] A. Guerrieri, G. Ghiani, A. Manni, A tourist advisor based on a question answering system, in: 2017 Intelligent Systems Conference (IntelliSys), IEEE, 2017, pp. 1173–1176.

[9] D. Contractor, K. Shah, A. Partap, P. Singla, et al., Large scale question answering using tourism data, arXiv preprint arXiv:1909.03527 (2019).

[10] S. Aghaei, E. Raad, A. Fensel, Question answering over knowledge graphs: A case study in tourism, IEEE Access 10 (2022) 69788–69801.

[11] D. I. Af'idah, S. F. H. Dairoh, Comparative analysis of deep learning models for retrieval-based tourism information chatbots, Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI) 11 (2025) 53–67.

[12] M. Rahim, Z. Turabee, Q. Rajput, S. A. Khoja, Semantic based question answering system on travel ontology, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2019, pp. 67–74.

[13] P. Do, T. H. Phan, B. B. Gupta, Developing a vietnamese tourism question answering system using knowledge graph and deep learning, Transactions on Asian and Low-Resource Language Information Processing 20 (2021) 1–18.

[14] Y. Sui, Question answering system based on tourism knowledge graph, in: Journal of physics: Conference series, volume 1883, IOP Publishing, 2021, p. 012064.

[15] A. Kongthon, S. Kongyoung, C. Haruechaiyasak, P. Palingoon, A semantic based question answering system for thailand tourism information, in: Proceedings of the KRAQ11 Workshop, 2011, pp. 38–42.

[16] D. K. Malladi, P. Mannem, Statistical morphological analyzer for hindi, in: Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013, pp. 1007–1011.

[17] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).

[18] P. Gatla, Anushka, N. Kanwar, G. Sahoo, R. K. Mundotiya, Tourism question answer system in indian language using domain-adapted foundation models, arXiv preprint (2025).