Journal of Theoretical and Applied Information Technology
30th June 2024. Vol.102. No. 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HEALTHCARE QUESTION ANSWERING SYSTEM IN BENGALI – A PROPOSED MODEL

ARGHYADEEP SEN[1], DR. SATYA RANJAN DASH[2], DR. MANAS RANJAN PRADHAN[3], SHANTIPRIYA PARIDA[4]

[1]PhD Student, School of Computer Engineering, KIIT Deemed-to-be University, Odisha, India

[2]Associate Professor, School of Computer Applications, KIIT Deemed-to-be University, Odisha, India

[3]Associate Professor, School of Information Technology, Skyline University College, University City of Sharjah, Sharjah, UAE

[4]Silo AI, Finland

E-mail: [1]2081012@kiit.ac.in, [2]sdashfca@kiit.ac.in, [3]manas.pradhan@skylineuniversity.ac.ae, [4]shantipriya.parida@silo.ai

ABSTRACT

The rule-based Question Answering System was part of the manual input of rules and the development of rule-based models to implement an answering engine. However, manually formulated rules and knowledge of each language became the disadvantages of the rule-based approach, whereas current studies on low-resource languages lack expertise and suitable datasets in such languages. There are existing studies to obtain quantity and historical information from a given text; however, these studies were developed for educational purposes, and these models can respond only to simple questions. This study is focused mainly on obtaining healthcare related queries from patients' perspectives. The proposed model of the Question Answering System would be able to answer patients' queries from a set of prescriptions and medical reports. Moreover, this model can provide the necessary details about the patient to the medical personnel to understand the patient's condition without having to comb through a long medical history of reports. The Question Answering System would be able to resolve the issues with doctors' messy handwriting, recommend specific medicine and dosage to the patients, and help common patients understand their condition and provide support for their well-being.

Keywords: *Question Answering System, Low Resource, Health Care, Medical Question Answering Dataset*

1. INTRODUCTION

The question-answering system and implementation of QA systems are two of the oldest research fields in Computer Science. QA research works were initiated in the late 1960s and early 1970s, whereas recent works on QA development are based on free text, where the text findings are relevant to the query processed and the answer to the question returns a set of relevant documents [22]. The primary motivation of this study was to provide Bengali-speaking patients with a system that can provide relevant responses to understand prescriptions. Moreover, the problem lies not only in Bengali being a low-resource language but also in the significant lack of NLP studies in the Bengali medical domain. We have developed a benchmark dataset as a test dataset for multilingual tasks such as large language modeling, question answering, zero-shot learning, and others. We have manually developed a benchmark dataset with the help of a native Bengali speaker for Bengali responses to frequently asked questions from patients (table 1).

Table 1: Sample from Medical QA dataset in instruction format of question and answer

Instructions	Answers
"Bengali Question"	"কম ক্যালোরিযুক্ত খাবার কি ওজন কমাতে সাহায্য করে?"
"English Question"	"Does a low calorie diet help in losing weight?"
"Bengali Answer"	"হ্যাঁ। খাদ্য গ্রহণ কমিয়ে এবং নিয়মিত ব্যায়ামের মাধ্যমে ওজন কমানো যায়। একটি কম ক্যালোরি খাদ্য একটি কম কার্বোহাইড্রেট, উচ্চ ফাইবার, মাঝারি প্রোটিন এবং একটি কম চর্বিযুক্ত খাদ্য গঠন করা উচিত।"

"English Answer"	"Yes. Weight reduction can be achieved by reducing food intake and by regular exercise. A low calorie diet should constitute a low carbohydrate, high fiber, moderate protein and a low fat diet."

1.1 Background of QA System and Different Approaches Used

A QA system based on the retrieval of documents and information is known as an information retrieval-based QA system, whereas, in the previous two decades, the development of the semantic web made structured data available to the web, where there are several existing knowledge bases (KBs) [17]. Knowledge bases regarding publications, media, historic data, geography, and others QA developed in several different knowledge bases is identified as an important repositories for information retrieval [21]. Detailed information availability and significant knowledge gain are more relevant to the fact that QA systems developed over knowledge bases are useful for finding information that a particular user requested using natural language.

A natural language query or question is translated into another format to address the answer or relevant set of documents to the user [24]. Hence, QA systems developed over knowledge bases are referred to when information retrieval and relevant QA systems are identified. Several different types of QA systems have been developed in recent years, where development requires extensive work over in-depth analysis of the information. Initially, the rule-based approach is mainly chosen to develop the most potential QA systems. Grammatical semantics is used to derive specific rules relevant to a specific language, and these rules are used for selecting an appropriate answer for a question [10]. These rules are mainly heuristic and written by users depending on the context, lexical hints, and semantics of the language.

Predefined patterns in a language can be overlooked with rule definition, and these patterns could have been useful for question classification on behalf of answer type. However, these rules can be used in a decision tree so that an accurate path can be found to reach the correct answer to a question [6]. The primary disadvantage of a rule-based QA system was the need to develop rules manually. Rules definition would require detailed knowledge about the language, and as soon as textual content availability increased, the use of a statistical approach became more important than before [4]. The statistical approach can avoid the use of structured query languages, and this approach can work with heterogeneous data along with predicting correct answers from given data [5]. The statistical approach requires setting a knowledge base depending on a specific training dataset for the model; the model has self-learning capability for QA systems. The statistical approach provides freedom from the rule-bound approach; however, statistical approaches can also suffer from the bias of the particular training dataset.

Neural Networks in QA systems brought several possibilities into reality, as machine learning techniques can only process natural data in raw format. However, a machine learning system requires very specific engineering and detailed domain knowledge for designing a feature extraction process [7]. The feature extraction would transform raw data into an internal representation so that the classifier could detect patterns in the input dataset. Deep learning methods are useful for allowing the user to feed raw data and discover representations for detection, classification, and prediction. Deep learning techniques can transform the models into higher and more abstract modules so that complex functions can be learned [9] [14]. Classification tasks can be conducted with increasing aspects of inputs and support variations. Dynamic Memory Networks and reinforced memory networks can be used for providing results in the question-answering system so that artificial intelligence can be incorporated with the human perspective.

1.2 Basic QA System and Its Components

The Basic QA system includes common modules, each of which is adjourned by a core element named the Question Processing Module. The Question Processing Module requires question classification, and the Document Processing Module requires information retrieval so that the Answer Processing Module can extract the answers [20] [28]. The question Processing Module includes three interconnected parts known as the interface, analyzer, and classification of questions. The document retrieval module is used for retrieving the appropriate documents based on keywords for looking into the correct answers [22]. The solution Processing Module is used for identifying useful information as the answer to the question. The following diagram of proposed model 1 shows the parts of a QA system with individual aspects relevant to the QA system's operations.

1.3 Major Issues in Visual Question Answering Models and Datasets

Most state-of-the-art VQA approaches can be used to evaluate multiple-choice questions based on conventional accuracy measurement metrics. However, these approaches are not compatible with evaluating open-ended questions [2]. There is always an issue in the evaluation of multiple-choice questions, as the correct answer should be determined, reducing the problem instead of providing the answer to the question. There must be some sort of reasoning or logical explanation given to the system so that it can determine the answer from the choices rather than simply interpret the choices.

The DAQUAR[1] and VQA datasets are considered benchmark datasets for most of the VQA models. Existing models cannot meet the requirements of real-world applications such as providing necessary insights from the massive amount of data to a data analyst, supporting blind people, teaching children a particular concept using smart devices, and making communication with a robot [29]. Hence, there must be publicly available datasets such as PubMedQA [11] or VizWiz [8]. Moreover, existing datasets are highly biased, and the majority of the models are dependent on the models rather than the image content. There are image quality issues for low-quality images [32]. Existing datasets are a relatively small collection of data, and the creation of a massive dataset still requires enormous time and cost [16]. A large proportion of data is required to train the deep VQA models to obtain reasoning capabilities and learn concepts about the topic.

1.4 Major Applications of Visual Question Answering Systems

Visual Question Answering has several potential applications with different purposes. Visual Question Answering has a basic application in that it can help blind or visually challenged people identify objects. Visual content and language both can be used together for communication, and visually challenged people would have access to a language interface [30]. They cannot access visual content; here, the VQA interface would help them connect vision with language. VQA systems can be useful for having interaction with, organizing, or analyzing massive unstructured visual data [16]. For instance, a person may have a large collection of images and videos that they have collected over the past few years. These old images and videos can be used to relive memories and moments, and the VQA system can be used to achieve this kind of interaction [26]. VQA systems can be used as teaching modules for children to learn new concepts and activities. Most of the contents on the internet are multimodal, such as text, video, or images, and hence, a suitable system can be used to connect vision with language. In this case, the VQA system serves the very purpose of establishing a bridge between vision and language [33]. The VQA system can be used to make insightful summaries from a collection of several visual data points to support analysts [34]. These are major applications of VQA systems that are part of current research.

1.5 Aim and Objectives

The research in this paper focuses on a specific domain and uses a specific low-resource language, Bengali. Bengali is the seventh most widely spoken language in the world; several researchers are working on developing a Bengali language-based information retrieval, query-response system, and question-answering system. Bengali is also known as the most vernacular language among other low-resource Indic languages. Different states in India, such as West Bengal, Tripura, Assam, and the Andaman and Nicobar Islands, have used Bengali, and Bangladesh has Bengali as its national language. The majority of the citizens in the mentioned states and areas may not be able to understand or access e-commerce platforms due to language barriers, as most Indians speaking the vernacular language are not comfortable using English as their language medium.

The proposed system is mainly focused on processing medical-related queries in Bengali and providing responses or answers in Bengali. The domain is mainly based on queries frequently asked when a patient gets a prescription from a doctor or when medical personnel try to get relevant information from a set of medical documents or reports. The research rationale is based on the problem domain, where most of the patients are unaware of medical terms in English and instructions from the doctor are not clear to read when they are handed over the prescription. Hence, the proposed system should be able to provide answers regarding their medical conditions from the prescription, and it can obtain relevant information from the prescription in Bengali. On the other hand, new

[1] https://www.kaggle.com/datasets/dotran0101/daquar/data

medical personnel can read the patient records and find the relevant priorities from the prescription or medical records. Hence, transferring one patient to other medical personnel, where debriefing patient conditions would become easier than usual.

2. LITERATURE REVIEW

As the simplest form of Visual Question Answering Antol et al. [3], proposes the Vanilla VQA model, and this model can be a benchmark for performance evaluation of other VQA models. This VQA model includes VGG, which functions as a Convolutional Neural Network for recovering image features, and LSTM or GRU as a Recurrent Neural Network for recovering the linguistic embedding of the queries [12]. These features are then passed to the multi-layer perceptron classifier. The neutral-symbolic approach is introduced for understanding language based on reasoning and a logical base [31]. As a positive benefit, the Neural-symbolic approach for VQA can improve overall accuracy by reducing computational and memory costs. The NS-VQA model includes three components: a scene parser, a question parser, and a program executor. The model traverses the workflow from one sequence to another with the help of an encoder and decoder [1]. The LSTM encoder is used for the conversion of the questions in this NS-VQA approach and the CLEVR, CLEVR-CoGen, and CLEVR-Humans datasets, along with Minecraft, which uses a neural symbolic technique for VQA. Kafle and Kanan [13] opined that embedding strange and irrational questions with evaluation metrics can improve the VQA model with a detailed understanding of the questions. They included the TDIUC dataset with 12 different question types and proposed an evaluation metric for compensating for the bias that exists in the dataset. Marino et al. [19] obtained a benchmark for OK-VQA that can address the challenges of knowledge-based visual question answering. While comparing other knowledge-based visual question-answering datasets, this study mentioned that their knowledge-based VQA work is more diverse, huge, and difficult. In another study, they discussed the scenario that requires outside knowledge to answer questions that are not available within the dataset. They have stated this scenario as the most difficult form of VQA questions and knowledge gathering challenge [18]. In the OK-VQA model, they have shown that their Knowledge Reasoning with Implicit and Symbolic Representation (KRISP approach) can outperform existing state-of-the-art VQA models. The KRISP approach considers symbolic knowledge and implicit knowledge together successfully. Shah et al. [25] stated another form of the VQA model known as the KVQA model that enables VQA for named objects. KVQA includes 183,000 question-and-answer pairs with more than 18,000 named entities and 24,000 pictures.

The Literature Review section is prepared in a tabular format, mentioning relevant works in the Bengali Question Answering System (Table 3 in Appendix I). This table includes columns highlighting research objectives, methodology used, results obtained, and research gaps identified in the review. From this table, it can be found that Banik and Rahman [4] aimed to use Named Entity Recognition (NER) for answering questions using Bengali newspapers, and the model was limited in its performance as there was a lack of pre-trained word embedding. Simple sentences were parsed to extract Bengali answers; however, there is an issue with complex sentences [15]. Similarly, simple factoid sentences were parsed using a statistical approach, and descriptive questions were avoided [23]. Other approaches were limited to relatively smaller datasets in Bengali language [9, 27] and a lack of scoring metrics was evident in the results for baseline QA systems [7, 20], [5] had aimed to identify semantically relevant answers in the Bengali dataset, and they have used a shallow parser for POS tagging of the dataset. Their model had achieved 97.32% accuracy using the confusion matrix, and their system precision had reached 98.14% considered higher than other papers.

CliCR [27] is considered as a dataset that includes different domain specific gap-filling questions about several medical cases and the authors have identified that domain specific knowledge base is a successful aspect for development of medical question-answering system. CliCR dataset contains 105,000 question answer pair in medical domain and the answer type of this dataset is entity.

BioASQ [31] is an extraction dependent question answering dataset that is developed using biomedical corpus. This dataset is directly focused on biomedical questions such as “What are the physiological manifestations of disorder Y?” The dataset is relatively small with 287 instances of questions and answers in the domain of biomedical. The answer type for the dataset is span and subjective relevance where the medical articles are indexed to prepare the dataset.

PubMedQA [11] is a multi-choice biomedical Question-Answering dataset (around

1,000 question answer pair) that uses PubMed abstract. For individual question-answer pair, the question is generated from title or extracted from sentence from an article. The context is considered as the article's abstract excluding the conclusion from the abstract. The conclusion is considered as the long answer to the question as per the context; hence, the dataset depends on summarization of context to retrieve accurate answer from the conclusion. The deployed models would provide yes/no/maybe answer to state whether the conclusion can answer the question or not.

3. PROPOSED MODEL

The Proposed model includes three parts for the QA system's development; these parts are shown in figure 2 (Appendix I). The left side of the diagram shows a process flowchart in the format of a block diagram to give an idea of the working principles. There is a separate section in the diagram showing the development phases and major particularities of the QA system. This section discusses the process flowchart, mentioning the relevance of each segment and their working processes in the entire QA system.

Question Analysis Module: The first part of the proposed model includes the "Question Analysis Module", where question analysis would be performed when the prescriptions and medical reports are scanned. We aim to understand doctors' handwriting and identify relevant information from the prescription. So that patients can understand the handwriting of the prescription and they can find out the medicines and instructions by themselves [24]. Sometimes, doctors, even Bengali native speakers, prefer to write prescriptions in English. Patients sometimes cannot understand the exact meaning of the instructions due to the handwriting and also because of English terminologies; therefore, this system should be able to process prescriptions in English or Bengali and provide answers only in Bengali for patients' convenience and clear understanding. The proposed system would take inputs from prescriptions when the text and handwriting are collected through PyTesseract Optical Character Recognition (OCR). PyTesseract OCR would be useful in this part, where it would automatically analyze the printed or handwritten text to turn it into a machine-readable format. On the other hand, the proposed model would be essential for doctors or physicians, as it would be able to parse through medical reports and store them for patient health condition-related question-answering tasks [22]. In that field, too, OCR would be able to collect the relevant information about a patient so that a question from medical personnel could be answered from the system interface. Patient and medical staff queries should be stored as search queries; those queries should be tagged with Part-of-Speech tags, and only keywords such as nouns and proper nouns should be stored for query formulation.

Information Retrieval Module: The module would provide relevant information or answers to specific queries from patients or medical personnel. The module would have a prescription dataset and a medical dataset as repositories. The QA system should be trained with question-and answer (QA) training datasets. The prescription dataset would include detailed features about the disease, medications, instructions, dosage frequency, dosage details, and next visit information. Whereas the medical report dataset would have more detailed features such as patient health data, medical diagnosis, admission date, admission reasons, treatment duration, treatment progress, current clinical conditions, and consultant remarks. These details in the medical report dataset would be necessary for another physician to understand the current patient's condition and be able to transfer them to their treatment schedule and process. The QA system would be able to provide them with these details so that physicians would be aware of the patient's condition and medical personnel would be able to carry out their work fluently. Queries from patients or medical personnel can be carried out on the system, and answers will be collected or extracted from the provided datasets. With the help of deep learning techniques, the queries and relevant datasets would be converted into vectors for expressing the meaning of the queries and answers. In the embedding phase, the word from each text would be converted into Word Embedding, and this embedding format is a correct representation of the word in a set of vectors. A similar meaning based word would be converted into a similar representation of a vector. GloVe (Global Vector) Word Embedding would be used in the embedding phase. For Encoders, we need to prepare a representation of the dataset. CNN-based encoders and RNN (GRU/LSTM)-based encoders are to be used here. Output from the encoder would be a hidden vector in both forward and backward directions (Bidirectional Attention Flow).

Output Module: The output module should include a deep learning-based model where each phase is shown in figure 2 (Appendix I). The

output module would extract the answer from the QA system, and it would evaluate the answer based on accuracy scoring. The answer to the query would be displayed to the users. First step in the deep learning model would be processing the QA datasets. This step would require word embedding for encoders. The dataset would be encoded for the attention phase. The attention phase is necessary for finding the match between the hidden vector for a question and the hidden vector for the dataset, and the similarity matrix should be computed. Context Attention for each Question and Question Attention for each Context would be combined for predicted answers; the final answer would be displayed for users.

4. DATASET

There are very limited sources of datasets that are available for use in Medical Question Answering in Bengali. Therefore, we chose to prepare our own benchmark dataset. The dataset contains 42 categories of common and critical diseases and their frequently asked questions. The dataset is created based on the source[2] and this benchmark dataset is translated manually into Bengali. The dataset is stored in Github[3] for future references in Large Language modeling tasks. The dataset is in JSON format, and it includes five columns: "Categories", "Instruction (Question)", "English_Instruction (Question)", "Output (Answer)", and "English_Output (Answer)". The dataset includes 300 rows of Medical Question-Answer data, both in Bengali and English. The statistics of the dataset are shown as follows in table 2:

Table 2: Details of the Benchmark Medical Question Answer Dataset created for as testing material for Large

Column Name	Content
Categories	42 Different Disease Names
Instruction (Question)	300 FAQs in Bengali
English_Instruction (Question)	300 FAQs in English
Output (Answer)	300 Answers in Bengali
English_Output (Answer)	300 Answers in English

Here is a sample instruction set shown from the dataset. {"Categories": "Coronavirus FAQs", "Instruction (Question)": করোনাভাইরাস সংক্রমণের জন্য আমার কোন ডাক্তারের সাথে পরামর্শ করা উচিত?", "English_Instruction (Question)": "Which doctor should I consult for coronavirus infections?", "Output (Answer)": "আপনি যদি সর্দি, গলা ব্যথা, শ্বাসকষ্ট এবং জ্বর অনুভব করেন, তাহলে আপনার পারিবারিক চিকিত্সকের সাথে যোগাযোগ করা ভাল, যিনি উপসর্গগুলি মূল্যায়ন করবেন এবং আপনাকে উপযুক্ত চিকিৎসা সুবিধার কাছে পাঠাবেন যদি সেখানে করোনাভাইরাস সংক্রমণের সন্দেহ থাকে।", "English_Output (Answer)": "If you experience a runny nose, sore throat, respiratory difficulties, and fever, it is best to contact your family physician who will then assess the symptoms and refer you to the appropriate medical facility if there is a suspicion of coronavirus infection."}

This instruction set falls under "Coronavirus FAQs" category and the question is stated as which doctor a patient should consult for coronavirus infections. The answer is collected from mentioned source[4]. The Bengali questions and answers are translated manually with appointing native Bengali speaker as annotator.

Following figure 1 (Appendix I) shows the length of Bengali instruction text and length of answer text for the dataset. Most common question length is ranging between 5 (12%), 6 (almost 15%), 7 (11%), 8 (almost 8%), 9 (9%) and 10 (14%) as per bar chart. Box plot suggests median lies at nearly 7 for Bengali question length and question length ranges from minimum (2) to maximum (15). However, two outliers are at 18 and 21 question length. Similarly for the case of Bengali answer length, most common answer length is from 10 to 20 where, distribution of answer percent is diverse and it suggests that most of the answers are detailed and contains at least a sentence. Box plot for answers length suggests that median lies slightly upper than 20 and minimum length of Bengali answer is from 5 to 65 (approximately). However, there are so many outliers in answer length at 70 to max range of 120.

4.1 Human Evaluation of the Dataset

To assess the quality of the dataset and translation quality, we have manually evaluated the medical instructions sets with answers. As this dataset is first of its kind, there is very limited

[2] https://www.medindia.net/patients/diseasefaq.asp
[3] https://github.com/ArghyaDS/Sample_Med_QA_Pair/blob/main/MedicalQAInstructionSets.json
[4] https://www.medindia.net/patients/diseasefaq.asp

existence of Bengali Medical Question-answering dataset; we needed an expert opinion about the dataset quality. Hence, for this dataset the human evaluation was conducted to consider quality. Currently, the dataset size is relatively small and we have sought help from a medical professional who is a native Bengali speaker. He had helped us to conduct manual evaluation on the dataset to ensure:

1. translation quality of the instruction sets
2. accurate instructions in terms of queries
3. informative answers to the instructions

4.2 Experiment with Gemma 2B LLM

We conducted a small experiment of fine-tuning Google's Gemma 2B LLM[5] using our medical question-answering dataset. Google's Gemma series of open-source foundational large language model is developed recently and Gemma is available in two sizes 2B and 7B where the models are pre-trained with two billions and seven billions of instruction-sets respectively. Our medical question-answer dataset has instructions in Bengali and corresponding answers in Bengali. So, we chose lightweight Gemma 2B model and fine-tuned it with our dataset using LoRA[6] technique.

Another reason to choose Gemma 2B for fine-tuning is that this model is trained on 2 billion parameters and it is lightweight so that it can be fine-tuned in Google Colaboratory GPU runtime easily. LoRA technique is chosen as LoRA (Low-Rank Adaptation) technique for fine-tuning large language models (LLMs) can offer a parameter-efficient approach. LoRA technique does not require fine-tuning entire base model; that would require massive magnitude of computational resources, time and cost for training. The process of fine-tuning a LLM is shown in following figure 1. LoRA technique significantly reduces computational resource requirement for fine-tuning. This technique decomposes weight matrix of pre-trained language model into smaller matrices that can be combined to approximate the original matrix. Smaller matrices of LoRA adapters can be fine-tuned rather than fine-tuning the original weight matrix and these adapters can be multiplied to form a new matrix of same size of the original weight matrix.

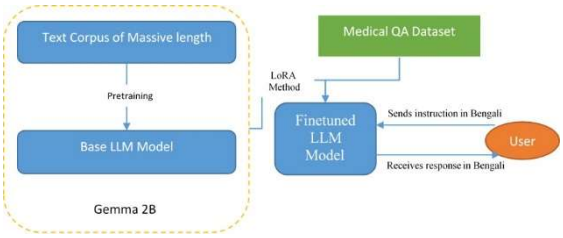

Figure 1: LLM fine-tuning Architecture using LoRA Method

4.3 Experiment Results

The dataset is preprocessed into instruction-response format so that the dataset can be used for fine-tuning the Gemma 2B LLM. Here is a sample row of pre-processed dataset format.

Instruction: "মস্তিষ্কের টিউমারের কারণ কী?" Response: "মস্তিষ্কের টিউমারের চূড়ান্ত কারণ জানা যায়নি তবে বিভিন্ন কারণ যেমন রেডিয়েশন এক্সপোজার এবং পলিভিনাইল ক্লোরাইড মস্তিষ্কের টিউমার হওয়ার সম্ভাবনা বাড়িয়ে তোলে।"

When the model is not fine-tuned, the LLM provided a wrong and confusing response as follows.

Instruction: "কনজাংটিভাইটিস কি?" Response: "কনজাংটিভাইটিস হলো একটি রোগ। কনজাটিইটিস হলো একটি রোগ। কজাংটিভিস হলো একটি রোগ কনজাংটিভাইটিস।"

Later, we have fine-tuned the LLM using rank-4 (low-rank matrices for LoRA method) for controlling approximation level of LoRA method, input sequence length was set at 512 tokens and the training was conducted for five epochs with batch size of five. When we have asked the same question, the LLM provided almost correct response as follows.

Instruction: "কনজাংটিভাইটিস কি?"

Response: "কনজাংটিভাইটিস হলো চোখের একটি অসুস্থতা।"

Our dataset includes the actual response as: "কনজাংটিভাইটিস হল কনজাংটিভা এর প্রদাহ, একটি পাতলা, স্বচ্ছ স্তর যা চোখের পাতা এবং চোখের সামনের

[5] https://blog.google/technology/developers/gemma-open-models/

[6] https://github.com/microsoft/LoRA

পৃষ্ঠকে আবৃত করে। এটি সব বয়সের মানুষকে প্রভাবিত করে।"

It can observed that fine-tuned model provided almost accurate response. This experiment shows that Gemma 2B model can be fine-tuned to learn Bengali conversations and the fine-tuning process was successful. However, the training of model was conducted at a low rank of LoRA and using a lower quantity of data. This experiment is conducted to understand the capability of the dataset so that it can be used for fine-tuning Large Language Models. In future, more volume of dataset would be collected and preprocessed in order to fine-tune and obtain perfect responses.

5. SURVEY RESULTS AND DISCUSSION

In this section, results, findings and limitations of the previous studies are discussed. Based on these outcomes, the research gaps are discussed in this area of 'question-answering system development for low-resource languages'.

Results from the survey suggests that there is significant research progress in the field of visual question answering system for English. Development of VQA is part of deep learning research discipline and continuous development of language corpus is helping the research towards advanced experiments. Advanced experiments such as large language models, training with instructions and zero-shot learning are becoming next research trends based on massive language corpus sources. Some successful VQA models are mentioned as follows,

Vanilla VQA model was stated as simplest benchmark model that can be useful for evaluation of other VQA models. It was developed using CNN for recovering image features along with LSTM or GRU as a Recurrent Neural Network for recovering the linguistic embedding of the queries. Language can be understood using reasoning and a logical base; hence, VQA model would require these two components as basic parts to understand any language.

Another model named as NS-VQA model that was developed using three components: a scene parser, a question parser, and a program executor. Here, encoder and decoder helped to follow the workflow from one sequence to another sequence.

A benchmark for OK-VQA is developed that can address the challenges of knowledge-based visual question answering. Working with knowledge-based questions or anomalies in irrational questions can be difficult to deal with and therefore, existing research works only focused on simple questions or questions that can be answered using knowledge within the dataset provided. When it comes to outside knowledge that should be used to answer a question; this scenario is known as a challenge in existing papers. KRISP approach was stated that can outperform existing state-of-the-art VQA models; as it considers symbolic knowledge and implicit knowledge together successfully.

Another successful VQA model was mentioned that is known as the KVQA model that enables VQA for named objects. KVQA model includes 183,000 question-and-answer pairs with more than 18,000 named entities and 24,000 pictures.

In the context of *limitations and research gaps* from the survey, a model was developed using a Bengali newspapers dataset, the model was limited in its performance as there was a lack of pre-trained word embedding. The dataset was limited so the experiment provided F1-score of 69%. In another paper, their proposed method achieved 60% accuracy compared to naïve approach. This model parses only simple sentences to extract Bengali answers; however, there is an issue with complex sentences. Hence, limited Bengali corpus and challenges in using complex sentences hindering the development of VQA model in this research context. Another successful model parses simple factoid sentences using a statistical approach, and descriptive questions were avoided. Another research work was conducted on QA system development for single or multiple document. The limitation was found that large dataset was used as input, the system might retrieve lesser relevant document. Some of the models lack in proper use scoring metrics for the Bengali QA system evaluation.

Based on literature survey, it is found that VQA model development for low-resource language is still an emerging area. The VQA model can involve the integration of both visual and textual information and the study shows that multimodal understanding is a necessity for progression in this domain. Combining images and text as form of input can be essentially translated into meaningful answers for the users. However, in medical domain

and in the context of low-resource language such as Bengali, the visual question answering system lacks in above-mentioned aspects.

6. CONCLUSION AND FUTURE WORKS

Our work has emphasized developing a question-answering system in Bengali, and the motivation of this study is to provide Bengali-speaking patients with a system that can provide relevant responses to understand prescriptions. The proposed model would be able to respond to general patients as well as medical personnel. Our medical QA system would be helpful for patients to understand medical information and doctors instructions in Bengali. Doctors or medical personnel can understand patients' conditions from summarized documents, and the QA system can detect common diseases in patients.

The medical VQA system faces some existing challenges in this research domain. Firstly, the system should be able to answer a range of comprehensive question categories. Secondly, the medical features should be combined with the task. Thirdly, there must be some sort of validation work to verify the evidence of an answer for the patients and the benefits of the medical VQA system should be improved over existing medical VQA systems. Therefore, real-world scenarios should be included in idea making experiments in the medical domain. Real-time conversations between a patient and medical personnel should be considered to collect the necessary questions for the medical VQA systems. The proposed system should be able to answer questions from general patients; hence, real-time conversation inputs would be useful for making the dataset more realistic. Evidence verification, bias over the answers, and incorporation of an external knowledge database would be essential to developing medical VQA, along with interactive answering assistants as well.

When we tried to develop our own dataset of Medical question-answer questions, we faced the challenges of accurate translation quality from English to Bengali. Therefore, we have employed manual translation of Medical Questions and Answers into Bengali by a native Bengali speaker. In the future, this medical benchmark dataset can be used as a zero-shot learning, large language model, and medical chat-bot test dataset.

REFERENCES:

[1] Agrawal A, Batra D, Parikh D. Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356. 2016 Jun 23.

[2] Agrawal A, Batra D, Parikh D, Kembhavi A. Don't just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 4971-4980).

[3] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 2425-2433).

[4] Banik N, Rahman MH. Gru based named entity recognition system for bangla online newspapers. In2018 International Conference on Innovation in Engineering and Technology (ICIET) 2018 Dec 27 (pp. 1-6). IEEE.

[5] Das A, Mandal J, Danial Z, Pal A, Saha D. A novel approach for automatic Bengali question answering system using semantic similarity analysis. International Journal of Speech Technology. 2020 Dec;23:873-84.

[6] Debnath A, Rajabi N, Alam FF, Anastasopoulos A. Towards more equitable question answering systems: How much more data do you need?. arXiv preprint arXiv:2105.14115. 2021 May 28.

[7] Faisal F, Keshava S, Anastasopoulos A. SD-QA: Spoken dialectal question answering for the real world. arXiv preprint arXiv:2109.12072. 2021 Sep 24.

[8] Gurari D, Li Q, Stangl AJ, Guo A, Lin C, Grauman K, Luo J, Bigham JP. Vizwiz grand challenge: Answering visual questions from blind people. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 3608-3617).

[9] Islam ST, Huda MN. Design and development of question answering system in bangla language from multiple documents. In2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) 2019 May 3 (pp. 1-4). IEEE.

[10] Jha BK, Akana CM, Anand R. Question answering system with indic multilingual-bert. In2021 5th International Conference on Computing Methodologies and Communication (ICCMC) 2021 Apr 8 (pp. 1631-1638). IEEE.

[11] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146. 2019 Sep 13.

[12] Johnson J, Hariharan B, Van Der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 2901-2910).

[13] Kafle K, Kanan C. An analysis of visual question answering algorithms. InProceedings of the IEEE international conference on computer vision 2017 (pp. 1965-1973).

[14] Keya M, Masum AK, Majumdar B, Hossain SA, Abujar S. Bengali question answering system using seq2seq learning based on general knowledge dataset. In2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2020 Jul 1 (pp. 1-6). IEEE.

[15] Khan S, Kubra KT, Nahid MM. Improving answer extraction for bangali q/a system using anaphora-cataphora resolution. In2018 International Conference on Innovation in Engineering and Technology (ICIET) 2018 Dec 27 (pp. 1-6). IEEE.

[16] Lin Z, Zhang D, Tao Q, Shi D, Haffari G, Wu Q, He M, Ge Z. Medical visual question answering: A survey. Artificial Intelligence in Medicine. 2023 Jun 8:102611.

[17] Bhaskar Majumdar and Mumenunnessa Khan. 2021. Bangla Question Answering Based Model Using Sequence to Sequence Learning. (2021).

[18] Marino K, Chen X, Parikh D, Gupta A, Rohrbach M. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021 (pp. 14111-14121).

[19] Marino K, Rastegari M, Farhadi A, Mottaghi R. Ok-vqa: A visual question answering benchmark requiring external knowledge. InProceedings of the IEEE/cvf conference on computer vision and pattern recognition 2019 (pp. 3195-3204).

[20] Masum AK, Abujar S, Akter S, Ria NJ, Hossain SA. Transformer based bengali chatbot using general knowledge dataset. In2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) 2021 Dec 13 (pp. 1235-1238). IEEE.

[21] Muller B, Soldaini L, Koncel-Kedziorski R, Lind E, Moschitti A. Cross-lingual open-domain question answering with answer sentence generation. arXiv preprint arXiv:2110.07150. 2021 Oct 14.

[22] Saha A, Noor MI, Fahim S, Sarker S, Badal F, Das S. An approach to extractive bangla question answering based on bert-bangla and bquad. In2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI) 2021 Jul 8 (pp. 1-6). IEEE.

[23] Sarker S, Monisha ST, Nahid MM. Bengali question answering system for factoid questions: A statistical approach. In2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 2019 Sep 27 (pp. 1-5). IEEE.

[24] Sen A, Parida S, Kotwal K, Panda S, Bojar O, Dash SR. Bengali visual genome: A multimodal dataset for machine translation and image captioning. InIntelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021) 2022 Feb 28 (pp. 63-70). Singapore: Springer Nature Singapore.

[25] Shah S, Mishra A, Yadati N, Talukdar PP. Kvqa: Knowledge-aware visual question answering. InProceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 8876-8884).

[26] Srivastava Y, Murali V, Dubey SR, Mukherjee S. Visual question answering using deep learning: A survey and performance analysis. InComputer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5 2021 (pp. 75-86). Springer Singapore.

[27] Šuster S, Daelemans W. CliCR: a dataset of clinical case reports for machine reading comprehension. arXiv preprint arXiv:1803.09720. 2018 Mar 26.

[28] Tahsin Mayeesha T, Md Sarwar A, Rahman RM. Deep learning based question answering system in Bengali. Journal of Information and Telecommunication. 2021 Apr 3;5(2):145-78.

[29] Teney D, Anderson P, He X, Van Den Hengel A. Tips and tricks for visual question answering: Learnings from the 2017 challenge. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 4223-4232).

[30] Teney D, Van den Hengel A. Visual question answering as a meta learning task. InProceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 219-235).

[31] Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y. An overview

of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics. 2015 Dec;16:1-28.

[32] Wu J, Mooney R. Self-critical reasoning for robust visual question answering. Advances in Neural Information Processing Systems. 2019;32.

[33] Yi K, Wu J, Gan C, Torralba A, Kohli P, Tenenbaum J. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. Advances in neural information processing systems. 2018;31.

[34] Zhang D, Cao R, Wu S. Information fusion in visual question answering: A survey. Information Fusion. 2019 Dec 1;52:268-80.

Journal of Theoretical and Applied Information Technology
30th June 2024. Vol.102. No. 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5 APPENDIX

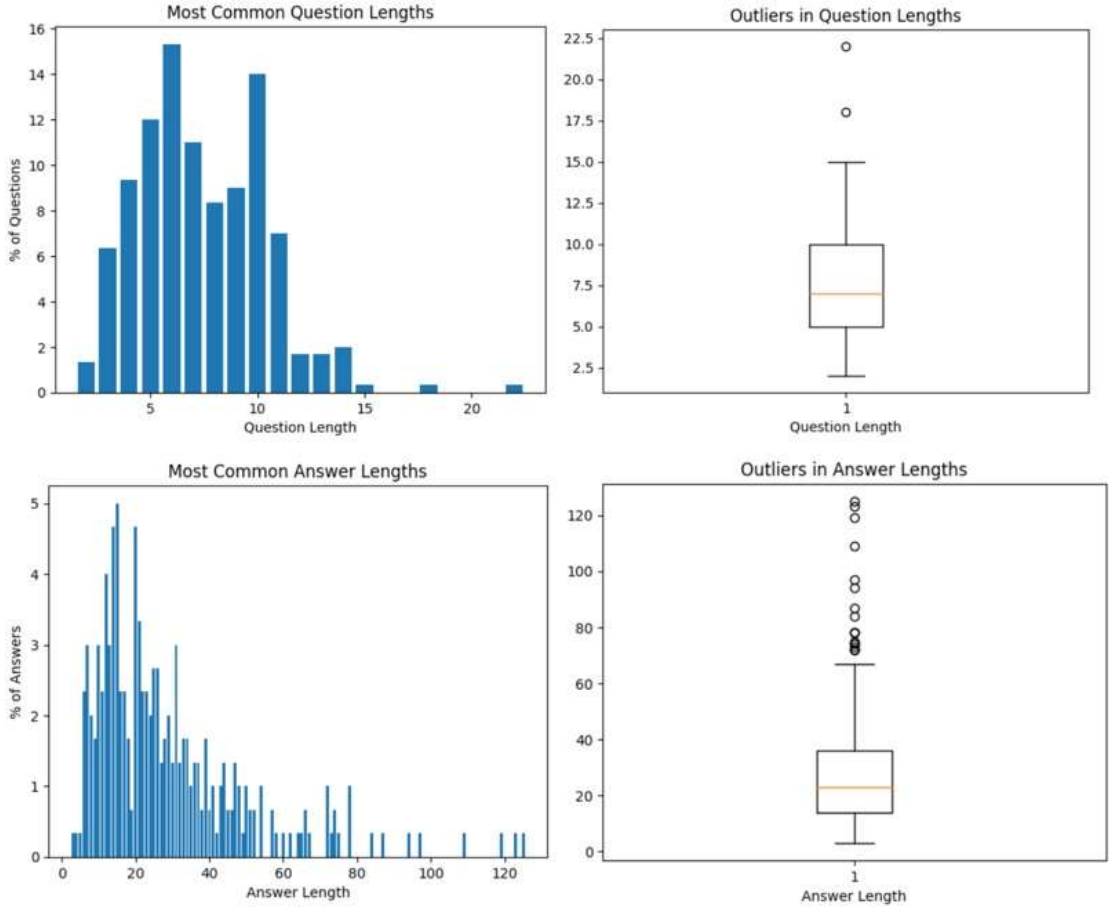

Fig. 1: Benchmark Medical Question Answer Dataset Analysis

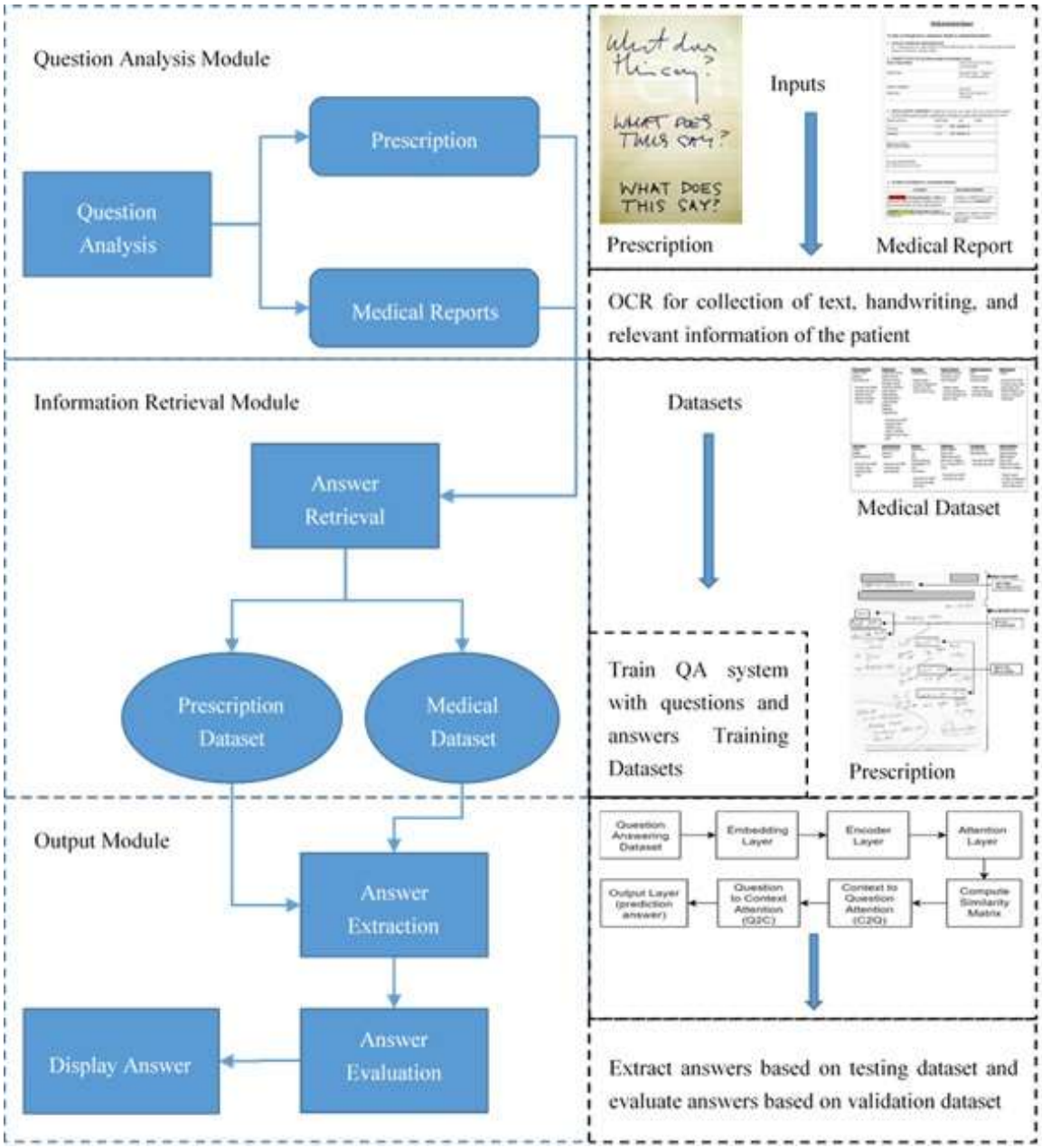

Fig. 2: Proposed model for Question Answering System. The left side of the diagram shows a process flowchart in the format of the block diagram to give an idea about the working principles

Table 3: Summary of literature gaps in Bengali QA System

Authors	Objectives	Methodology and Dataset used	Results Obtained	Research Gaps
Banik and Rahman, 2018 [4]	To use Named Entity Recognition (NER) for Bengali online newspapers for Question Answering	Gated Recurrent Unit (GRU) based model is used for development of Bengali NER task with manually annotated dataset	Based on this limited dataset, the experimental system provided F1-score of 69%.	Proposed model can perform better using more data with pre-trained word embedding.
Khan, Kubra and Nahid, 2018 [15]	To explore Anaphora Cataphora resolution to extract answers for Bengali QA system	Parsing simple sentences in the system and experimenting with both Bengali and English	Compared to naïve approach, proposed method achieved 60% accuracy	Works only with simple sentences and provides incorrect

		language along with systematic and semantic overview		answers to complex sentences.
Sarker, Monisha, and Nahid, 2019 [23]	To use a statistical approach to experiment with Bengali Question Answering System for factoid questions.	Document collection, mapped question and internet source; classify document with questions, identify minimum length answers, and exact answers for evaluation; used a corpus with 15,355 questions and 220 documents for the system	Proposed system achieved 66.2% accuracy with mentioning object name and 56.8% without mentioning the object name. Question classifier accuracy is 90.6% and document classifier accuracy is 75.3%	The system only works for simple factoid questions and in future, complex and descriptive questions should be considered
Islam, and Huda, 2019 [9]	Question Answering system in Bengali allowing answers from single or multiple documents	Identify keywords and retrieval of answers	Precision, recall, and F-score were 0.35, 0.65, and 0.45 respectively based on testing 500 questions	Using large dataset as input, the system might retrieve lesser relevant document
Keya et al. 2020 [14]	Develop and experiment with context-based QA system	Deep learning-based Seq2Seq model for Bengali context-based QA system with 2,000 general knowledge dataset	Achieved 99% accuracy from the model and 89% accuracy for validation	No scoring or evaluation criteria was introduced to evaluate model performance
Das et al. 2020 [5]	To identify semantically relevant answers in Bengali dataset	Part-of-speech (POS) to use index matching between a question and probable answers and similarity and entropy as metrics; Sense score for ranking relevant answers from 275,000 Bengali sentences	System accuracy was 97.32% and using confusion matrix, the system precision was 98.14%.	Shallow parser is used to conduct the POS tagging for the dataset
Faisal, Keshava, and Anastasopoulos, 2021 [7]	The researchers aimed to extend QA dataset for developing a multi-dialect QA benchmark in five different languages (Bengali, Arabic, Kiswahili, English, and Korean).	Authors have tried to introduce dialects into QA system, as there are wide varieties of Question Answering (QA) Systems with several commercial applications. However, existing QA system does not consider errors from speech recognition models or those QA systems can consider the language variations through dialects.	The dataset includes more than 68k audio data with 24 different dialects from 255 different speakers.	The paper provides a baseline showing how normal QA system performance can decrease with different dialects introduced to speech data.
Masum et al. 2021 [20]	The researchers have used Seq2Seq model for learning Question Answering system and transformer model is introduced for resolving sequence related dilemmas.	It reduced training time compared to RNN-based model and transformer model is used for Bengali general knowledge ChatBot on Bengali general knowledge QA dataset.	The experiment resulted into 85.0 BLEU score and in comparison, Seq2Seq model with attention to dataset resulted into 23.5 BLEU score.	Only BLEU score metrics used for the analysis; the researchers should explore metrics such as RIBES (Rank-based Intuitive Bilingual Evaluation Score) and ChrF (character n-gram F-score for automatic evaluation) for comparative results.
Tahsin Mayeesha, Md Sarwar, and Rahman, 2021 [28]	The researcher have surveyed different Question Answering systems for Bengali (non-English), English languages, and explored different deep learning methods to conduct the research.	They have considered BERT, DistillBERT, RoBERTa as three basic BERT model for exploration and performance comparison.	They have used SQUAD dataset and QA task and modified it with Bengali translation as well.	Zero-shot model is more convenient as per their findings and in future work, the researchers would like to explore more advanced language models.