Indian Institution of Industrial Engineering

BOMBAY

CERTIFICATE OF PUBLICATION

This is to certify that the article entitled

QA IN INDIAN LANGUAGES: AN EXTENSIVE SURVEY OF CURRENT RESEARCH

Authored By

Nikitha Reddy Bitla,
Computer Science and Engineering, Maturi Venkata Subba Rao Engineering College, Hyderabad, India

Published in

Industrial Engineering Journal : ISSN **0970-2555**

Volume : 53, Issue 2, No. 2, February : 2024

UGC Care Approved, Group I, Peer Reviewed Journal

with IF=6.82

Editor in Chief

(Kuldip Singh Sangwan)

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

QA IN INDIAN LANGUAGES: AN EXTENSIVE SURVEY OF CURRENT RESEARCH

Venkataramana Battula, Aishwarya Chirravuri, Tansin Taj, Nikitha Reddy Bitla, Computer Science and Engineering, Maturi Venkata Subba Rao Engineering College, Hyderabad, India
venkataramana_cse@mvsrec.edu.in,

Abstract
Thoroughly reviewing the landscape of Question Answering (QA) in Indian languages, the survey meticulously examines the extensive body of work contributed by various researchers in the field. It underscores the significance of QA research in India, given the country's rich linguistic diversity, showcasing how QA technology has the potential to transform information access, education, healthcare, and cultural preservation within this diverse linguistic tapestry. The paper addresses the unique challenges researchers encounter, including data scarcity, linguistic diversity, code-switching, dialectal variations, and limited linguistic resources, serving as the backdrop for evaluating these research contributions. Additionally, the survey provides an in-depth analysis of the datasets tailored for QA in Indian languages, shedding light on their intricacies and diversity. Furthermore, it reviews the works of researchers, highlighting innovative strategies such as data augmentation, transfer learning, crowdsourcing, and the use of multilingual models, elucidating their specific applications and substantial contributions to the burgeoning field of QA in Indian languages. In conclusion, the paper emphasizes the profound impact of QA research in India, underlining its role in preserving linguistic heritage, enhancing healthcare access, improving education, and supporting linguistic studies. It serves as an indispensable resource for researchers, practitioners, and policymakers navigating the dynamic realm of QA in NLP within India's diverse linguistic and cultural milieu, highlighting the collective efforts to bridge linguistic gaps and promote digital inclusivity in this multifaceted nation
Keywords: Question-Answering, Extractive QA, ODQA, Low-resource Language QA, QA datasets

I. INTRODUCTION

Question Answering (QA) is a pivotal field within Natural Language Processing (NLP) that seeks to bridge the gap between human language and machine understanding. It has emerged as a cornerstone in the development of intelligent systems, with profound implications for various applications, from enhancing search engines and virtual assistants to revolutionizing education and healthcare. In this era of burgeoning information, where vast volumes of unstructured textual data abound, the ability to ask questions and receive meaningful responses is not merely a convenience but a necessity. This research paper delves into the realm of QA in NLP, exploring its significance, the myriad of challenges it presents, the key techniques that underpin its operation, and its diverse applications across industries. As QA continues to evolve, it promises to reshape the way we interact with information, and this paper endeavors to provide a comprehensive overview of its current state and future directions.

The importance of QA in NLP cannot be overstated. In a world where data has become the lifeblood of decision-making, QA serves as the conduit through which individuals, organizations, and even machines access, understand, and act upon this data. From the precision of search engine results to the convenience of voice-activated personal assistants, QA forms the backbone of these technologies. It offers transformative potential in healthcare, where QA systems can aid medical professionals in diagnosing and treating patients based on a wealth of medical knowledge. In education, it has the power to make learning more interactive and engaging, providing instant answers to student queries and enabling educators to create automated assessments. Furthermore, QA is indispensable in content generation, as it can generate coherent, informative articles, reports, and summaries on a multitude of topics. As such, this paper aims to shed light on the various domains where QA has a profound impact, reinforcing its significance as an ever-expanding field in NLP.

Different QA variants exist based on inputs and outputs. In extractive QA, models extract answers from a context, which could be provided in the form of text, tables, or HTML, often using BERT like models. Open-generative QA involves models generating free text directly based on the provided context, while closed-generative QA, without a specified context, entirely generates answers through model-generated content. Question-answering (QA) systems play a crucial role at the intersection of natural language processing (NLP) and information retrieval. The standard OpenQA architecture comprises three key steps: Question Analysis, Document Retrieval, and Answer Extraction. In Question Analysis, the system comprehends the user's natural language question, facilitating subsequent stages. Document Retrieval involves searching for relevant materials using self-built information retrieval (IR) systems or web search engines. Finally, Answer Extraction extracts the final responses from identified documents. Various QA techniques evolve, including IR-based Factoid QA, Knowledge-Based QA utilizing structured databases, and approaches that leverage multiple information sources to predict and rank candidate answers.

Open Domain Question Answering (ODQA) models, such as DrQA, Anserini, and Multi passage BERT QA, employ diverse strategies to retrieve relevant information from large knowledge bases like Wikipedia. These models often incorporate techniques like dense representations, dense parse phrase indexing, and retrieval-augmented language model pretraining. While each model has its strengths and intricacies, they collectively contribute to the evolving landscape of ODQA, addressing challenges like context understanding, passage retrieval, and efficient information extraction.

Closed Domain Question Answering (CDQA) systems, on the other hand, specialize in extracting answers from specific domains using predefined sets of documents or paragraphs. They analyse given paragraphs or documents to answer questions accurately within a confined subject area. The ongoing advancements in QA systems showcase their significance in various applications, from personal assistants and chatbots to search engines, e-invoicing, and digital billing. As these systems continue to mature, their capabilities expand, making them even more integral to human-computer interaction and information retrieval.

There are a few challenges in question-answering systems that will be discussed in this paragraph. The first challenge is ”ambiguity,” which means that the same word can have different meanings. For example, a bank may be a riverbank or a bank, which is a financial establishment. The question could only be answered if the exact meaning is identified according to the context. Successfully tackling this limitation or gap could increase the accuracy of the system capable of answering the question. The second one is the “lexical gap,” which refers to the same meaning that can be expressed in different ways. And the last one is 'Multilingualism', as the name suggests, many languages. So the language should also be taken into consideration before the process. If all these limitations are handled accordingly, then there is a freeway to an efficient and precise QA system. In recent years, Natural Language Processing (NLP) has made remarkable strides, reshaping the way we interact with information and technology. However, within the vast realm of linguistic diversity, there exists a segment often referred to as ”low-resource languages.” These languages, predominantly spoken by marginalized communities, pose unique challenges and opportunities in the context of question-answering (QA) in NLP.

Our main contribution exhibits research work conducted on the Question Answering Task in Indian languages. We also present the following for current research concerns.

- Challenges in Low-Resource Language QA
- Strategies and Techniques for Low-Resource Language QA
- Question Answering in NLP for Indian Languages
- QA Datasets for Indian languages

II. SURVEY OF QA IN INDIAN LANGUAGES

In our literature review, we explore Question-Answering (QA) systems, scrutinizing Google search limitations and unfolding diverse approaches to develop language-specific QA models for Hindi,

Bengali, Telugu, and Marathi. This emphasizes the imperative for advancing Indian language processing and fostering cross-linguistic comprehension. Sharma et al [1] addressed the limitations of Google search results. Their Question-Answering (QA) dialogue system aimed to quickly provide concise answers to user queries. The methods they used included partial semantic and syntactic analysis, along with a keyword-based approach. A key component of their system was the Dialogue Manager, which helped resolve communication challenges. The findings show that their QA dialogue system delivers specific and quick responses compared to the time-consuming clicking-based system of the Indian Railway website. These initial results encourage further research and potential extensions, such as providing real-time information like arrival and departure times via phone.

Poonam Gupta et al [2] present how advancements in technology make question-answering in natural language a crucial part of text mining research. This process simplifies user queries by providing concise answers, instead of making users navigate through extensive documents. However, there is a challenge due to the limited availability of language resources for Indian languages like Hindi, Punjabi, Bengali, Kannada, Telugu, and Marathi. This paper surveys various question answering techniques designed for these languages while considering India's unique linguistic diversity. Limited research and language resources exist for Indian languages, particularly in the domain of question-answering systems for Hindi, Punjabi, Telugu, and Bengali, with only nine papers identified. In contrast, there is more extensive research happening for other languages worldwide. This highlights the urgent need for increased investment and attention in Indian language processing to advance technology for these languages.

Poonam Gupta et al [3] explored the importance of QA systems for providing accurate responses. They stressed the value of QA systems in education, particularly during the Covid-19 pandemic when online learning became prevalent. They concentrated on a specific area and noticed a shortage of QA systems for Indian languages like Hindi, Punjabi Bengali, Malayalam, and especially Marathi.

Kumar [4] develops a Hindi QA System that allows users to ask questions in Hindi and receives answers from Hindi documents. This system utilizes a unique approach, including a Hindi search engine, to effectively retrieve context-based information, especially in fields such as agriculture and science. In their paper, they introduce a Hindi QA system designed for specific tasks within domains. This system makes use of NLP techniques and a Unicode-compliant search engine. It is automated and context-based, resulting in a high level of accuracy. Their future plans involve further enhancing the extraction of semantic information and expanding the system's capabilities to support multiple languages. This extension aims to benefit a broader audience, particularly non-English speakers and rural communities. Praveen Kumar [5] introduces the Hindi QA system, employing NLP techniques for agriculture and science queries. The system ensures Unicode compliance and boasts a 75% accuracy rate. Kumar and his team aim to enhance semantic extraction, expand capabilities to different domains and languages, potentially revolutionizing local language information access.

Arijit Das [6] develops a Bengali QA system using supervised learning techniques, achieving high accuracy. This system has practical applications in chatbots, virtual agents, and e-governance for Bengali-speaking populations and can potentially be adapted to other languages with minimal modifications. Arijit Das has created a Bengali Question Answering (QA) system using lightweight supervised learning methods. This system categorizes questions and retrieves answers from those categories. It offers a cost-effective solution suitable for personal computers, making it relevant during budget constraints, such as those imposed by the COVID-19 pandemic.

Maithili Sabane [7] focuses on creating QA datasets for Hindi and Marathi, both considered low-resource languages. This undertaking involves translating the SQuAD 2.0 dataset, resulting in the release of extensive QA datasets for each language, each comprising 28,000 samples. The paper also highlights the best-performing models in this context. The primary goal of this research is to provide valuable support for studies in Hindi and Marathi, while also fostering cross-linguistic comprehension. Furthermore, the datasets, models, and code generated through this initiative are made publicly accessible, facilitating further progress in conversational AI research, particularly in the domains of

Marathi and Hindi languages. The ultimate objective is to promote the development of enhanced models for Question-Answering tasks in these languages.
Bharat A. Shelker [8] focuses on the creation of QA datasets for Hindi and Marathi, which are low resource languages. The approach involves translating the SQuAD 2.0 dataset into these languages. This effort has resulted in the release of the largest QA datasets for Hindi and Marathi, each comprising 28,000 samples. Furthermore, the paper presents the best-performing models for these languages. The primary objective of this work is to bolster research in Hindi and Marathi and promote cross-linguistic understanding. Importantly, Bharat A. Shelker and the team plan to make the datasets, models, and code publicly available, facilitating further research in conversational AI, specifically within the contexts of Hindi and Marathi languages. The ultimate goal is to support the development of enhanced Question Answering models in these languages. Md. Aminul Islam [9] introduces a Bengali question classification system using the SGD classifier, with average precisions of 0.95562 for coarse questions and 0.87646 for finer ones. Pioneering Bengali question type classification for QA, it tackles low resource languages and aims to improve precision.
Jonathan H. Clark [10] introduces the TYDI QA dataset, which comprises 204K question-answer pairs spanning 11 linguistically diverse languages. The dataset's objective is to enable the development of models that can generalize across a wide array of global languages. Notably, this dataset includes both quantitative and qualitative linguistic analyses of language phenomena, collected directly in each language to avoid translation-induced biases. Rakesh Kumar [11] introduces TeQuAD, a Telugu QA Dataset generated from 82k parallel triples sourced from the Stanford QA Dataset. The authors present methods for crafting similar datasets tailored to low-resource languages and showcase their models' performance in both Monolingual and Cross-Lingual Machine Reading Comprehension scenarios, achieving an impressive F1 score of 83% and an EM score of 61%. This research contributes to the development of a quality assurance dataset for Indian languages, with a particular focus on building MRC (Machine Reading Comprehension) corpora through translation methods. It highlights the utilization of high-resource languages like English to create datasets for low-resource languages such as Telugu. The work also introduces techniques for dataset quality and quantity enhancement, along with mechanisms for data augmentation.
Gowtham Ramesh [12] introduces Samanantar, a vast parallel corpora collection for 11 Indic languages and English, comprising 49.7 million sentence pairs. Its NMT models outperform existing benchmarks, enhancing multilingual NLP for Indic languages. Anjani Garg [13] discusses the evolution of automatic QA systems, with a specific focus on generating questions from Punjabi language documents. The paper emphasizes the remarkable improvement in performance witnessed over the last decade in the realm of QA systems for Punjabi language content. Partha Pratim Manna [14] presents a Bengali Semantic QA system using Bengali WordNet, achieving an 80% overall accuracy on 40 diverse questions. Future improvements may consider handling negative sense words like"no" and" not" to avoid contradictions. The paper also discusses encountered challenges and pitfalls.
Suket Arora [15] discusses the role of Question Answering Based Dialogue Systems (QABDS) in Natural Language Processing. The paper compares QABDS systems, identifies research gaps, and proposes an NLP-based QABD system for Punjabi. This system assists rural Punjabi users in English communication and represents the first of its kind for the language. It analyzes queries, identifies keywords, generates SQL queries, and provides answers in Punjabi.
Seena [16] addresses factoid question answering in Malayalam using the TnT tagger, highlighting the challenge posed by the language's agglutinative nature and the increasing web content in Malayalam. The paper proposes anaphoric resolution for efficiency and mentions potential sentence representation in future research. It introduces HindiRC, a Hindi question answering dataset, and compares traditional similarity metrics. Kaveri Anuranjana [17] introduces HindiRC, the first grade-based Hindi reading comprehension dataset. It evaluates similarity metrics and acknowledges the research gap in Malayalam question answering, proposing anaphoric resolution. They suggest exploring specific

sentence representations, machine learning-based solutions, and automatic question generation using abundant Hindi news and Wikipedia articles.

Hrishikesh Terdalkar [18] presents research focusing on building knowledge graphs from Sanskrit texts and employing them for factoid question answering. They achieve a 50% accuracy rate in answering such questions related to human and synonymous relationships in texts like Mah¯abh¯arata and A¯yurveda. The paper also outlines plans for improving accuracy through better word analyzers, dictionaries, and expanding the framework to handle more complex questions in Sanskrit texts. Adhitya Thirumala [19] conducts a project to tackle the underrepresentation of natural language processing (NLP) in Hindi and Tamil. They aim to enhance extractive question-answering (QA) in these languages, critical for non-English speakers. Three models are employed, with the third, finetuned on the Indic dataset, performing the best. This model achieves a remarkable word-level Jaccard score of 0.958 in Hindi and 0.829 in Tamil, showcasing the potential to improve search engine QA methods for diverse languages. Bharat A. Shelke [20] emphasizes the significance of Marathi language QA systems in education during COVID-19. Their system achieves high precision and recall, making it suitable for educational content retrieval.

III. PROPOSED WORK

We explore challenges in answering questions in languages with limited resources, discussing solutions and methods. Our focus is on Indian languages, addressing challenges in NLP question answering. We delve into strategies and techniques to improve question-answering in these languages. Additionally, we examine available datasets specifically designed for question-answering in Indian languages, aiming to enhance research and development in this domain.

1) Challenges in Low-Resource Language Question Answering

QA in low-resource languages presents a formidable set of challenges. The scarcity of digital text data in these languages hinders the development of QA systems, making it difficult to train and fine-tune models effectively. Additionally, lowresource languages often exhibit intricate linguistic features, including complex morphology, dialectal variations, and a lack of linguistic resources, such as dictionaries and language models. Moreover, QA systems addressing these languages must demonstrate cultural sensitivity and context awareness to prevent misinterpretations and biases

2) Strategies and Techniques for Low-Resource Language QA

Despite these formidable challenges, there exist strategies and techniques that researchers and practitioners employ to confront QA issues in low resource languages. Data augmentation and synthetic data generation can be employed to create larger and more diverse datasets, overcoming the scarcity of data. Transfer learning from high resource languages or domains is highly effective, allowing pre-trained models to be fine-tuned using limited data from low-resource languages. Collaborating with speakers and local experts from these communities through crowdsourcing contributes significantly to dataset development. Furthermore, the deployment of multilingual models and domain-specific adaptation aids in addressing linguistic diversity.

3) Applications and Implication

Low-resource language QA holds significant applications across diverse domains, which underscores its importance. QA systems for low-resource languages play a pivotal role in preserving cultural knowledge and traditions, granting access to digital information in native languages, and facilitating communication between local governments and communities in regions where these languages predominate. Furthermore, the provision of healthcare information and guidance in native languages enhances healthcare outcomes in underserved regions. In education, low-resource language QA supports e-learning initiatives by offering educational content and interactive lessons in native languages. Additionally, QA systems assist linguists in the study of low-resource languages, facilitating the documentation of linguistic features and variations.

4) Question Answering in NLP for Indian Languages

Natural Language Processing (NLP) has witnessed significant progress in recent years, particularly in major languages like English. However, addressing linguistic diversity remains a challenge, especially in a linguistically rich and diverse country like India. This article explores the landscape of Question Answering (QA) in Indian languages, highlighting its significance, the unique challenges it poses, strategies for development, and its potential applications.

5) Challenges in QA for Indian Languages

Question Answering (QA) in Indian languages presents a complex set of challenges that must be navigated to ensure effective language understanding and response generation. One of the most significant obstacles is the scarcity of digital data available in many of these languages. Unlike major languages, Indian languages often lack extensive datasets, making it challenging to train QA models effectively. This data scarcity hampers the development of robust QA systems, as these models heavily rely on large and diverse datasets for training and fine-tuning. Another critical challenge arises from the diverse nature of Indian languages. India boasts a wide range of scripts and writing systems, including Devanagari, Tamil, Kannada, and many more. Each language possesses its unique intricacies, from grammar and vocabulary to script styles. Consequently, developing QA systems for Indian languages requires language-specific handling and processing to accommodate these idiosyncrasies effectively. Furthermore, the prevalence of code-switching and multilingualism in India adds an additional layer of complexity. Many individuals seamlessly switch between languages within a single sentence or conversation. This linguistic phenomenon poses a significant challenge to QA systems, demanding the ability to understand and respond accurately to code-switching, reflecting the linguistic diversity and dynamism of the region. Compounding these challenges is the limited availability of linguistic resources for Indian languages. Many lack comprehensive linguistic resources such as well-annotated corpora, pre-trained word embedding's, named entity recognition models, and syntactic parsers. These resources are essential for linguistic analysis and are often readily available for major languages. Their absence further complicates the development of QA systems for Indian languages. Moreover, several Indian languages exhibit dialectal variations and regional differences. QA systems must be versatile enough to comprehend and respond accurately to these variations, adding yet another layer of complexity to the task

6) Strategies for Developing QA in Indian Languages

Addressing these multifaceted challenges demands the implementation of innovative strategies tailored to the unique linguistic landscape of India. Data augmentation emerges as a vital strategy to overcome the scarcity of data. By applying techniques such as data augmentation and synthetic data generation, it becomes possible to create additional data points, thereby expanding the available datasets for low-resource Indian languages, and ultimately improving QA system performance. Transfer learning is another valuable approach in the development of QA systems for Indian languages. Pre-trained models, initially trained on high-resource languages or domains, can be fine-tuned with limited data from Indian languages. This allows the model to leverage the knowledge captured in larger, related datasets and adapt it to specific Indian languages, mitigating the challenges posed by data scarcity. Engaging native speakers and local experts through crowdsourcing efforts can contribute significantly to addressing the scarcity of linguistic resources. Crowdsourcing can be instrumental in data collection, annotation, and translation, thus enriching linguistic resources and improving QA system accuracy. Leveraging multilingual models, such as mBERT (Multilingual BERT), proves highly effective. These models are designed to handle multiple languages, including Indian languages, by capturing linguistic commonalities. Integrating multilingual models enables the development of QA systems capable of comprehending a wide array of Indian languages, even in code-switching scenarios. Lastly, domain specialization plays a crucial role in improving QA system performance. By training QA models on domain-specific data, such as medical or legal texts, these models can provide more accurate and context-aware answers, addressing specific needs and applications.

IV. DATASETS FOR QA

Natural Language Processing (NLP) heavily relies on high-quality datasets to train and evaluate models. In this section, we provide an overview of several important NLP QA datasets that are curated to support various research tasks. These datasets cover a wide range of languages, domains, and challenges, making them valuable resources for the NLP community. Below, we present a summary of these datasets in table 1.

Table 1: Datasets for QA

Dataset	Description
TyDi QA datasets	Introduces TYDI QA, a question answering dataset covering 11 diverse languages with 204K question-answer pairs. Emphasizes its suitability for challenging and trustworthy multilingual model evaluations without translation bias[10].
HindiRC	Divided by primary education grade levels, this Hindi Reading Comprehension dataset contains natural questions and assesses traditional similarity metrics for question answering efficiency. Offers potential for resource-scarce languages, with future directions including machine learning-based methods and automated question generation [17]
Chaii	Addresses the underrepresentation of Indian languages like Hindi and Tamil on the web. Aims to improve NLU models' performance for these languages using a new question answering dataset called chaii-1[19].
MMQA	Curates 500 articles in six different domains, forms a comparable corpus of 250 English and 250 Hindi documents, and generates 5,495 question-answer pairs in both English and Hindi, covering various entities as answers. Designed for monolingual, cross lingual, and multilingual question-answering systems in English and Hindi [21].
XQuAD	Consists of 240 paragraphs and 1190 question-answer pairs translated into ten languages by professional translators. Serves as a comprehensive benchmark for cross-lingual evaluation, enabling researchers to assess the performance of models across different languages and linguistic variations [22].
XQA	A multilingual dataset designed for cross-lingual Open-domain Question Answering (OpenQA). It includes a training set in English and development/test sets in eight other languages, totaling 90,000 question-answer pairs. This dataset enables research into cross lingual OpenQA systems and language understanding [23].
XOR QA	Introduces Cross-lingual Open Retrieval Question Answering, a novel task that addresses information scarcity and asymmetry in multilingual question answering. Contains a large dataset with 40,000 questions in seven diverse non-English languages [26].
csebuetnlp Bangla QA	Presents BanglaBERT, a BERT-based NLU model pretrained on a substantial amount of Bangla data. Introduces downstream task datasets to create the Bangla Language Understanding Benchmark (BLUB)[27].
Facebook Multilingual QA datasets	Introduces MLQA, a multi-way aligned extractive question answering evaluation benchmark in seven languages, aiming to promote research in multilingual QA systems[28].
IITH HiDG	A Hindi Discourse Graph bank from the Indian Institute of Technology Hyderabad, useful for various NLP tasks. It aids in understanding context and meaning, making it valuable for language understanding and generation research in Hindi.
IITB HiQuAD	Focuses on solving the challenging problem of automatic question generation (QG) for languages with limited training data. Proposes a cross-lingual QG model and demonstrates its effectiveness for Hindi and Chinese.

v. CONCLUSION

This survey explores Question Answering (QA) in Indian languages, emphasizing its transformative impact on education, healthcare, and culture. It discusses challenges, analyses QA datasets, and highlights its role in linguistic preservation, healthcare, and global information access. The paper advocates for addressing low-resource languages in Natural Language Processing.

REFERENCES

[1] Sharma, Lovely, Vijay Dhir, and Kamaljeet Kaur. ”A New Model for Question-Answer based Dialogue System for Indian Railways in Hindi Language.” Indian Journal of Science and Technology 8.32 (2015): 1-4.

[2] Gupta, Poonam, and Vishal Gupta. ”A survey of existing question answering techniques for Indian languages.” Journal of emerging Technologies in web intelligence 6.2 (2014): 165-169.

[3] SHELKE, BHARAT A., and C. NAMRATA MAHENDER. ”SIMILARITY MEASUREMENTS OF CANDIDATE’S ANSWERS TO MARATHI QA SYSTEM.”

[4] Kumar, Praveen, et al. ”A Hindi question answering system for E-learning documents.” 2005 3rd International Conference on Intelligent Sensing and Information Processing. IEEE, 2005.

[5] Kumar, Praveen, et al. ”A query answering system for Elearning Hindi documents.” South Asian Language Review 13.12 (2003): 69-81.

[6] Das, Arijit. ”An alternate approach for question answering system in Bengali language using classification techniques.” INFOCOMP Journal of Computer Science 19.1 (2020).

[7] Sabane, Maithili, Onkar Litake, and Aman Chadha. ”Breaking Language Barriers: A Question Answering Dataset for Hindi and Marathi.” arXiv preprint arXiv:2308.09862 (2023).

[8] Shelke, Bharat A., and Ramesh R. Naik. ”Database Creation for Marathi QA System.” Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021). 2021.

[9] Islam, Md Aminul, et al. ”Word/phrase based answer type classification for bengali question answering system.” 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV). IEEE, 2016.

[10] Clark, Jonathan H., et al. ”Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages.” Transactions of the Association for Computational Linguistics 8 (2020): 454-470.

[11] Vemula, Rakesh, Mani Nuthi, and Manish Shrivastava. ”TeQuAD: Telugu Question Answering Dataset.” Proceedings of the 19th International Conference on Natural Language Processing (ICON). 2022.

[12] [Ramesh, Gowtham, et al. ”Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.” Transactions of the Association for Computational Linguistics 10 (2022): 145-162.

[13] Garg, Anjani, and Bharti Grover. ”Review on Punjabi Question Answer.”

[14] Manna, Partha Pratim, and Alok Ranjan Pal. ”Question answering system in Bengali using semantic search.” 2019 International Conference on Applied Machine Learning (ICAML). IEEE, 2019.

[15] Arora, Suket, Sarabjit Singh, and Simrandeep Singh Thapar. ”Question Answering based Dialogue System in Punjabi language.” Think India Journal 22.14 (2019): 14160- 14168.

[16] [Seena, I. T., G. M. Sini, and R. Binu. ”Malayalam question answering system.” Procedia Technology 24 (2016): 1388- 1392.

[17] Anuranjana, Kaveri, Vijjini Rao, and Radhika Mamidi. ”Hindirc: A dataset for reading comprehension in hindi.” 0th International Conference on Computational Linguistics and Intelligent Text. 2019.

[18] Terdalkar, Hrishikesh, and Arnab Bhattacharya. ”Framework for question-answering in Sanskrit through automated construction of knowledge graphs.” Proceedings of the 6th International Sanskrit Computational Linguistics Symposium. 2019.

[19] Thirumala, Adhitya, and Elisa Ferracane. ”Extractive Question Answering on Queries in Hindi and Tamil.” arXiv preprint arXiv:2210.06356 (2022).
[20] Shelke, Bharat A., and C. Namrata Mahender. ”Development of Question Answering System in Marathi Language.” Specialusis Ugdymas 1.43 (2022): 10176-10185.
[21] Gupta, Deepak, et al. ”MMQA: A multi-domain multilingual question-answering framework for English and Hindi.” Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
[22] Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. ”On the cross-lingual transferability of monolingual representations.” arXiv preprint arXiv:1910.11856 (2019).
[23] Liu, Jiahua, et al. ”XQA: A cross-lingual open-domain question answering dataset.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
[24] Maurya, Kaushal Kumar, et al. ”ZmBART: An unsupervised cross-lingual transfer framework for language generation.” arXiv preprint arXiv:2106.01597 (2021).
[25] Kumar, Vishwajeet, et al. ”Cross-lingual training for automatic question generation.” arXiv preprint arXiv:1906.02525 (2019).
[26] Asai, Akari, et al. ”XOR QA: Cross-lingual open-retrieval question answering.” arXiv preprint arXiv:2010.11856 (2020).
[27] Bhattacharjee, Abhik, et al. ”BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla.” arXiv preprint arXiv:2101.00204 (2021).
[28] Lewis, Patrick, et al. ”MLQA: Evaluating crosslingual extractive question answering.” arXiv preprint arXiv:1910.07475 (2019).