State of the Art and Emerging Trends in Indian Languages' Question-Answering Systems: An Overview

Mamta S. Bendale
Research Scholar, S.S.V.P.S's L.K. Dr. P. R. Ghogrey Science College Dhule, Maharashtra, India

Rupali H. Patil
S.S.V.P.S's L.K. Dr. P. R. Ghogrey Science College, Dhule, Maharashtra, India

Bhausaheb V. Pawar
KCE's Institute of Management & Research, Jalgaon, Maharashtra, India

ABSTRACT
As digital technologies continue to evolve, the demand for effective natural language processing (NLP) systems tailored to Indian languages has grown significantly. Most of the researchers has been concentrated on languages with rich digital resources, frequently failing to acknowledge the Indian subcontinent's linguistic diversity. This review paper seeks to address this deficiency by offering a thorough summary of the progress and improvements in Question Answering Systems(QASs), particularly for Indian languages. With 22 official languages recognized under the eighth schedule of the Indian Constitution, India has a vast linguistic variety. QASs are an essential part of NLP that try to understand user inquiries and provide human-like answers. In this study, an overview of QAS research for several Indian languages is presented. The analysis and outcomes regarding Categories of Indian Languages, Necessity of Indian Languages' Question-Answering Systems, applications of QAS and approaches of QAS's are reported. This review distinguishes itself from previously available reviews by including languages such as Odia, Sanskrit, Kannada, Assamese, and others. The study also highlights that BERT and SVM are the most used models for developing QAS in various Indian languages. All conclusions are drawn based on a review of the literature.

Keywords
Automatic Question Answering, Categorization of Indian Languages, Applications and Approaches of Question Answering System, Indian Languages in Question Answering System

1. INTRODUCTION
The quantity of information on the Internet has significantly increased in the last few years. Many times, people arrive with certain inquiries in mind that they wish to have answered. They constantly want to ask questions through their own language, without having knowledge to a certain query language, set of rules for query construction, or even a particular knowledge area. They would want to know the replies to be brief and accurate. To better fit the user's demands, the most recent method is to really investigate what user actually want with linguistic perspective and try to figure out what the user actually means.[1].Although common users lack skills in computer science, statistics or linguistics, they do have some familiarity with their native tongue and prefer it when interacting with computers.[2].

Using questions in natural languages as input, the QAS finds the exact response to queries in natural languages by searching through a collection of texts. It's not the same as information extraction (IE) or information retrieval (IR). The information retrieval system (IR) provides consumers with a collection of papers pertaining to their inquiries, but it does not provide precise answers.[1]. While using a QAS, the user requires a succinct, understandable, and accurate response. This response might be related to a particular word, sentence, paragraph, image, audio clip, or complete document. [3].

According to [4] there are three modules in QAS: 1) Question analysis 2) Document retrieval and 3) Answer extraction. The question processing module is a significant component of QA systems. This module processes the question. If there are issues with this module, it will affect other parts.[1]. Processed question given as an input to the second component document retrieval which tries to retrieve documents related to question. Retrieved documents are finally sent to the third component to get accurate answer.

Question-answering systems that support Indian languages are becoming more and more necessary as digital technology spreads throughout India. There is a need to enhance question-answering systems for Indian languages so that accurate information can be easily accessible by a larger population in their own language. Comprehensive reviews that concentrate only on Indian languages in the context of QAS is limited. This paper could close this gap by giving a summary of previous studies, the current development and potential future directions in this field,

Writing a review paper on Indian language QASs would involve examining the techniques, algorithms, methodologies and advancements in the development of QASs specifically tailored for Indian languages. In this paper, we have examined previous research, particularly for Indian language QASs. The structure of this paper is as follows: The procedure for gathering data for this research is outlined in Section 2. Section 3 presents, Categories of Indian Languages. Section 4 outlines the Languages of Indian States. Section 5 reports the Necessity of Indian Languages' Question-Answering Systems. Section 6 reports Applications of QAS. Question Answering Approaches of QAS is described in section 7, Literature Review is present in section 8 and Section 9 presents the findings of the research.

2. INFORMATION SOURCES
To collect the data for this study, several online digital libraries such as IEEE, ACM, ArXiv, Google Scholar, Springer Link, ACL Anthology and Science Direct. For searching we use keywords like QAS for Indian language, Answer extraction for Indic language, automatic answer generation etc., combinations of keywords and their synonyms. The title and abstract of the publication were searched for these terms.

The categorization of Indian languages plays a crucial role in understanding the nuances of linguistic research, language policy, and language technology. The following section provides a comprehensive breakdown of the categorization of Indian languages.

3. CATEGORIES OF INDIAN LANGUAGES

Four general categories can be used to categorize Indian languages. These are:

1. Indo-Aryan 2. Dravidian 3. Sino-Tibetan 4. Austric

3.1 Indo-Aryan

It is a member of the Indo-European language family, which the Aryans brought to India. With roughly 74% of all Indians speaking it, it is the largest language group in the country. The language group includes all of the major languages spoken in northern and western India including Bengali, Marathi, Gujarati, Hindi, Oriya, Pahari, Bihari, Kashmiri, Gujarati, Punjabi, Sindhi, Rajasthani, Assamese, and Urdu [5].

3.2 Dravidian

Most of the languages in this group are those spoken in southern India. Centuries before the Indo-Aryan language, the Dravidian language arrived in India. Approximately 25% of Indians are covered by it. Three general groupings of Dravidian languages can be distinguished: The Northern group, the Central group, and the Southern group. The major languages of the Dravidian group are: Telugu, Tamil, Kannada and Malayalam.

3.3 Sino-Tibetan Group

The Sino-Tibetan, also known as Mongoloid, speech family is found throughout the sub-Himalayan regions of India, including North Bihar, North Bengal, Assam, and the country's north-eastern borders. The earliest Sanskrit literature refers to these languages called Kiratas as being older than the Indo-Aryan languages. The proportion of speakers of these languages in India is about 0.6%. The two principal Sino-Tibetan languages are Bodo and Manipuri.

3.4 Austric Group

Two language families within the Austric family are Austroasiatic and Austronesian (the latter was formerly known as Malayo-Polynesian). They are spoken in the Pacific Islands, Southeast Asia, and India. With over 5 million speakers, the most important Austric language and the most commonly spoken Adivasi language is Santhali. Following figure show categorization of Indian languages

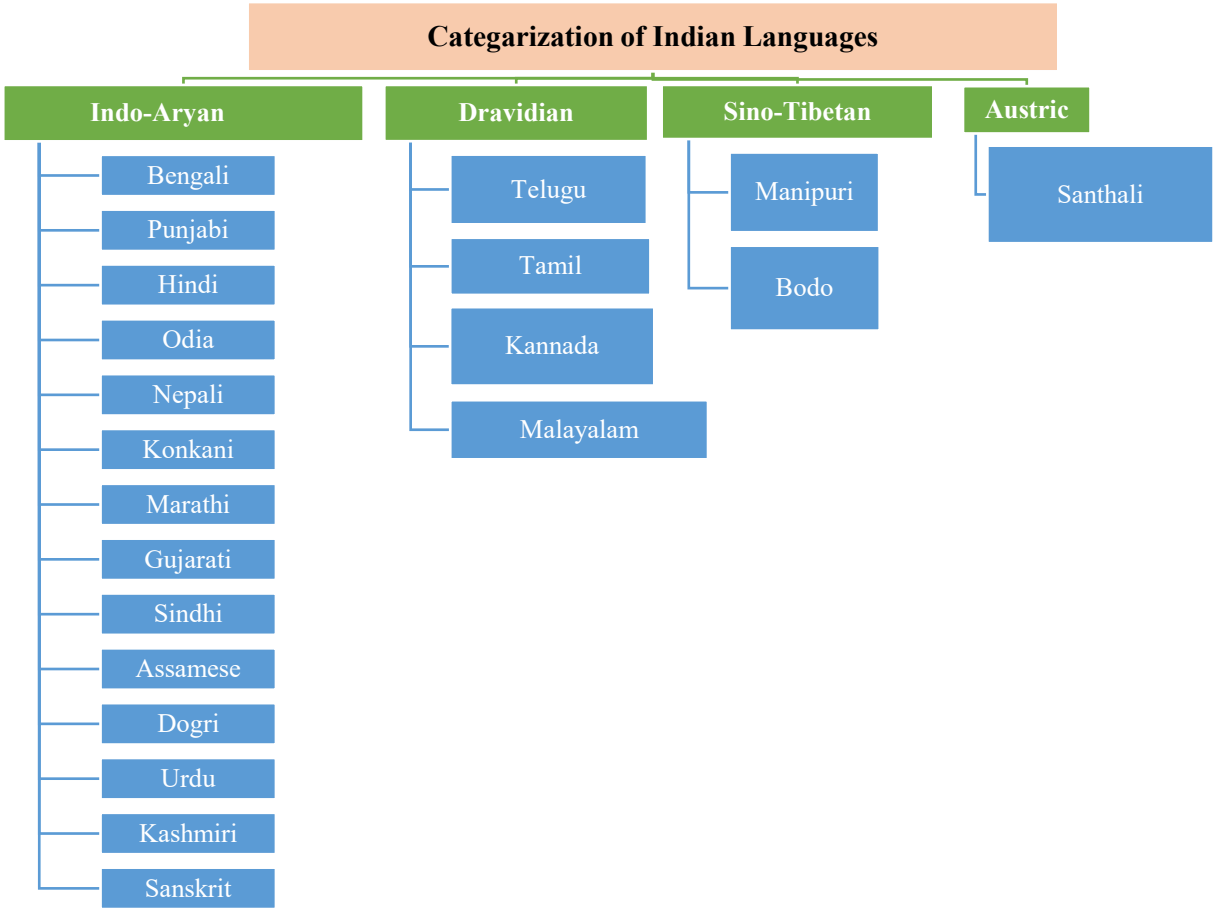

Figure 1: Categorization of Indian Languages

4. LANGUAGES OF INDIAN STATES

Language is an essential component of communication, as it enables the clear expression and understanding of thoughts, feelings, and ideas on various topics. There is multiple languages spoken in different states of India[6]-[7]. The list of languages and the states in which they are spoken are displayed in the table 1 and the figure 2 displays the percentage of languages spoken in India

Table 1: Languages of Indian States

Sr.No.	State	Language	Sr.No.	State	Language	Sr.No.	State	Language
1	Andhra Pradesh	Telugu	11	Karnataka	Kannada	21	Rajasthan	Hindi
2	Arunachal	English	12	Kerala	Malayalam	22	Sikkim	English, Nepali
3	Assam	Assamese	13	Madhya Pradesh	Hindi	23	Tamil Nadu	Tamil
4	Bihar	Hindi	14	Maharashtra	Marathi	24	Telangana	Telugu
5	Chhattisgarh	Hindi	15	Manipur	Manipuri	25	Tripura	Bengali, English
6	Goa	Konkani, English	16	Meghalaya	English	26	UttarPradesh	Hindi
7	Gujarat	Gujarati	17	Mizoram	Mizo	27	Uttarakhand	Hindi
8	Haryana	Hindi	18	Nagaland	English	28	West Bengal	Bengali, English
9	Himachal	Hindi	19	Odisha	Odia			
10	Jharkhand	Hindi	20	Punjab	Punjabi			

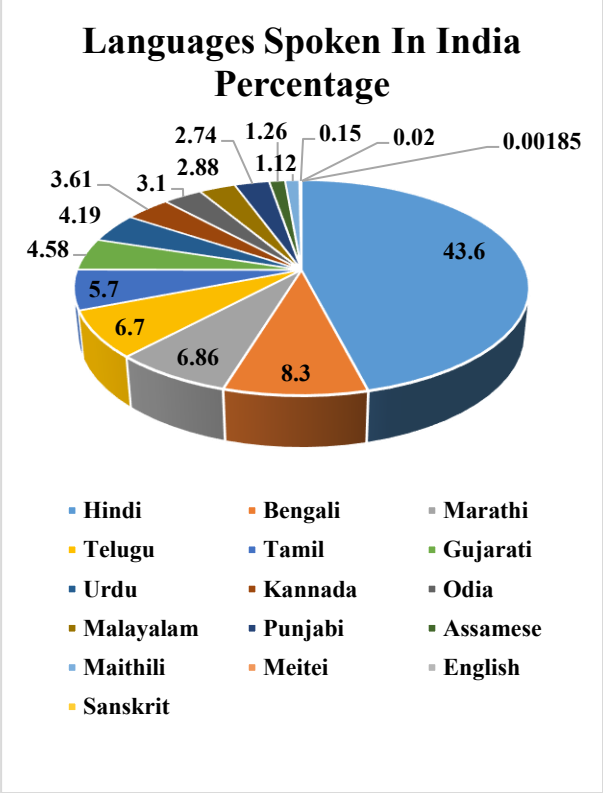

Figure 2: Languages Spoken in India Percentage[6]

5. NECESSITY OF INDIAN LANGUAGES' QUESTION ANSWERING SYSTEMS

Indian Languages' Question-Answering Systems are crucial for several reasons:

1.**Accessibility and Inclusivity**: India is a multicultural country where several languages spoken across the whole country. Indian language support for QASs improves inclusivity and accessibility by making information easily accessible to speakers of different Indian languages.

2.**Empowerment of Non-English Speakers**: India has a wide range of English proficiency levels, with many people feeling more at ease speaking their original tongues. Non-English speakers gain more power when a system for answering questions is available in their native tongue, enabling them to obtain information and services more efficiently.

3.**Preservation of Cultural Heritage:** Indian languages hold a wealth of cultural heritage in addition to being useful for communication. By making information in regional languages more accessible, an Indian Language QAS contributes to the preservation and advancement of this legacy.

4.**Support for Education**: In India, many students study in their mother tongues. Students who want educational resources and support in their native language can greatly benefit from a QAS that supports Indian languages.

5.**Enhanced Collaboration and Communication:** Language barriers can impede effective communication and collaboration in a country like India, which has a diverse population of languages. An Indian language question and answer system makes it easier for people and organizations all over the nation to collaborate and communicate.

6.**Market Expansion and Economic Growth**: With a sizable population that prefers to communicate in their native tongues, India boasts a sizable and expanding market. Offering services in Indian languages, such as QASs, can open up new market niches and boost the economy.

7.**Government Initiatives and Policies**: Through several programs and regulations, the Indian government has placed a strong emphasis on the promotion and preservation of Indian languages. These initiatives, as well as the government's objectives of linguistic diversity and inclusivity, are supported by the development of Indian Languages' Question-Answering Systems.

Overall, an Indian Language QAS helps the nation's population with its varied linguistic needs and advances the larger social objectives of inclusivity, cultural preservation, education, and economic growth.

6. APPLICATIONS OF QUESTION ANSWERING SYSTEM

QASs have a wide range of applications across various domains:

1. **Information Retrieval**: Systems for answering questions can be used to retrieve particular information from vast amounts of text. This is especially helpful in fields like customer service,

where customers can ask questions and get prompt responses without having to browse through large amounts of paperwork or databases.

2. **Education**: Systems that respond to questions can be used as teaching aids by giving students prompt feedback and answers to their inquiries. They can be included into e-learning systems to improve the way students learn about everything from history to arithmetic.

3. **Healthcare**: Healthcare workers can obtain medical information, guidelines, and research findings more quickly with the aid of question-answering systems in the healthcare industry. These systems can also be used by patients to research questions and concerns they may have about their health.

4. **Legal and Compliance**: Aside from helping legal professionals with research into statutes, case law, and legal documents, QASs can also help guarantee regulatory compliance by responding to inquiries regarding policies and regulations.

5. **Content Curation**: QASs can be used by publishers and content creators to automatically create FAQs (Frequently Asked Questions) or to suggest products, videos, or articles to users based on their interests and queries.

6. **Virtual Assistants**: Virtual assistants, such as Chabot and voice assistants, can be powered by QASs, which allow them to comprehend and reply to user inquiries in natural language. These virtual assistants can be used in a variety of settings, such as virtual receptionists, smart home appliances and customer service.

7. **Research and Data Analysis**: QASs can be used by researchers and analysts to better perform literature reviews or to glean insights from large datasets. These tools are useful for finding pertinent research papers, gathering important data, and finding answers to particular research queries.

8. **Language Translation**: By providing answers to questions in one language based on information available in another language, QASs can help with language translation. This can be especially helpful for communication and information retrieval across linguistic boundaries.

9. **Search Engines Enhancement**: By permitting users to enquire questions in natural language, question answering systems can enhance conventional keyword-based search engines. As a result, search results are more relevant and the user experience is improved.

10. **Customer Support**: QASs are widely used by businesses to offer automated customer support. By understanding consumer inquiries and responding appropriately, these systems can speed up response times and minimize the need for human intervention.

11. **Compliance and Regulatory Assistance**: By offering responses to inquiries concerning rules, guidelines, and practices, QASs can assist companies in maintaining regulatory compliance.

7. QUESTIONANSWERING APPROACHES

In literature several approaches are available to build question-answering systems. Some common approaches are:

1.Rule-Based Systems: These systems make use of pre-established rules and patterns. In this approach to match questions with relevant answers, regular expressions or custom rules are frequently used. Although this approach can be accurate in well-defined domains, it may struggle with complex questions and are unable to deal with ambiguity[8]. By using predefined patterns, these rules are able to classify questions according to the type of answer. The path that leads to the right answer was determined by using these grammatical rules, which represent the context as decision trees[9]. The requirement of manually creation of the heuristic rules is a significant disadvantage of rule-based QASs. A thorough understanding of a language's semantics was required to create these rules[10].

2.Statistical Approach- Statistical techniques not only work with formal query languages but may also create natural language questions. When properly learned, these approaches outperform other state-of-the-art methods in terms of results. Basically, they require enough data for precise statistical learning [11]. Generally speaking, statistical methods have so far been effectively implemented across the various phases of a quality assurance system. Several techniques have been used for question classification, including maximum entropy models, Bayesian classifiers, and support vector machine (SVM) classifiers[12]. These statistical techniques evaluate queries to forecast the kind of response that users are likely to provide. These models are trained using a corpus of queries or documents annotated with the specific categories listed in the system[13].

3.Machine Learning Approach: The answer to a question can be obtained by applying different machine learning techniques, including deep learning, unsupervised learning and supervised learning. Utilizing training data, these models pick up patterns and relationships that enable them to tackle increasingly challenging queries and situations. Algorithms in the QA domain are now able to comprehend linguistic features without explicit instruction thanks to the introduction of machine learning. This strategy, which allows the system to analyse an annotated corpus and subsequently create a knowledge base, was made possible by statistical techniques.In linguistic and sentiment-related domains, machine learning is frequently integrated with statistical methods [14].

4.Deep learning Approach- Because deep learning uses neural networks to learn underlying features in data, it is fundamentally different from machine learning. Each simple, connected unit in a standard neural network (NN), known as a neuron, generates a series of real-valued activations[15]. Sensors that detect their surroundings activate input neurons, and weighted connections from previously active neurons activate other neurons[15]. Some neurons can trigger responses that change their environment. Every input has a weight attached to it that indicates how important it is in relation to other inputs[15]. Deep learning models have demonstrated notable success recently in a range of natural language processing tasks, including text summarization, machine translation, and semantic analysis [16]. Textual context is mapped into logical representations by neural network architectures, which are then applied to answer prediction. Bidirectional Long Short-Term Memory (LSTM) units are used by these neural networks for question processing and response classification [16].

5.Hybrid Approaches: A lot of contemporary QASs combine the methods mentioned above to take their individual advantages. A system can combine rule-based or IR techniques with machine learning models to extract precise answers by understanding context and relevance.

The accuracy, scalability, and complexity of each method are trade-offs. The degree of precision required, the accessibility of training data and the difficulty of the questions all influence the approach that is selected.

8. STATUS OF QA WORK FOR INDIAN LANGUAGES

The research on question-answering systems in the Indian language shows a consistent development with notable turning points and ongoing improvements. Following section shows the summary of literature review on different Indian languages.

8.1 Bengali

A pioneering study by researchers [17] have developed first Bengali factoid QAS. Their system consists of three modules: question analysis, sentence extraction, and answer extraction. The question analysis step consists of five steps: question type (QType) identification, named entity identification, question topical target (QTT) identification, keyword identification and expected answer type (EAT) identification. The output of the question analysis module is used to create a query. The sentence is then extracted from the paragraph using this query, and the results are ranked according to the answer score value. For their experiment, they have gathered information related to geography and the agricultural sector from Wikipedia.[18] presented an Informative QAS for Bangla language. They performed their experiment using cosine similarity, Jaccard similarity, Naïve Bayes and achieved 93.22%,84.64% and 91.31% accuracy respectively. Study by [19]describes a method to identify semantically relevant answers in the Bengali dataset. Several statistical parameters, including part-of-speech (POS), frequency, and index, were used to compared similarity between a question and answer in the first section of the algorithm. In separate modules, entropy and similarity were computed in the second stage. Lastly, a sense score was produced to order the responses. A repository having 275000 sentences in total was used to test the algorithm. The Language Research Unit of the Indian Statistical Institute, Kolkata, provided resources for this Bengali repository as part of the Technology Development for Indian Languages (TDIL) project, which is funded by the Indian government. The Language Technologies Research Centre LTRC group at IIIT Hyderabad developed the shallow parser, which was used for POS tagging. For their experiment, they presented an approach that consists of seven modules: question preprocessing, frequency calculation, matching index calculation, entropy difference calculation, POS matching, cosine similarity calculation, and sense matching. They achieve 97.32% accuracy with this approach.[20]proposed Bengali QAS based on Deep learning. For their experiment they have used transfer learning to transfer standard English data SQuAD 2.0. into Bengali, they implemented fuzzy matching to reserve the superiority of answer after translation. For zero shot transfer learning they have used multilingual BERT model and fine tune it for Bengali reading comprehension. They also used RoBERTa and Distil BERT for comparison purpose. They conducted a study of third and fourth grade students at Cambrian College in Dhaka and compared the model's performance on reading comprehension tests in Bengali QA with human children. They achieved F1 score of 66.67% in their study. Research by [21] proposes an effective query answering system for text retrieval in Bengali. To improve grammatical similarity and get suitable Bengali textual resources for user inquiries, this system incorporates word embedding clustering and deep level feature representation. Utilizing a deep belief network, the pre-trained word embedding module is generated. POS tagging, semantic similarity estimation, ranking, inverse filtering, DBN, and various pre-processing stages are all included in this system. Pre-processed data is provided as an input to the global word representation for their research. This global word representation is produced by combining TF-IDF based pre-trained word embedding, character based Bi-LSTM word embedding and similarity-based affix level embeddings. By utilizing semantic similarity, the answer is provided by the answer retrieval or answer ranking module. The user receives this response and if the user is not satisfied, the Deep Belief Network (DBN) module receives the ranked sentences and determines which Bengali sentences are the best based on previously collected data. They attain 97% accuracy for the TDIL dataset and 98.5% accuracy for the SQuAD 2.0 dataset.

8.2 Hindi

In a ground-breaking effort, researchers [22] developed cross lingual QAS which comprise English and Hindi language. Their system receives an English query, interprets the answer into Hindi, and then translates the Hindi response back into English. Their architecture contains four modules Question examiner, cross-lingual information retrieval (CLIR) system, answer finder, and MT system. To identify the answer input query is analysed by using chunker part-of-speech tagger and predefined key question patterns. To easily access the information from Hindi Documents.[2] developed Hindi QAS. The Automatic Entity Generator module in their architecture creates entities automatically based on a given question. Another three modules are question classification, question parsing, and query formulation. These modules help classify, parse, and generate questions, respectively. The answer extraction module uses the generated query as an input to find relevant passages. The answer selection module then chooses the best-ranked response and displays it to the user based on the final score. Their architecture archives 75% accuracy.[23]developed a Hindi QAS "PRASHNOTTAR". Query Pre-processing, the first part of their architecture, processes and analyses the input question. The query is generated by the next query generation module using the Query Logic Language. This query is used by the database search module to look up relevant documents, and the answer display module shows the result to the user at the end. When, where, what time, and how many kinds of questions they deal with for their experiment. Among from those 4 categories what time type questions give highest accuracy 80%. Overall accuracy of their architecture is 68%.[24]developed Graphical User Interface to convert Hindi Language into equivalent SQL query. Their architecture contains four modules Tokenizer, Mapper, Query generator and DBMS.[25] developed an Interface which is Domain-Independent to convert Hindi sentence into equivalent Relational Database. This work is an expansion of [24]. The language processing module and the database module are the two separate components of their system. Language processing module contains four sub module like Query analyser, POS tagger, Morphological Analyser and Semantic Analyser. Database module contains three sub module like SQL Query Generator, Domain Identifier and SQL Query Executer. In database module there is a domain-identifier component which uses knowledge base to recognize exact domain. Database module generates and executes SQL query corresponding to Hindi query provided by the user. Overall accuracy of their application is 87%.[26] developed a Hindi QAS which uses Machine Learning Approach. Their architecture mainly contains three phases Accessing NL Query phase, Classification, Feature and Extraction Phase. The input query is read and pre-processed in the first phase using stemming, stop word removal, and tokenization techniques. Second phase again divided into two sub modules: Entity Detection and Feature Extraction. In Entity Detection entity can be identified using similarity measure and feature vector can be created using Term Frequency. In last Classification phase the correct label of class can be identified by using Naive Baye's Classifier. Their research achieves 92% accuracy for test set 1 in which user know about domain and 88% accuracy for test set 2 in which user do not know about

domain.[27] developed a Multi-lingual and Multi-Domain Question-Answering Framework for Hindi and English language. For their experiment they created their own dataset MMQA. Dataset creation done in three different stages: Comparable Article Curation followed by Question Answer Formulation and finally last stage is Validation. For dataset creation they collect the data from different web sources by using web crawler. They collect 5,495 Question Answer pairs from 500 articles. Question processing module of their architecture contains Question Classification and Query Formulation phase. Next Passage Retrieval module uses Lucene's text retrieval functionality to extract the passage from given document. Coarse class and finer class of a question is utilized to extract candidate answer. Last module Answer Scoring and Ranking in which scores are calculated using Term coverage (TS), Proximity score (PS), N-Gram coverage score (NS), Semantic Similarity Score (SS) and Pattern matching score (MS) methods. Maximum score answer is displayed as a final answer.[28] developed an internet based application for Hindi QAS. Their application includes Question Interface, Question Classification, Query Formulation, Answer Extraction and Display answer modules to perform different task for question answering. The CYK parser, query formulation, NLP building modules, and concept of entities are all integrated into the Java implementation, which enables the system to function in any domain. [29]developed Extractive QAS for Hindi and Tamil queries . For their experiment they have used three models XLM-RoBERTa, XLM-RoBERTa+finetune and RoBERTa + Hindi finetune/Tamil finetune Hugging face, an NLP company, provided them with a package called transformers, which they used to import all these models into their code. (Wolf et al. [2019]). A computationally effective deep learning model called a transformer adds attention, enabling it to assign varying degrees of significance to various input components. No convolutions are used in it. Depending on which model is chosen, these transformers are pre-trained models that Tensorflow or PyTorch can use. They got word-level Jaccard score for XLM-RoBERTa is 65.6%, XLM-RoBERTa+finetune is 74.9% and RoBERTa + Hindi finetune/Tamil finetune 95.8 % for Hindi and 82.9% for Tamil, which clearly shows that RoBERTa + Hindi finetune/Tamil finetune model gives highest score.

8.3 Malayalam

A Malayalam QAS for answering factoid questions was designed by researchers [30]. Three modules that comprise their architecture are Question Type Analysis, Document Selection and Processing, and Answer Extraction. Question type analysing phase find question word from the question and except question word all remaining words are selected as keywords and lemmatized. In next phase sentence tokenizer and word tokenizer is used to separate out sentence and word from documents. Pattern matching method is used to find out rank answer. After receiving highly ranked sentence TnT tagger uses second order Markov model to tag the words. Using Vibhakthi and POS Tag Analysis [31] developed a Rule Based Malayalam QAS. Their architecture analyses the question type and provides word-level answers by using the POS and Vibhakthi tags of the questions. The most similar sentences from the entire document are found using the keyword-matching technique. These sentences are then broken down into their constituent words, and the sentence with the highest weight and word level matching is shown as an answer. For Improving Malayalam Question Answering [32] developed a Neural Word Embedding Based Transformer Model based on Health Domain. Their architecture finds the answer of factoid type question. Their architecture contains pre-processing, retriever, and reader module. Pre-processing phase involves Tokenization and Lemmatization. Retriever module retrieves most possible documents by converting query and documents into a unique file format, applying TF/IDF and document vectorization by using word embedding. Potential documents are sent to the reader module, which employs a logical score to display an accurate response. For implementation they used Distil BERT model with 6 layers and 66 million parameters.

8.4 Marathi

Using NLP techniques, ontology extraction for the agriculture domain has been carried out in the Marathi language by [33] .The goal of the agriculture domain system's model ontology is to find pertinent responses to the farmer's questions. Their proposed system has three modules Pre-processing, Keyword identification and Knowledge extraction. Part-of-Speech Tagging, Tokenization, Stop Word Removal, Stemming and Syntactically driven parsing methods are used for pre-processing. In Keyword identification module they find 1 to 3 keywords for each question. Rule-based and Conditional Random Fields (CRF's) Methods are used for knowledge Extraction by researchers. Researchers [34] presented Marathi QAS using ontology that uses the idea of ontology as a formal way to represent a knowledge base from which answers can be extracted. To express domain-specific knowledge about restrictions and semantic relations in the specified domains, ontology is used. Experts in the field assist in developing the ontologies, and a syntactic and semantic analysis is conducted on the query to extract the answer from the database. General Objective of their research is to find accurate and systematically correct answer. Their architecture uses sequence of processes like Tokenization, word grouping, POS Tagging and Chunking. After chunking query triples are extracted and then generate onto triples which matches with ontology to generate exact answer.[35]uses Transfer Learning to developed QAS for low resource language Marathi. The system tries to provide a paragraph to the user's query. For this work, Marathi has been selected as the low resource language. The questions and paragraph are written in natural language. Their framework uses multilingual BERT with 110M parameters. In their work answer having two fields, answer text and starting position of answer. As the part of pre-processing, they use BERT tokenizer for tokenization.[36] designed Marathi Language QAS. In their architecture answer is retrieved from model answer database and text corpus. Fetched answers are compared using sequence model. For pre-processing Punctuation Removal, Word tokenization and Sentence tokenization methods were used by researchers. For Corpus development they collect the information from 2nd ,3rd and 4th classes of Balbharati Marathi text book.[37] designed a Marathi QAS using deep learning .For their experiment they created MrSQuAD dataset by translating SQuAD1.1 dataset into Marathi . They used different multilingual and monolingual models over Marathi QA dataset like DistilBERT Multilingual , mBERT , XLM-RoBERTa , IndoAryanXLM,RoBERTa , MuRILMahaBERT ,Indic BERT ,MahaRoBERTa ,MahaAlBERT , Marathi DistilBERT , DevBERT , DevRoBERTa , DevAlBERT , DevBERT-scratch, MahaBERT , Indic BERT , MahaRoBERTa , MahaAlBERT , Marathi DistilBERT , DevBERT , DevRoBERTa , DevAlBERT , DevBERT-scratch.They achieve best performance by MuRIL multilingual model with an F1 score of 0.74.and EM score of 0.64

8.5 Punjabi

An online Question Answering approach for English and Punjabi was designed by [38] . He takes a question from user and pre-processed it by applying stop word removal techniques. He extracts important terms from the remaining question, uses the

Vector Space Model to find synonyms for these keywords, and then creates a query. Query is used to retrieve web pages and top 20 answers with highest score are displayed.[1] presented Hybrid Approach for Punjabi QAS for factoid Punjabi questions to improve accuracy in terms of precision and recall. Initially they find out the factoid questions from the given text. They classify the questions into different category and designed procedures for each category. For their experiment they design scoring architecture to calculate the answer score by using similarity matrix and pattern matching techniques to identify best answer from the set of answers given by the system.[39] designed Punjabi QAS based on Gravity. They extracted numerical features from question like Lexical Density, Epistemic Score, Point of Gravity, Readability Index and Matching Gravity Score. They achieve 91% accuracy in terms of precision.

8.6 Tamil

A Tamil QAS parser was introduced by researchers [40]. For their experiment they use Stanford parser modal and modify POS tag sets, dependency sets and chunk sets to gain expected outcome. MALT (Models and Algorithms for Language Technology) parser tool is used for dependency parsing. LIBSVM and LIBLINEAR algorithms of MALT tool used for classification purpose. Their system is used to parse the question and parse near about 50 thousand sentences preserved in the structured data base.[41] designed a Answer Prediction system for Tamil and Hindi Questions. For their experiment they follow four different approaches with four datasets. Approach 1 employs zero shot transfer with the Chaii dataset; Approach 2 fine-tunes the model using the multilingual model and performs k-fold validation with the Chaii dataset; Approach 3 increases the size of the dataset using the same multilingual model as Approach 2; and Approach 4 uses the SQuAD dataset and transfers it to Tamil in order to fine-tune the model. Jaccard Similarity, EM and F1 Score are used for evaluation. Their experiment shows that MuRIL Model gives better result than XLM-RoBERTa.

8.7 Assamees

To prepare an Assamese text for analysis, the following pre-processing methods are used by [42]: Sentence separator, token generator, removal of stop words, root word generator, and POS tagging. Logical representation and the Structured Text Generator modules receive the pre-processing phase's output as an input. The module known as Structured Text Generator produces structured text that includes subject(s), verb(s), instance(s), and object(s). The Parse Tree is used by the Logical Representation Module to generate logical rules that lead to the solution.

8.8 Odia

While searching the information on search engine names plays an important role but there are different variations of names are available due to this searching process becomes problematic. So to solve this issue [43] proposed an Automatic Approximate Matching Techniques Based on Phonetic Encoding for Odia Query. For their experiment they assign Soundex phonetic codes for each English alphabet and achieve 92% accuracy.

8.9 Kannad

To Answer Agricultural-Related Enquiries in the Kannada Language [44] proposed a Few-Shot Setting BERT Model Krishiq-BERT. For creating Krishiq Agricultural dataset they collected information of 24 crops along with government schemes for farmers. They collect this information from University of Agricultural Sciences, Dharwad. They developed closed domain QAS for Kannada language which can give answer of agricultural queries. Their model uses retriever–reader system, retriever module filter down 10 relevant documents according to question asked and reader module fine-tuned which is HuggingFace DistilBERT model searches topmost document out of the k filtered documents which are provided by the retriever module to give final answer.

8.10 Sanskrit

A framework for automatically constructing knowledge graphs in Sanskrit QAS was proposed by researchers [45] . For construction knowledge graph they pre-processed the text. After pre-processing identify the relationship between words and then identify the triplets. Triplets are enhanced and given to SPARQL querying language to form query pattern to build a knowledge graph. Final answer was extracted from knowledge graph.

Table 2 provides an overview of the observations derived from a review of previous research on QASs for non-Indian languages. It displays QASs created by different researchers for a number of non-Indian languages, along with the dataset, methodologies, and approaches utilized to produce the QAS and the accuracy that was attained.

Table 2: **Literature summary**

Sr. No	Language of Research and Reference	Methods/Approach /Model	Domain/Dataset	Accuracy
1	Bengali [17]	--	14 documents from geography and agriculture domain acquire from Wikipedia	MRR-32%
2	Bengali [18]	Cosine similarity, Jaccard similarity and Naïve-Bayes	Hall information, department information, teacher information, library, NSTU nature, bus schedules etc. of Noakhali Science and Technology University (NSTU)	Cosine similarity-93.22% Jaccard similarity - 84.64% Naïve Bayes algorithm-91.31%
3	Bengali [19]	Naive Bayes, Artificial Neural Network, Decision Tree and Support Vector Machine (SVM),Semantic Similarity	Technology Development for Indian Languages (TDIL)Bengali corpus	97.32
4	Bengali [20]	BERT, Distil BERT, RoBERTA	Bengali Wikipedia Bengali culture, SQuAD	F1 Score-66.67%
5	Bengali [21]	Bi—LSTM, TF-IDF, Semantic Similarity, Deep Belief Network(DBN)	TDIL and SQuAD 2.0.	TDIL Dataset- 97% SQuAD 2.0.- 98.5%
6	Hindi [22]	HMM pattern-matching	Hindi BBC news and dictionary	MRR-25%
7	Hindi [2]	Locality-based similarity heuristic	Documents available on the site of LTRC from Agriculture and Science domain	75%
8	Hindi [23]	POS	Stored Hindi text data on web	68%
9	Hindi [24]	Tokenizer, Mapper, Query generator	NLIDB	----
10	Hindi [25]	Shallow Parser for POS	Employee Payroll, Railway Enquiry and Student Information database	87%
11	Hindi [26]	Naïve Baye's, Machine Learning Approach	-------	TS1- 92% TS2- 88%
12	Hindi [27]	CNN-RNN based model for question classification	MMQA dataset Tourism, History, Diseases, Geography, Economics, Environment	MRR-49.10% for Factoid Questions BLEU-41.37% for Short descriptive question
13	Hindi [28]	similarity heuristic	Open Domain	----
14	Hindi [29]	XLM-RoBERTa XLM-RoBERTa+finetune RoBERTa + Hindi finetune/Tamil finetune	Challenge in AI for India (chaii) Dataset	XLM-RoBERTa-65.6% XLM-RoBERTa+finetune-74.9% RoBERTa + Hindi finetune/Tamil finetune-89.3%
15	Malayalam [30]	Pattern Matching, approach ,TnT Tagger, Second order Markov Model	Personality in Kerala sports	70 %

16	Malayalam [31]	Rule Based Approach	---	----
17	Malayalam [32]	BERT, DistilBERT, word embeddings	SQUAD dataset related to health	F1 Score-86%
18	Marathi [33]	Conditional Random Fields, Rule based Method	Agriculture	F-Score -76.98%
19	Marathi [34]	Ontology	History, Sports, City, Entertainment, Political, Festival	Accuracy -89.28%
20	Marathi [35]	Multilingual BERT	Wikipedia and news dataset	F1-score-56.7% Bert-score- 69.08%
21	Marathi [36]	N-gram Sequence model	2nd, 3rd and 4th Marathi book of Balbharti	70%
22	Marathi [37]	DistilBERT Multilingual , mBERT , XLM-RoBERTa , IndoAryanXLM,RoBERTa , MuRILMahaBERT ,Indic BERT ,MahaRoBERTa ,MahaAlBERT , Marathi DistilBERT , DevBERT , DevRoBERTa , DevAlBERT ,DevBERT-scratch, MahaBERT , Indic BERT , MahaRoBERTa , MahaAlBERT , Marathi DistilBERT , DevBERT , DevRoBERTa , DevAlBERT , DevBERT-scratch	MrSQuAD	MuRIL EM score -64% F1 score -74%
23	Punjabi [38]	Vector Space Model	Open Domain	---
24	Punjabi [46]	Pattern Matching and mathematical expressions	Sports	Precision-85.66% Recall-65.28% F-score-74.06% MRR-43%
25	Punjabi [39]	Random Sampling Method	9th and 10th textbook of physics	91%
26	Tamil [40]	Stanford parser model, Machine Learning ,SVM	Tourist domain	------
27	Tamil [41]	Multilingual-BERT, XLM-RoBERTA, MuRIL	Challenge in AI for India (chaii), XQuAD, MLQA, SQuAD	MuRIL with 5-fold Validation F1 Score-82.6%
28	Assamees [42]	Pattern Based Approach	--	--
29	Odia [43]	Matching Techniques	Collect data from Odia Language	92%
30	Kannada [44]	Krishiq-BERT	Agriculture	F1score-61.16%
31	Sanskrit [45]	Rule based Approach , Knowledge graph	Mahabharat,Ramayna, Bhavaprakasa Nighaṇṭu	50%

Following figure provides an overview of the research paper publication status for different Indian languages, offering insights into the number of research papers published across different Indian languages.

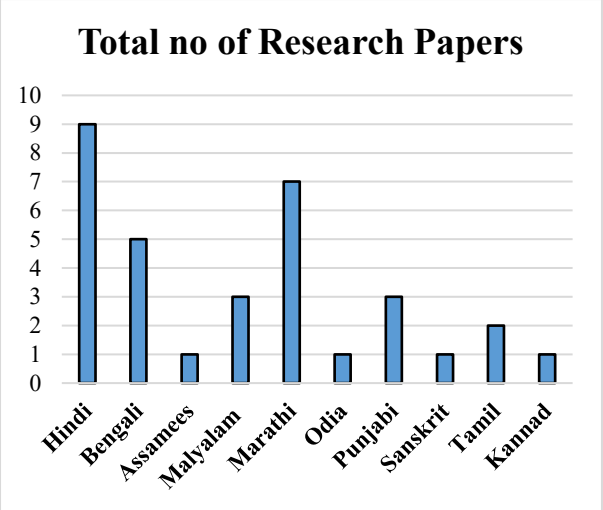

Figure 3: **Research status for various Indian languages**

Over the past few years, the development and application of QASs for Indian languages have gained significant growth. Following figure illustrates the year-wise publications for different Indian languages, specifically focusing on research related to QASs.

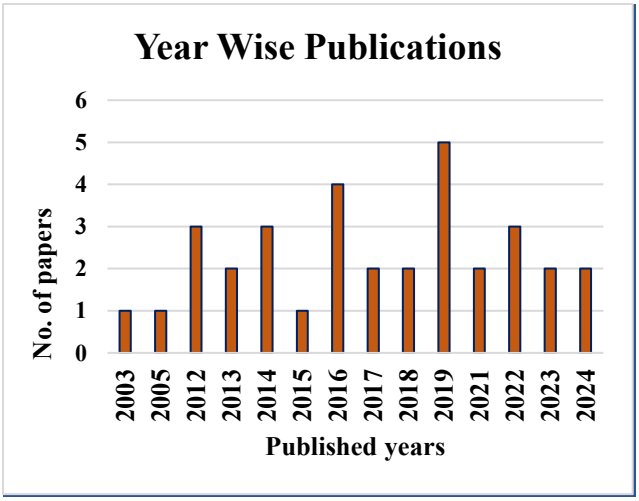

Figure 4: **Year Wise Publications for QAS**

According to the literature review, BERT and SVM are widely used at various stages of Indian language QAS development, as illustrated in Figure 5

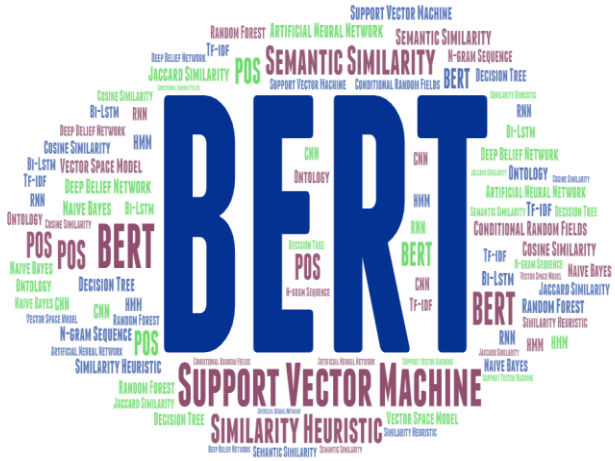

Figure 5: Overview of the literature

9. CONCLUSION

This paper focuses on an extensive overview of the state of Indian languages question-answering systems. It begins with an exploration of the categories of Indian languages, followed by a discussion on linguistic diversity in India. The paper also examines the necessity, applications, and approaches related to QASs. According to the findings of this study, it is found that the research work in the Indian languages QAS is still not matured as compared to resource rich languages. Some memorable researches are there in some Indian languages like Hindi, Bengali, Marathi, Malayalam, Punjabi and Tamil, very few research present in languages like Gujrati, Sindhi, Maithili, Kashmiri, Manipuri, Santhali, Konkani etc. Additionally, it is observed that most researchers used the SQuAD standard dataset, and very few create their own. This suggests that large-scale dataset creation is necessary to produce results that are more reliable and accurate. According to the study, BERT and SVM are more frequently used to develop QAS in many Indian languages. Building domain-specific QASs in low-resource Indian languages and producing high-quality, annotated training datasets are potential areas for future research. Additionally, investigating multilingual transfer learning strategies and integrating speech-based input present encouraging avenues for enhancing the performance and accessibility of Indian language QASs.

10. DECLARATIONS

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

11. REFERENCES

[1] P. Gupta and V. Gupta, “A survey of text question answering techniques,” *International Journal of Computer Applications*, vol. 53, no. 4, 2012, Accessed: Jul. 15, 2024.

[2] Kumar, Praveen, Shrikant Kashyap, Ankush Mittal, and Sumit Gupta. "A Hindi question answering system for E-learning documents." In 2005 3rd International Conference on Intelligent Sensing and Information Processing, pp. 80-85. IEEE, 2005.

[3] O. Kolomiyets and M.-F. Moens, “A survey on question answering technology from an information retrieval perspective,” *Information Sciences*, vol. 181, no. 24, pp. 5412–5434, 2011.

[4] M. M. Biltawi, S. Tedmori, and A. Awajan, “Arabic question answering systems: gap analysis,” *IEEE Access*, vol. 9, pp. 63876–63904, 2021.

[5] “Indian Languages: Classification of Indian Languages.” Available at: https://www.yourarticlelibrary.com Accessed on: 27 November 2024.

[6] “Official Languages of India List with States PDF Map Classical,” Guidely. Available at: https://www.indianetzone.com Accessed on:15 July 2024.

[7] R. P. Patil, R. P. Bhavsar, and B. V. Pawar, “Automatic marathi text classification,” *Int. J. Innovat. Technol. Expl. Eng*, vol. 9, no. 2, pp. 2446–2454, 2019.

[8] H. T. Madabushi and M. Lee, “High accuracy rule-based question classification using question syntax and semantics,” in *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers*, 2016, pp. 1220–1230. Accessed: Jul. 15, 2024.

[9] E. Riloff and M. Thelen, "A rule-based question answering system for reading comprehension tests," in *ANLP-NAACL 2000 workshop: reading comprehension tests as evaluation for computer-based language understanding systems*, 2000. Accessed: Jul. 15, 2024.

[10] S. M. Humphrey, A. Névéol, A. Browne, J. Gobeil, P. Ruch, and S. J. Darmoni, "Comparing a rule-based versus statistical system for automatic categorization of MEDLINE documents according to biomedical specialty," *J. Am. Soc. Inf. Sci.*, vol. 60, no. 12, pp. 2530–2539, Dec. 2009, doi: 10.1002/asi.21170.

[11] K. S. D. Ishwari, A. K. R. R. Aneeze, S. Sudheesan, H. J. D. A. Karunaratne, A. Nugaliyadde, and Y. Mallawarrachchi, "Advances in Natural Language Question Answering: A Review," Apr. 10, 2019, *arXiv*: arXiv:1904.05276. Accessed: Jul. 15, 2024.

[12] R. P. Patil, R. P. Bhavsar, and B. V. Pawar, "A NOTE ON INDIAN LANGUAGES TEXT CLASSIFICATION SYSTEMS," *Asian Journal of Mathematics and Computer Research*, vol. 15, no. 1, pp. 41–55, 2017.

[13] S. K. Dwivedi and V. Singh, "Research and reviews in question answering system," *Procedia Technology*, vol. 10, pp. 417–424, 2013.

[14] S. R. Muthutantrige and A. R. Weerasinghe, "Sentiment Analysis in Twitter messages using constrained and unconstrained data categories," in *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, IEEE, 2016, pp. 304–310. Accessed: Jul. 15, 2024.

[15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[16] M. Tan, C. dos Santos, B. Xiang, and B. Zhou, "LSTM-based Deep Learning Models for Non-factoid Answer Selection," Mar. 28, 2016, *arXiv*: arXiv:1511.04108. Accessed: Jul. 15, 2024.

[17] Banerjee, S., Naskar, S. K., & Bandyopadhyay, S. (2014). Bfqa: A bengali factoid question answering system. In Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17 (pp. 217-224). Springer International Publishing.

[18] Md. Kowsher, M. M. M. Rahman, S. S. Ahmed, and N. J. Prottasha, "Bangla Intelligence Question Answering System Based on Mathematics and Statistics," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh: IEEE, Dec. 2019, pp. 1–6. doi: 10.1109/ICCIT48885.2019.9038332.

[19] A. Das, J. Mandal, Z. Danial, A. Pal, and D. Saha, "A novel approach for automatic Bengali question answering system using semantic similarity analysis," *Int J Speech Technol*, vol. 23, no. 4, pp. 873–884, Dec. 2020, doi: 10.1007/s10772-020-09760-5.

[20] T. Tahsin Mayeesha, A. Md Sarwar, and R. M. Rahman, "Deep learning based question answering system in Bengali," *Journal of Information and Telecommunication*, vol. 5, no. 2, pp. 145–178, Apr. 2021, doi: 10.1080/24751839.2020.1833136.

[21] A. Das and D. Saha, "Deep learning based Bengali question answering system using semantic textual similarity," *Multimed Tools Appl*, vol. 81, no. 1, pp. 589–613, Jan. 2022, doi: 10.1007/s11042-021-11228-w.

[22] S. Sekine and R. Grishman, "Hindi-english cross-lingual question-answering system," *ACM Transactions on Asian Language Information Processing*, vol. 2, no. 3, pp. 181–192, Sep. 2003, doi: 10.1145/979872.979874.

[23] S. Sahu, "Prashnottar: A Hindi Question Answering System," *IJCSIT*, vol. 4, no. 2, pp. 149–158, Apr. 2012, doi: 10.5121/ijcsit.2012.4213.

[24] M. Dua, S. Kumar, and Z. S. Virk, "Hindi Language Graphical User Interface to Database Management System," in *2013 12th International Conference on Machine Learning and Applications*, Miami, FL, USA: IEEE, Dec. 2013, pp. 555–559. doi: 10.1109/ICMLA.2013.176.

[25] R. Kumar, M. Dua, and S. Jindal, "D-HIRD: Domain-independent Hindi language Interface to Relational Database," in *2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, Chennai, India: IEEE, Apr. 2014, pp. 81–86. doi: 10.1109/ICCPEIC.2014.6915344.

[26] G. Nanda, M. Dua, and K. Singla, "A hindi question answering system using machine learning approach," in *2016 international conference on computational techniques in information and communication technologies (ICCTICT)*, IEEE, 2016, pp. 311–314. Accessed: Dec. 05, 2023.

[27] Gupta, Deepak, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. "MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi." In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.

[28] Stalin, Shalini, Rajeev Pandey, and Raju Barskar. "Web based application for hindi question answering system." International Journal of Electronics and Computer Science Engineering 2, no. 1 (2012): 72-78.

[29] A. Thirumala and E. Ferracane, "Extractive Question Answering on Queries in Hindi and Tamil," Sep. 26, 2022, *arXiv*: arXiv:2210.06356. Accessed: Dec. 05, 2023.

[30] I. T. Seena, G. M. Sini, and R. Binu, "Malayalam Question Answering System," *Procedia Technology*, vol. 24, pp. 1388–1392, 2016, doi: 10.1016/j.protcy.2016.05.155.

[31] S. M. Archana, N. Vahab, R. Thankappan, and C. Raseek, "A Rule Based Question Answering System in Malayalam Corpus Using Vibhakthi and POS Tag Analysis," *Procedia Technology*, vol. 24, pp. 1534–1541, 2016, doi: 10.1016/j.protcy.2016.05.124.

[32] L. S. K and M. I. P, "'A Neural Word Embedding Based Transformer Model for Improving Malayalam Question Answering on Health Domain,'" *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 14s, Art. no. 14s, Feb. 2024.

[33] P. Dalvi, V. Mandave, M. Gothkhindi, A. Patil, S. Kadam, and S. Pawar, "ONTOLOGY EXTRACTION FOR AGRICULTURE DOMAIN IN MARATHI LANGUAGE USING NLP TECHNIQUES.," *ICTACT Journal on Soft Computing*, vol. 7, no. 1, 2016, Accessed: Dec. 05, 2023.

[34] S. S. Govilkar and B. J. W, "Question Answering System Using Ontology in Marathi Language," *IJAIA*, vol. 8, no. 4, pp. 53–64, Jul. 2017, doi: 10.5121/ijaia.2017.8405.

[35] A. Phade and Y. Haribhakta, "Question Answering System for low resource language using Transfer Learning," in *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, IEEE, 2021, pp. 1–6. Accessed: Dec. 05, 2023.

[36] B. A. Shelke and C. N. Mahender, "Development of Question Answering System in Marathi Language," *Specialusis Ugdymas*, vol. 1, no. 43, pp. 10176–10185, 2022.

[37] Amin, Dhiraj, Sharvari Govilkar, and Sagar Kulkarni. "Question answering using deep learning in low resource Indian language Marathi." arXiv preprint arXiv:2309.15779 (2023).

[38] V. Gupta, "A Proposed Online Approach of English and Punjabi Question Answering," *International Journal of Engineering Trends and Technology*, vol. 6, no. 5, 2013.

[39] G. S. Dhanjal, S. Sharma, and P. K. Sarao, "Gravity based Punjabi question answering system," *International Journal of Computer Applications*, vol. 147, no. 3, p. 21, 2016.

[40] R. Sankaravelayuthan, M. Anandkumar, V. Dhanalakshmi, and S. N. Mohan Raj, "A Parser for Question-answer System for Tamil," *QA System Using DL*, vol. 229, p. 230, 2019.

[41] R. V. Namasivayam and M. Rajan, "Answer Prediction for Questions from Tamil and Hindi Passages," *Procedia Computer Science*, vol. 218, pp. 1985–1993, 2023.

[42] Sarma, Shikhar Kr, and Rita Chakraborty. "Structured and Logical Representations of Assamese Text for Question-Answering System." In Proceedings of the Workshop on Question Answering for Complex Domains, pp. 27-38. 2012.

[43] R. C. Balabantaray, B. Sahoo, S. K. Lenka, D. K. Sahoo, and M. Swain, "An Automatic Approximate Matching Technique Based on Phonetic Encoding for Odia Query," vol. 9, no. 3, 2012.

[44] P. Ajawan, V. Desai, S. Kale, and S. Patil, "Krishiq-BERT: A Few-Shot Setting BERT Model to Answer Agricultural-Related Questions in the Kannada Language," *J. Inst. Eng. India Ser. B*, vol. 105, no. 2, pp. 285–296, Apr. 2024, doi: 10.1007/s40031-023-00952-6.

[45] Terdalkar, Hrishikesh, and Arnab Bhattacharya. "Framework for question-answering in Sanskrit through automated construction of knowledge graphs." arXiv preprint arXiv:2310.07848 (2023).

[46] P. Gupta and V. Gupta, "Hybrid Approach for Punjabi Question Answering System," in *Advances in Signal Processing and Intelligent Recognition Systems*, vol. 264, S. M. Thampi, A. Gelbukh, and J. Mukhopadhyay, Eds., in Advances in Intelligent Systems and Computing, vol. 264. , Cham: Springer International Publishing, 2014, pp. 133–149. doi: 10.1007/978-3-319-04960-1_12.