A Diverse and Explainable Multi-hop QA Dataset for Bengali Language

Authors
Md. Aseer Intiser 180042128
Mohammad Munimul Islam 180042136
Md. Reyanus Salehin 180042148

Supervised By
Mohammad Anas Jawad, Lecturer, CSE
Dr. Abu Raihan Mostofa Kamal, Professor, CSE

Academic Year: 2021-2022
20 May 2023

A thesis report submitted to the Department of CSE in partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering

Department of Computer Science and Engineering
Islamic University of Technology

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by *Md. Aseer Intiser*, *Mohammad Munimul Islam* and *Md. Reyanus Salehin* under the supervision of *Mohammad Anas Jawad*, Lecturer of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh and *Dr. Abu Raihan Mostofa Kamal*, Professor of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. We are very grateful to our supervisors *Mohammad Anas Jawad*, Lecturer of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), and *Dr. Abu Raihan Mostofa Kamal*, Professor of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), for their supervision, knowledge, and support, which has been invaluable to us.

It is also declared that neither this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Md. Aseer Intiser

Md. Aseer Intiser
Student ID: 1800421428

Mohammad Munimul Islam
Student ID: 180042136

Md. Reyanus Salehin

Md. Reyanus Salehin
Student ID: 180042148

Supervisors:

Mohammad Anas Jawad
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology

Dr. Abu Raihan Mostofa Kamal
Professor
Department of Computer Science and Engineering
Islamic University of Technology

Acknowledgement

We would like to express our gratitude towards the IUT authority for providing the assistance required in making our proposed dataset. We are indebted to our supervisors, Mohammad Anas Jawad and Dr. Abu Raihan Mostofa Kamal for providing us with insightful knowledge and guiding us at every stage of our journey, Finally, we would like to express our heartiest appreciation towards our family members for their continuous support, motivation, suggestions, and help, without which we could not have achieved this progress in our work.

Special thanks to Dr. Md. Azam Hossain, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology and Md. Mezbaur Rahman, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology for their help throughout our research

Abstract

Bengali is a resource-scare language with a scarcity of quality data sets both in single and multi-hp question answering. In an approach to fill that gap, we want to take a little step by generating a reading comprehension-based open-domain multi-hop question answering which will be explainable and diverse. We will generate about 100 passages from news and Wikipedia articles and 500 question-answer pairs. We will maintain the diversity in selecting domains of contexts and also in generating questions and answers. Our data set will be explainable in generating the answer to a given question by providing supporting facts and showing the reasoning chain.

Keywords: QA, QA in Bengali, RC based QA in Bengali, Open-domain QA, Open-domain QA in Bengali, Multi-hop QA, Multi-hop QA in Bengali, Reasoning Chain

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Question Answering (QA) is a sub-domain of Natural Language Processing (NLP) that deals with the generation of questions or answers in human language (natural language). As mentioned by authors in [1] Question Answering (QA) aims to provide precise answers in response to the user's questions in natural language. In this paper, we will focus on the reading comprehension (RC) based answer generation part of QA. RC involves reading a text, understanding its meaning, and then being able to answer questions about it. In [2] authors have mentioned the main challenges for machines regarding RC-based answer generation are to understand the natural language and knowledge about the world.

Natural language poses a challenge as humans can ask questions in various ways and it varies from person to person and even if it is for the same question. The machines have to understand the texts, the questions, and some other knowledge to generate the answers. The texts, questions, and answers are written or generated in natural language. In this study we have focused to generate question-answer pairs (QA Pairs) solely from the texts, in this case, the machines have to understand the questions and the texts to generate the answers. The machine, here Artificial Intelligence (AI) model, has to be trained on texts, questions, and related answers as initially they are not capable of understanding the natural language. This training will make the model understand the texts, questions, and answers in its own way and also lead to forming reasonings over the texts to generate answers in natural language.

Based on the scope of texts, the RC-based QA is divided into two sections. The first section is called Close Domain QA. Here texts are generated based on one particular domain like medicine, law, history, etc. This domain can even be more specific like a certain period of history, certain or several certain fields of medical studies, specific sections of law (like laws regarding land, legal rights, criminal offenses, etc.), etc. The other section called Open-domain QA involves generating text from multiple domains like sports, laws, medicine, history, fictional and non-fictional stories, general knowledge, news, etc. Like the scope of texts, the RC-based QA is also divided into two sections based on question generation. The first section is Single-hop QA. Single-hop QA involves the generation of questions in natural language from a single sentence or a context (interchangeably used with the words "paragraph" and "passage"). The other section called Multi-hop QA involves generating questions from multiple sentences or multiple contexts. The RC-based QA is also divided into three sections based on answer generation. The first section is Generative Answering. The model will generate the answer in the natural language of the given question based on the reasoning or reasonings over the context. The second one is called Extractive Answering. This involves indicating the sentence or sentences, or part or parts of the context or contexts which includes the answer to the given question. Extractive answering also needs reasoning to indicate the answer. The third section is Selective Answering, it selects the answer options given the question and options for answers like MCQ.

As QA involves many divisions and sections if we specify our study it will be, we want to create an RC-based open-domain multi-hop QA dataset which will produce generative answering. This type of study has already been done in English, but there is no such dataset has been built in Bengali Language. This will be the first-ever attempt to create a multi-hop QA dataset in Bengali.

1.1 Difference between English and Bengali Language

As RC-based open-domain multi-hop QA datasets already exist in the English Language, it can be a question why we need to build a dataset in the Bengali Language. Because Bengali and English are two distinct languages with different grammatical structures and semantic systems. The syntactic and semantic differences between

Bengali and English have been discussed below.

Word order. Bengali is a subject-object-verb (SOV) language, which means that the subject comes first, followed by the object, and then the verb. In contrast, English is a subject-verb-object (SVO) language, which means that the subject comes first, followed by the verb, and then the object.

Verb forms. Bengali has a rich verb system with different verb forms for different tenses and moods. In contrast, English has a relatively simpler verb system with fewer inflections for tense and mood.

Case marking. Bengali is a highly inflected language with different cases for nouns, pronouns, and adjectives. In contrast, English has lost most of its case inflections over time.

Pronouns. Bengali has different pronouns for different levels of formality and respect, while English has a relatively simpler pronoun system.

Vocabulary. Bengali and English have different vocabularies, with Bengali having many words borrowed from Sanskrit and other languages, and English having many words borrowed from Latin and Greek.

Semantic categories. Bengali and English have different semantic categories, with Bengali having more words for social relationships, family members, and food, while English has more words for scientific and technical concepts.

Overall, Bengali and English are two distinct languages with different grammatical structures and semantic systems. Learning Bengali requires a deep understanding of its syntax and semantics.

1.2 Motivation

Based on the following problem statements we have got our motivations for this work.

From the study of the existing datasets of the Bengali language, we can see that all the available datasets are based on single-hop. The quality of the datasets is also not up to the mark. In some of the cases, the size of the datasets doesn't match the mentioned in the paper. In some cases, we have seen the context is missing for some contexts. The biggest dataset is a translation from English datasets. As translation is done through AI models it decreases its quality as the dataset generated in Bengali will include more diversity and include the naturalness of the human. In most cases, the supporting facts have not been indicated and also the reasoning chains have not been shown for generating answers. The datasets also don't contain information about the diversity of contexts, questions, and answers.

To progress in Question Answering in Bengali Language there is a huge need for quality datasets from close domain to open domain, single-hop to multi-hop. To make the AI models understand the natural language, there is no alternative to quality datasets. The availability of very few datasets (less than 10), the lack of quality in them, and their small sizes indicate that we need quality datasets if we want to progress in this field. Single and multi-hop in close-domain and in open-domain, generative and extractive answering in these variations, etc., indicates the variations of QA and the datasets that will be needed for them. This a field that is in its infant stage for Bengali Language and needs to go a long way.

As there are many works to be done we have selected this specific task of generating an open-domain multi-hop dataset based on reading comprehension which will contain generative answering.

Time restriction is the biggest challenge for us as most of the work has to be done manually due to a lack of tools in the Bengali language. Most of the correction and cleaning of the dataset has to be done manually. Making sure diversity of the contexts, questions, and answers and also open domains of the contexts is going to be tough. Continuous documentation is also not an easy job, including the necessary things and leaving the unnecessary ones. At times, it is tough to determine which ones are unnecessary.

1.3 Problem Statements

- Lack of robust datasets for low-resource languages like Bengali.
  - Lack of datasets.
  - The number of available datasets is less than 10
- Existing QA datasets are designed to answer questions over a single paragraph or document as the context, thus failing to test a system's ability to answer complex questions spanning multiple contexts.
- Systems trained on existing single-hop datasets lack supervision of the generated answers, thereby lacking explainability.
- Existing datasets lack diversity.
- Supporting fact and reasoning chains are absent in the datasets.
- There is no multi-hop dataset in the Bengali Language.

1.4 Thesis Objective

Establish a diverse and explainable QA dataset that requires reasoning over multiple contexts while providing supporting facts to enhance a model's explainability.

Diverse means the contexts will be generated from multiple domains (sports, news, history, significant events or incidents, law articles, general knowledge, etc.), rich diversity in question and answer types (having multiple types of QA pairs: WH questions, boolean questions, answers including different parts of speech), requiring different types of reasoning to answer the question. Explainable indicates the fact that the QA Pairs generated from multiple contexts will have supporting facts for the answers from the contexts, and the models will be able to build a reasoning chain from these facts to generate the answers.

Chapter 2

Background

2.1 Dataset Comparison

For English, we have discussed the multi-hop datasets only. As there is no multi-hop dataset for Bengali, we have discussed the available datasets that we have found so far.

2.1.1 English Datasets

In the paper [3], the authors have given a comprehensive guideline to generate a multi-hop dataset. This is our main paper and we have mainly followed it to generate our dataset. For some Wikipedia articles, they have made a graph, an article as a node, and each of the other articles as another node whose links have been contained in that article. Each of the links is considered as an edge. Thus they have developed multi-hop contexts. Thus, their database is open-domain. They have generally used the whole of Wikipedia to generate the contexts, generating 1,12,779 QA Pairs based on 1, 2, or 3 contexts. Their answering method is extractive.

In [4] the authors have used tables, which contain links to articles and keywords. Thus a table links between different keywords and articles and passages. It also contains passages from Wikipedia. They have used a total of 13k tables and 293k passages. The database is open-domain and they have generated 69,611 QA Pairs based on 2 or 3 passages from 1 table. The extractive answering method is also used here.

In [5] authors have generated a close-domain dataset based on fiction. They summarized a story and then generated QA pairs from the summary. They used 783 books and 789 movies to generate retrieve the stories and generated a total of 46,765 QA Pairs. As they have given the story and related QA Pairs to the model. The summary is condensed from multiple sentences or passages of the story which makes the dataset multi-hop. They have used the generative answering method.

In [6] authors have generated an open-domain dataset retrieving contexts from multiple sources which include children's story books, elementary science books, articles on history, anthropology, society, law and justice, 9/11 reports, and news. Their QA Pairs are MCQ type consisting of one question and multiple answer options related to the question. They have generated a question based on multiple sentences of one passage. In total, they have generated 9,872 QA Pairs from 871 passages. As QA Paris is MCQ, their answer selection type is selective.

Dataset	Total Size of Context	Context Granularity	Number of QA Pairs	Number of Hops	Domain	Answer Type
HotpotQA	|Wikipedia|	Passage	1,12,779	1/2/3	Open	Extractive
HybridQA	Passages:293k Tables:13k	Table, Passage	69,611	2/3	Open	Extractive
NarrativeQA	Books:783 Movies:789	Sentence	46,765	-	Close	Generative
MultiRC	Passages: 871	Sentence	9872	2.37	Open	Selective

Table 2.1: Comparison of different QA datasets in the English language

2.1.2 Bengali Datasets

In [7] authors have used a Bengali-translated version of the English dataset SQuAD 2.0. The SQuAD 2.0 dataset was introduced by the authors in [8]. It is an extension of the SQuAD 1.1 dataset with 50k more questions which is not answerable from the contexts. The authors have stated they have translated the whole dataset into Bengali. It is a single-hop dataset with 100k QA Pairs and 50k questions and the source of the Contexts is Wikipedia. And they have used extractive answering.

In [9] authors have collected their contexts from newspapers, books, and sonnets. This is a single-hop dataset with 1,676 paragraphs and 8,027 questions. They have tried to bring some variety in reasoning to reach the answer. As the contexts are not focused on a single domain, the dataset is open domain and they have produced generative answering. This dataset is better in quality than other ones.

In [10] authors have collected their contexts from the internet. They have not mentioned the domains or what type of writings they have collected, just mentioned the phrase "famous Bengali writings". Their dataset is single-hop and contains 3636 QA Pairs. Though they have vaguely mentioned 3636 RC with QA Pairs from the available dataset we have seen the QA Pairs is 3636 and one context is used for multiple QA Pairs. They have used generative answering.

In [11] authors have built a big dataset of 27.5 GB but most of the part of the dataset is for Natural Language Inference (NLI) as they have built a pre-trained model of Bangla. Their QA dataset is a translation of SQuAD 2.0 and Typologically Diverse Question Answering (TyDi QA) (a dataset introduced by the authors in [12]) datasets with a total of 354K QA Pairs of 150k in SQuAD 2.0 and 204K in TyDi QA. It has been translated from English. It is a single-hop dataset with generative answering. The dataset is not up to the mark as passages are missing for some QA Pairs, and not cleaned and organized properly.

None of these Bengali Papers have specifically focused on preparing Bengali Datasets. They have prepared the datasets to evaluate the performance of their models. In most of cases, these have resulted in very poor datasets.

Dataset	Translation	Total Size of Context	Context Granularity	Number of QA Pairs	Domain	Answer Type
Deep Learning Based	SQuAD 2.0	|Wikipedia|	Sentence	150k	Open	Generative
Factoid QA	No	Passages: 1,676	Sentence	8,027	Open	Generative
RC Based QA in Bengali	No	-	Sentence	3,636	Open	Generative
banglaBert QA Dataset	SQuAD 2.0 TyDi QA	|Wikipedia|	Sentence	354k	Open	Generative

Table 2.2: Comparison of available QA datasets in the Bengali language

2.2 Literature Review

2.2.1 SQuAD

The SQuAD dataset has two major versions SQuAD 1.1 introduced by the authors in [2] and SQuAD 2.0 introduced by the authors in [8]. This is a single-hop dataset and it is a pioneer for the later datasets for QA.

2.2.1.1 Their Goals

Based on the needs in 2016, they have tried to build a diverse single-hop dataset. They have curated their passages from Wikipedia articles of various domains. The answer types also vary from date, other Numeric, person, location, other Entity, common noun phrase, adjective phrase, verb phrase, clause, and other. They have in total curated 1,07,785 QA Pairs from 23,215 paragraphs of 536 articles. In SQuAD 2.0 they further included 50k more questions that are not answerable from the given paragraphs as the model should also know about the cases where the questions are not answerable.

2.2.1.2 Methodology

The dataset generation is done through 03 stages:

1. **Paragrahs Curation.** In this stage they have curated the paragraphs following the below steps:
   (a) Top 10,000 articles of English Wikipedia using PageRanks of Project Nayuki
   (b) 536 Articles were randomly selected
   (c) Extracted 23,215 individual paragraphs after stripping away images, figures, and tables, discarding paragraphs shorter than 500 characters
   (d) Randomly divided into training set (80%), development set (10%), test set (10%)
2. **QA Pair Collection.** In this stage they have collected the QA Pairs through crowd-sourcing by following the restrictions below:
   (a) Up to 5 QAs per paragraph
   (b) Asking Questions in their own words has been encouraged
   (c) Copy-Pasting has been discouraged
   (d) Highlight the answers in the paragraph
   (e) Not much diversity in reasoning in answering the questions
   (f) Ques
3. **Additional Answer Collection.** They have collected 02 additional answers for questions of test and dev sets from crow-workers and also asked them to mark unanswerable questions which was 2.6%.

2.2.1.3 Dataset Analysis

They analyzed the diversity of answer types mentioned in the "Their Goals" section above and calculated the percentage of answers in each type. They analyzed the difficulty of the questions in terms of reasoning and syntactic divergence between question and answer sentences of a QA Pair.

2.2.1.4 Result

For SQuAD 1.1, they have 86.8% accuracy for humans and the best 51% for the logistic regression model in terms of models. The score is measured in the F1 metric proposed in the paper. Later in [13] authors got 70.3% of the F1 score using their model in SQuAD 1.1 dataset.

2.2.1.5 Limitations

1. Single-hop Dataset
2. Reasoning Chain has not been clearly proposed in the dataset
3. Different sources could have brought more diversity to the paragraphs, questions, and answers.
4. They have not categorized the paragraphs, questions, and answers in terms of domains.
5. They have not studied the diversity of question structure. Diversity in question structure is required to make the models more understanding of the natural language.

2.2.2 The NarrativeQA Reading Comprehension Challenge

In this article [14], authors argued that existing multihop datasets perform multihop reasoning using just pattern matching or span selection or converting the multihop questions into a sequence of single hop questions and retrieving each missing information at a time. Existing datasets fail to inter the underline narrative, complex relations, and timelines. For this reason, They presented a dataset that contains questions with stories and movie scripts. Their dataset contains long self-contained stories and movie scripts where complex relations between entities, and complex timeline is present.

2.2.2.1 Proposed Solution: NarrativeQA Dataset.

NarrativeQA dataset 1567 stories and movie scripts were used as context. Book stories were collected from Project Gutenberg and movie scripts were collected from the web. The dataset contains a small collection of long contexts compared to other existing datasets where a big collection of small contexts was used. The advantage of this approach is that long contexts have diversity and deep relations, and timelines.

2.2.2.2 Dataset Creation Process.

NarrativeQA contains 46,765 QA pairs. It was implemented in the following steps:

1. **Context and Summary Curation.** They chose stories and movie scripts as context. Stories were chosen from Project Gutenberg. Movie scripts were collected through scraping form the web. Because of the limitation of available human written summaries, the collection size is not big. Each Summary was collected from Wikipedia using titles and then was verified if it actually summarizes the corresponding story or movie script by human annotators.
2. **QA Pair Collection.** In this stage they have collected the QA Pairs through crowd-sourcing on Amazon Mechanical Turk by following the restrictions below:
   (a) Annotators were given only the summary of the stories and movie scripts to avoid localized questions
   (b) Annotators were asked to write questions for testing students who have read the full story or movie script but not the summary.
   (c) Copy-Pasting had been prohibited
   (d) annotators were asked for answer for each question
   (e) extra, unnecessary information in the question were discouraged

3. **Explainability Check.** They further validate each question by asking annotators if a given question is answerable or not. Only 2.3% questions were marked as unanswerable.

2.2.2.3 Dataset Analysis.

The dataset contains 1567 contexts evenly split between stories and movie scripts along with 46,765 question-answer pairs. Almost 30.54% of questions are about persons, 24.50% are about the description of an entity and 9.73% are about locations. Most questions require reading several paragraphs of the context in order to find the answer. Moreover, questions from movie scripts require an iterative reasoning process to understand the dialogues.

2.2.2.4 Result

They achieved 10.48/10.75 in BLEU-1, 3.02/3.34 in BLEU-4, and 0.1760/0.171 in MRR. This dataset sets a new benchmark in the multi-hop QA domain. It overcomes the limitations of previous research works by incorporating long contexts and deep abstract questions where actual multi-hopness is ensured.

2.2.2.5 Limitations

We have found the following limitations with this dataset:

1. Dataset size is small
2. Supporting facts retrieval is missing

2.2.3 WebQA: Multihop and Multimodal QA

In this article [15], authors criticized existing open-domain QA datasets for considering the only text as the source of reasoning. But naturally, when we, humans try to search for something on the web, we retrieve information from both texts and images. Images contain many vital information. For Example, when searching to see if the color of a building is red or not, surfacing an image of that building is sufficient to answer the question rather than searching for a document or a text span where anyone happens to mention the color of that building. This is why, multi-hop multi-domain QA datasets are inefficient because they don't consider images as an important source of information in a document. Multi-modality is more natural because web search is a multi-modal experience for humans. They argue that there should be no discrimination between text and image.

2.2.3.1 Proposed Solution: WebQA Dataset

In order to overcome the mentioned limitations, they presented a multi-modal dataset named WebQA in [15]. It is a multi-hop, multi-modal, and open domain in nature. This dataset consists of both images and text. Each image is tagged with a description mentioning names or geographical or timeline information if it is not present in the image itself. The answers in this dataset are full-form sentences which are useful for voice assistants and conversational agents. The dataset also requires a model to retrieve supporting facts from the source. This dataset sets a new benchmark by using multi-modal in an open domain setting.

Dataset Creation Process WebQA was implemented in the following manners:

1. **Context Selection.** They have chosen both texts and images as context. But they have no intersection. No questions need both an image and an independent text snippet for answering questions. But for image-based questions, both images and their description is needed for answering the questions. Images and text snippets were collected this way:
   (a) **Image Source.** Images were collected from Wikimedia Commons using Bing Visual Search API. Wikimedia category list was filtered and only the categories labeled as interesting were selected. Categories like animals, plants, attractions, and architecture were removed.
   (b) **Text Source.** To make the dataset diverse, they constructed clusters of similar entities. A total of 8k clusters were created. Text snippets had low semantic overlap.
2. **QA pair generation** Rich multi-image questions are very rare to find in user search logs because users normally do not search for complex multi-image questions which they believe search engines can not answer properly. That is why, they moved towards crowdsourcing. Annotators were given a set of six related images and they produced three QA pairs by selecting one or two images that are necessary to answer the questions. For image distractors, they selected images with a high lexical overlap of the descriptions.
3. **Quality Control.** For producing good quality QA pairs, they followed these approaches:
   (a) Annotators were trained with a video tutorial and selected through a qualification task.
   (b) Annotation task was released batch by batch and checked quality after each batch
   (c) Sent feedback to the annotators
   (d) rewarded bonus for out-of-the-box questions

2.2.3.2 Dataset Analysis

This dataset has a total of 34k training QA pairs, 5k development pairs, and 7.5k testing pairs. 44% of image-based queries and 99% of text-based queries are multi-hop questions. Because of the open domain nature, questions are very diverse and a variety of questions are available eg. Yes/No, W/H, color, shape, numbers, properties, do/does, and is/are the major types of questions.

2.2.3.3 Limitations

1. Only text source or image source is used to answer a particular questions

2.2.4 MultiRC

In [6] authors tried to solve the reading comprehension (RC) problem for MCQ generating reasoning over multiple sentences. Given a context and a question, the question needs reasoning over more than one sentence of the context to answer.

Their main solution is to build a dataset. They have tried to make the dataset as unbiased as possible. In most of the previous cases for reading comprehension, the databases have been biased like assuming exactly one answer is possible, focusing on only one field or understanding level, size being too small, a missing indication of sentences supporting the answer of a question, being made for single-hop only, not verifying for multi-hop conditions, presenting continuous sub-strings of a paragraph as an answer, etc.

2.2.4.1 Proposed Solution

They have made a diverse dataset from contexts of seven different domains (News; Wikipedia Articles; Articles on society, law, and justice; Articles on history and anthropology; Elementary school science book; 9/11 reports; Fiction: Gutenberg project stories, children's stories, movie plots) for MCQ type questions which required reasoning over more than one sentences (two to four sentences) assuming the correct answer or answers will not be limited to the verbatim in a paragraph and can be one or more than one (from one option is correct to all options are correct) keeping the number of options a variable for each question.

2.2.4.2 Accepted Principles

They have accepted four principles before generating the dataset to make sure the dataset achieves the required goals they are targeting for.

2.2.4.3 Multi-sentenceness

A question must be answered from reasoning over multiple sentences, any question will be excluded if it can be answered from only one sentence.

2.2.4.4 Open-endedness

The answer to a question is not limited to the verbatim in a paragraph. Some answers need to be inferred from more than one sentence of a paragraph.

2.2.4.5 Answer to be judged independently

The total number of options, correct options, and incorrect options is variable. It's not possible to guess correct answers by a process of elimination or by choosing the best option.

2.2.4.6 Variability

Paragraphs have been taken from multiple domains, leading to linguistically diverse questions and answers. No restrictions are imposed on generating the question to generate different forms of reasoning.

2.2.4.7 QA Pair Generation

They have followed the following steps from passage curation to QA Pair generation:

1. Passage Curation
2. Generating Question from Passages using crow-workers
3. Verifying Multi-sentences of the Questions
4. Generating Answer-options for the Questions
5. Verifying Quality of Dataset

2.2.4.8 Result

Human has scored with 84.3% and the best model has scored with 66.7% accuracy.

2.2.4.9 Limitations and Future Work

1. The performance of AI models is not so good on dataset compared to Human performance, so there is a scope for improvement.
2. The dataset is explicitly made for Multiple Choice Questions where one or more answer options are selected, but the answers do not need to be generated.
3. The dataset is about generating QA with reasoning over multiple sentences of the same paragraph, but not for QA over multiple paragraphs.
4. The QA pair in the dataset doesn't require reasoning over more than 4 sentences. So, it is not applicable for more generation complex reasoning over more than 4 sentences that are contained in single or multiple paragraphs.

2.2.5 Dataset Design for Multi-hop Reasoning

In [16] authors have studied the WikiHop a multi-hop dataset that uses MCQ (selective answering) as the answer options and HotpotQA a multi-hop dataset that uses span or extractive answering as answer options. The authors tried to find out how MCQ or span as an answer affects multi-hop reasoning.

2.2.5.1 Datasets

As mentioned earlier they have used the WikiHop and HotpotQA datasets.

2.2.5.2 Methodology

They have used different models to evaluate both of the models. Their main models were simple factored models and factored BiDAF, in both cases they pair the questions with a sentence of the context and try to find the probability to find the answer. In each case, they use a different formula to calculate the probability. They have also conducted a study on no context baseline by combining question and answer and trying to measure the score for each question. They also used two more models MemNet and BiDAF++ and compare scores achieved in different models.

2.2.5.3 Results and Findings

They have found almost half of the QA-Pairs of the HotPotQA don't require multi-hop reasoning and the MCQ-based dataset is more vulnerable to multi-hop reasoning than span based. The single-hop dataset SQuAD score higher as it is single-hop. The below findings have been gained:

1. Models do not learn multi-hop reasoning over multi-choice datasets.
2. For multi-choice datasets adding more options doesn't qualitatively change the setting.
3. Span-based data is less vulnerable, but models still may not be doing multi-hop reasoning
4. Span-based training is more powerful

The table shows the result of predicting multi-hop QA using the single-hop model:

Method	Factored	Factored BiDAF
WikiHop	60.9	66.1
HotpotQA	45.4	57.2
SQuAD	70.0	88.0

Table 2.3: Score for datasets using single-hop models

2.2.5.4 Limitations

1. More datasets should have been studied.
2. It's not clear whether the models are using the knowledge for one question-context or question-answer pair for another or not.
3. There are different types of reasonings that exist, study for how different reasoning is affecting dataset can be done.
4. Different type of multi-hop dataset exists, only two types have been studied.

2.2.6 Factoid QA System in Bengali

In [9] authors have proposed a factoid system for the Bengali Language. Building a factoid question-answering system in the Bengali language with a dataset and a model.

2.2.6.1 Dataset Curation

The authors have made the dataset themselves by reading the newspaper, books, and sonnets accessible via the internet. They first curated 1,676 passages and from there 8,027 QA Pairs are generated. The dataset is generated in .xlxs format first and then it has been converted into .json format. The maximum size of their paragraph and questions are respectively 1585 words and 36 words. And the length of the answers varies from single word to multiple word.

2.2.6.2 Dataset Analysis

They have shown there are 03 types of relationships between the questions and the answers:

- **Lexical Match** Maximum questions and the text span have a lexical match between them
- **Same Answer** Multiple questions have same answer
- **Multiple Answer** A question has multiple answers for a given text span

They analyzed the answer types and get three variations: answers in number, some answers need the previous sentence to answer, and some answers need some words to be inserted or to be deleted from the text span for being accurate.

2.2.6.3 Methodology

They have used the architecture for the model shown in figure 2.1.

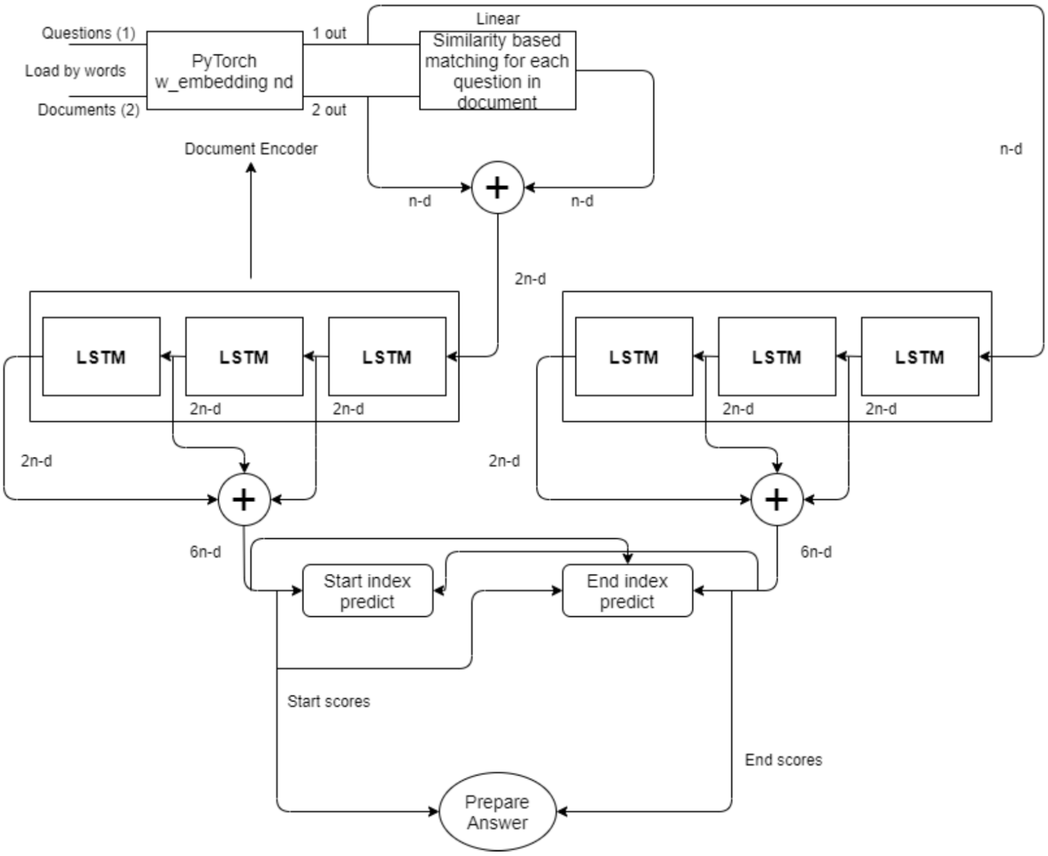


Figure 2.1: Bengali Factoid QA System

2.2.6.4 Result

They have achieved an F1 score of 92.16% for partial match and 76.8% for exact match utilizing LSTM. They have not shown any score for the human baseline.

2.2.6.5 Limitations

1. A bigger dataset should have been used.
2. The question structure has not been studied.
3. The diversity in answer types is only 03. This is too low.
4. Supporting factors and reasoning chains have not been shown in the database.
5. The F1 score shown for the given dataset is too good to be true, which indicates the overfitting of the model.

2.2.7 Bengali QA System based on GK Dataset

In [17] authors have tried to make a dataset based on general knowledge (GK) and proposed a model to build a QA system in Bengali based on the GK dataset.

2.2.7.1 Dataset

They have not given any link to the dataset or have not done any analysis on the dataset either. From the paper what we came to realize was they have collected a total of 2k QA Pairs in Excel format. And used the model proposed model on the dataset.

2.2.7.2 Methodology

They have used the model shown in figure 2.2.

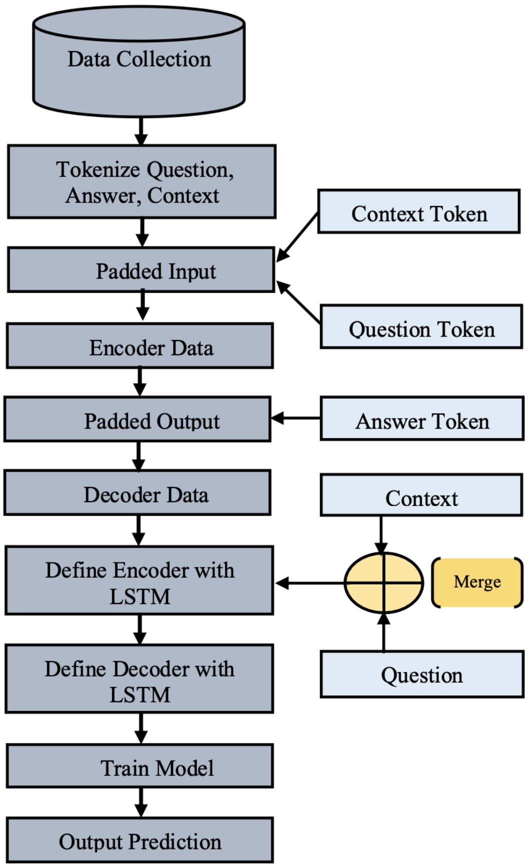


Figure 2.2: Model for Bengali GK-based QA System

2.2.7.3 Result

They have shown a 99% accuracy for the training set and 89% accuracy for the testing set. No human accuracy has been shown.

2.2.7.4 Limitations

1. No availability of the dataset.
2. The dataset is too small.
3. No diversity and explainability have not been described for the dataset.

4. No analysis of the dataset has been shown.
5. The result indicates the possibility of overfitting.

2.2.8 HybridQA

In [18] authors presented a large-scale multi-hop dataset for answering questions about heterogeneous data in both organized tabular and unstructured text formats. It does reasoning over heterogeneous data. Each question has numerous free-form corpora connected to the entities in the table as well as an alignment with a Wikipedia table. Because the questions are designed to combine both text and tabular data, their absence would make them impossible to answer.
Each crowd worker is given a table containing hyperlinked passages from Wikipedia during annotation so they can suggest queries requiring multi-hop reasoning over both types of knowledge.

2.2.8.1 Findings

- It is a high-quality dataset. They used Mechanical Turk to gather questions from the crowd workers. They also ensured strict quality control.
- It is a large-scale dataset that has 13k Wikipedia Tables, 293K hyperlinked passages, and 70K natural questions.
- It is a hybrid. It provides semantic understanding and symbolic reasoning.
- It does reasoning over open domain Wikitables.

2.2.8.2 Table/Passage Curation for Dataset

- Tables with 5-20 rows, 3-6 columns
- They limit the percentage of hyperlinked cells in the tables to no more than 35% of all the cells.
- We access the Wikipedia page for each hyperlink in the table and, for each, crop the first 12 sentences of the introduction.
- Thus collected 13,000 high-quality tables

2.2.8.3 Question/Answer Collection

- They released 13K HITs (human intelligence tasks) on the Amazon Mechanical Turk platform
- Each HIT gives the crowd worker a single Wikipedia table that has been crawled, together with all of its hyperlinked portions.
- They asked the employee to jot down six questions and their responses.
- They provided several examples in their Amazon Turker interface along with in-depth justifications to assist crowd-workers in comprehending "hybrid" questions.

2.2.8.4 Annotation De-biasing

- Table Bias
  - Questions about the top of the table are more frequently asked by annotators.
- Passage Bias
  - Asking questions on the passage's opening few phrases.
- Question Bias
  - Asking hybrid questions.

2.2.8.5 Data Analysis

Inference Types -

- Table → Passage chain (23.4%)
- Table → Passage chain (23.4%)
- Passage → Table → Passage chain (35.1%)
- Two parallel reasoning chain (3.1%)
- Two parallel reasoning chain (3.1%)
- Multiple reasoning chains (0.8%)

2.2.8.6 Result

Their estimated accuracy was greater than both SQuAD and HotpotQA at EM=88.2 and F1=93.5. The In-Table questions (almost 40%), which have far less ambiguity than the text-span questions, are accountable for better accuracy.

2.2.8.7 Future Scope

The experimental findings demonstrate that although the hybrid model may attain an EM of over 40%, the EM scores produced by the two baselines are below 20%. This gap indicates the need for HybridQA to aggregate heterogeneous data. The hybrid model's performance, meanwhile, is still far inferior to that of humans. As a result, HybridQA can act as a testing ground for research into question answering with heterogeneous information.

2.2.9 TAT QA

The goal of the TAT-QA (Tabular And Textual dataset for Question Answering) as mentioned by the authors in [19] is to advance QA research over increasingly challenging and realistic tabular and textual data, particularly those involving numerical reasoning.

2.2.9.1 Findings

- The presented context is hybrid and includes a semi-structured table.
- Humans with extensive financial understanding in practical fields are the ones that create the questions.
- Different answer formats, such as single span, multiple spans, and free-form, are available.
- Numerous numerical reasoning skills, such as addition (+), subtraction (-), multiplication (x), division (/), counting, comparing, sorting, and their compositions, are typically needed to solve problems.
- TAT-QA has a total of 16,552 questions linked to 2,757 hybrid contexts taken from actual financial reports.

2.2.9.2 Data Collection and Preprocessing

- They first downloaded about 500 financial reports
- Used table detection model in to detect tables
- Extracted the table contents
- Only kept tables with 3 - 30 rows and 3 - 6 columns

2.2.9.3 Dataset Annotation

- A valid hybrid context
- A table and at least two associated paragraphs
- First check whether there are ≥ 2 paragraphs around this table, and then check whether they are relevant. If yes, then this table will be linked to all the paragraphs nearby.

2.2.9.4 Question-Answer Pair Creation

- The annotators are then asked to create question-answer pairs
- Given one hybrid context, at least 6 questions are generated

2.2.9.5 Results

Different models on the dev and test set of TAT-QA performed differently, but the best among them was TagOp. On dev, the EM score was 55.2, and the F1 score was 62.7. And on Test, the EM score was 50.1, and the F1 score was 58.0. It is an absolute gain of 11:1% over the previous best baseline model.

2.2.9.6 Shortcomings

Their experiments on TAT-QA show that TAGOP achieves 58.0% in F1, which is an absolute gain of 11.1% over the previous best baseline model. However, this outcome still falls well short of human expert performance, which was 90.8% in F1. So there is a chance of improvement.

Chapter 3

Methodology

The methodology primarily can be divided into two parts. In the first part, the Dataset is generated which we are calling Dataset Generation. And in the second part, we will evaluate the dataset using a Bengali Question Answering model which we are calling Dataset Evaluation.

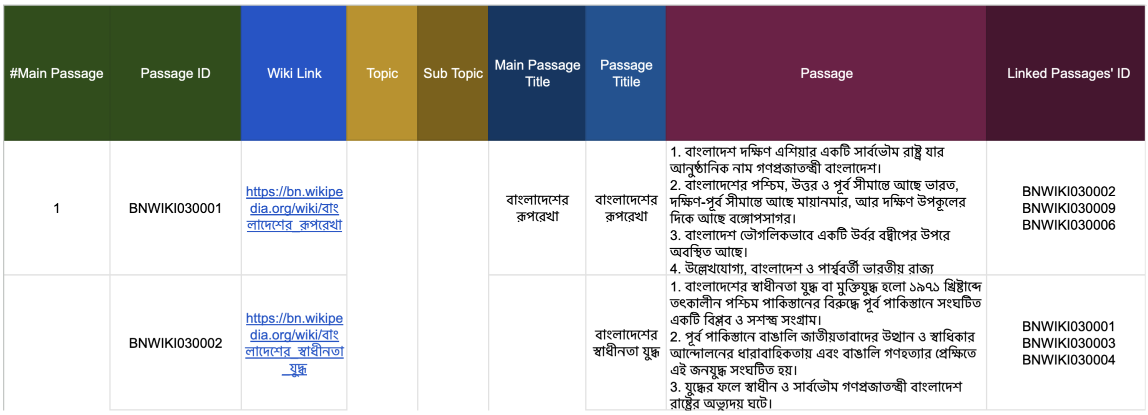
3.1 Dataset Generation

We have divided the process of Dataset Generation into four steps. We have studied several papers and almost all of them follow the same steps. But from them, we have mainly generated a process which is a combination of the processes the authors mentioned in the papers: [3], [2], and [6]. The four steps of our dataset generation process are Passage Curation, QA Pair Generation, Multi-hopness Verification, and Quality Verification.

3.1.1 Passage Curation

Our Passage Curation involves the generation of passages from Bengali Wikipedia. We have followed the mentioned steps and constraints for passage or context generation:

1. We have first selected an article from a topic of the Bengali Wikipedia. Generally, we have taken the first paragraph from the article as our passage.
2. We have stripped away images, figures, and tables from the paragraph if it contains any. In many cases, the paragraphs are very short and in some cases, we had to merge the first 2-4 paragraphs to make the passage.
3. For selecting the article we have made sure a passage can be made from the article and it contains at least one link to another article and QA generation is possible from the selected paragraphs of these articles.
4. The first passage which contains the link is called the main passage. And the passage generated from the linked article is called a linked passage. We have generated at least one and at most three linked passages for a main passage.
5. We have used Google Sheets to store the generated passages.
6. In each passage, we have numbered the sentences so that we can easily indicate which sentences of the passage are used to generate a question and these sentences can also be used to answer the question.



#Main Passage	Passage ID	Wiki Link	Topic	Sub Topic	Main Passage Title	Passage Titile	Passage	Linked Passages' ID
1	BNWIKI030001	https://bn.wikipedia.org/wiki/বাংলাদেশের_রূপরেখা			বাংলাদেশের রূপরেখা	বাংলাদেশের রূপরেখা	1. বাংলাদেশ দক্ষিণ এশিয়ার একটি সার্বভৌম রাষ্ট্র যার আনুষ্ঠানিক নাম গণপ্রজাতন্ত্রী বাংলাদেশ। 2. বাংলাদেশের পশ্চিম, উত্তর ও পূর্ব সীমান্তে আছে ভারত, দক্ষিণ-পূর্ব সীমান্তে আছে মায়ানমার, আর দক্ষিণ উপকূলের দিকে আছে বঙ্গোপসাগর। 3. বাংলাদেশ ভৌগলিকভাবে একটি উর্বর বদ্বীপের উপরে অবস্থিত আছে। 4. উল্লেখযোগ্য, বাংলাদেশ ও পার্শ্ববর্তী ভারতীয় রাজ্য	BNWIKI030002 BNWIKI030009 BNWIKI030006
	BNWIKI030002	https://bn.wikipedia.org/wiki/বাংলাদেশের_স্বাধীনতা_যুদ্ধ				বাংলাদেশের স্বাধীনতা যুদ্ধ	1. বাংলাদেশের স্বাধীনতা যুদ্ধ বা মুক্তিযুদ্ধ হলো ১৯৭১ খ্রিষ্টাব্দে তৎকালীন পশ্চিম পাকিস্তানের বিরুদ্ধে পূর্ব পাকিস্তানে সংঘটিত একটি বিপ্লব ও সশস্ত্র সংগ্রাম। 2. পূর্ব পাকিস্তানে বাঙালি জাতীয়তাবাদের উত্থান ও স্বাধিকার আন্দোলনের ধারাবাহিকতায় এবং বাঙালি গণহত্যার প্রেক্ষিতে এই জনযুদ্ধ সংঘটিত হয়। 3. যুদ্ধের ফলে স্বাধীন ও সার্বভৌম গণপ্রজাতন্ত্রী বাংলাদেশ রাষ্ট্রের অভ্যুদয় ঘটে।	BNWIKI030001 BNWIKI030003 BNWIKI030004

Figure 3.1: Excel Sheet of storing Generated Passages

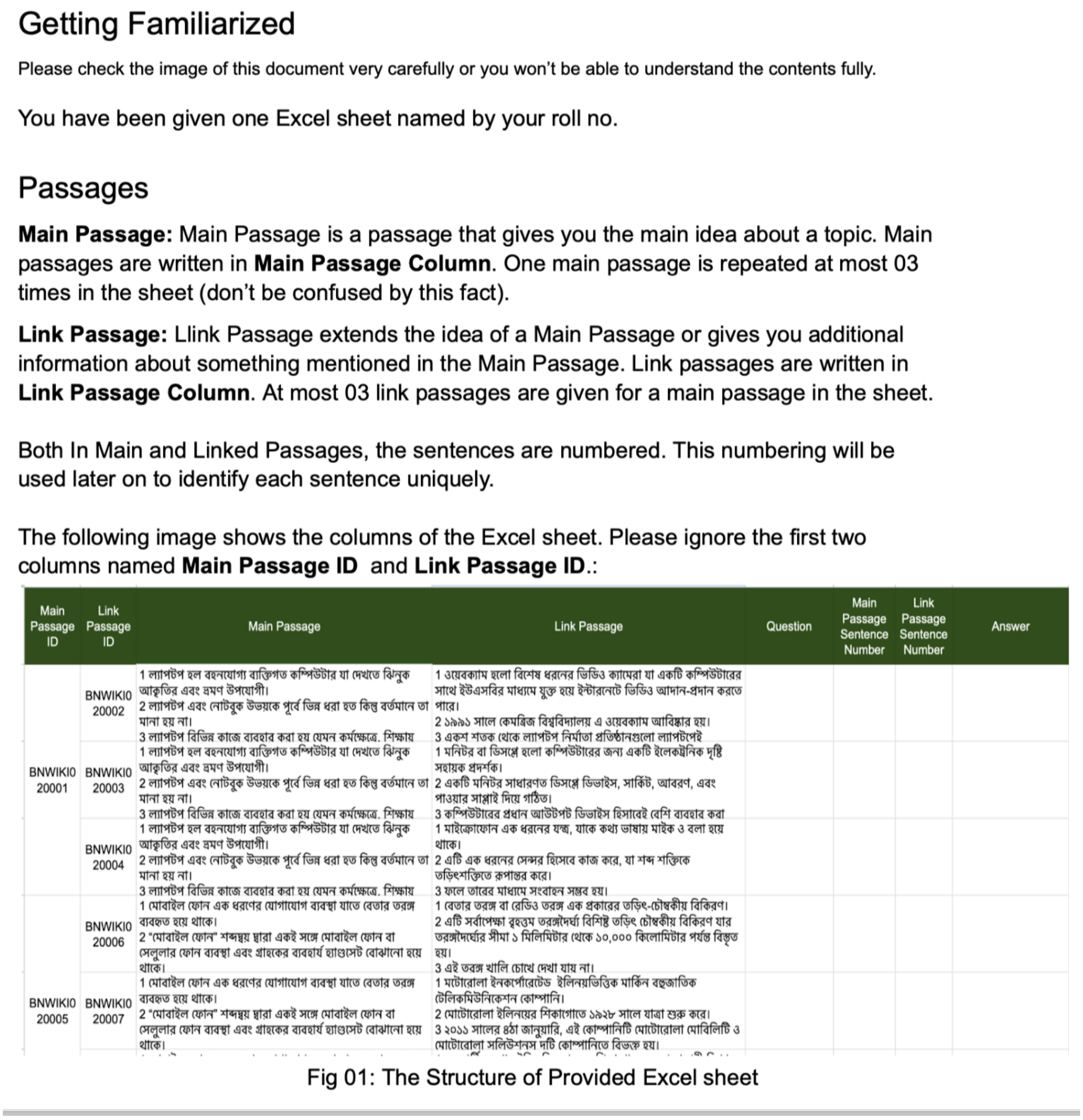
3.1.2 QA Pair Generation

After generating the passages QA pairs along with their supporting sentences are generated through crowd workers.

Crowd Worker

60 undergraduate engineering students from a prestigious institution and 15 people from different strata have volunteered for generating QA Pairs. We have used Google Sheets for assigning contexts and collecting the QA Pairs. We have given a specific Google sheet to each of the crowd-workers for QA Pair generation. Each volunteer has been assigned 15 main-linked Passage Pairs to generate QA pairs and their supporting sentences.

Crowd Worker Guidelines



Getting Familiarized

Please check the image of this document very carefully or you won't be able to understand the contents fully.

You have been given one Excel sheet named by your roll no.

Passages

Main Passage: Main Passage is a passage that gives you the main idea about a topic. Main passages are written in **Main Passage Column**. One main passage is repeated at most 03 times in the sheet (don't be confused by this fact).

Link Passage: Llink Passage extends the idea of a Main Passage or gives you additional information about something mentioned in the Main Passage. Link passages are written in **Link Passage Column**. At most 03 link passages are given for a main passage in the sheet.

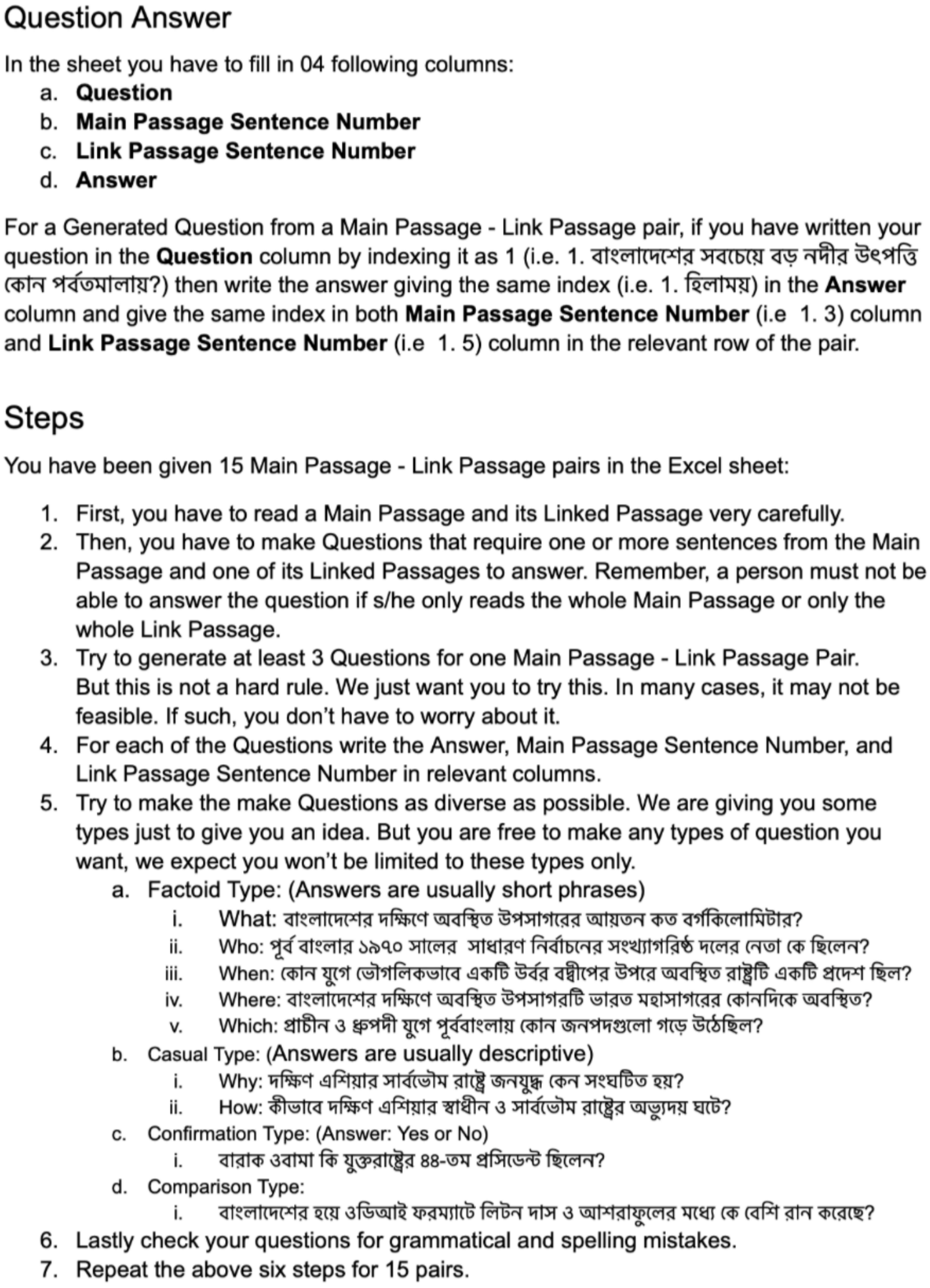
Both In Main and Linked Passages, the sentences are numbered. This numbering will be used later on to identify each sentence uniquely.

The following image shows the columns of the Excel sheet. Please ignore the first two columns named **Main Passage ID** and **Link Passage ID**.:

Main Passage ID	Link Passage ID	Main Passage	Link Passage	Question	Main Passage Sentence Number	Link Passage Sentence Number	Answer
BNWIKI0 20001	BNWIKI0 20002	1 ল্যাপটপ হল বহনযোগ্য ব্যক্তিগত কম্পিউটার যা দেখতে ঝিনুক আকৃতির এবং ভ্রমণ উপযোগী। 2 ল্যাপটপ এবং নোটবুক উভয়কে পূর্বে ভিন্ন ধরা হত কিন্তু বর্তমানে তা মানা হয় না। 3 ল্যাপটপ বিভিন্ন কাজে ব্যবহার করা হয় যেমন কর্মক্ষেত্রে, শিক্ষায়	1 ওয়েবক্যাম হলো বিশেষ ধরনের ভিডিও ক্যামেরা যা একটি কম্পিউটারের সাথে ইউএসবির মাধ্যমে যুক্ত হয়ে ইন্টারনেটে ভিডিও আদান-প্রদান করতে পারে। 2 ১৯৯১ সালে কেমব্রিজ বিশ্ববিদ্যালয় এ ওয়েবক্যাম আবিষ্কার হয়। 3 একশ শতক থেকে ল্যাপটপ নির্মাতা প্রতিষ্ঠানগুলো ল্যাপটপেই				
	BNWIKI0 20003	1 ল্যাপটপ হল বহনযোগ্য ব্যক্তিগত কম্পিউটার যা দেখতে ঝিনুক আকৃতির এবং ভ্রমণ উপযোগী। 2 ল্যাপটপ এবং নোটবুক উভয়কে পূর্বে ভিন্ন ধরা হত কিন্তু বর্তমানে তা মানা হয় না। 3 ল্যাপটপ বিভিন্ন কাজে ব্যবহার করা হয় যেমন কর্মক্ষেত্রে, শিক্ষায়	1 মনিটর বা ডিসপ্লে হলো কম্পিউটারের জন্য একটি ইলেকট্রনিক দৃষ্টি সহায়ক প্রদর্শক। 2 একটি মনিটর সাধারণত ডিসপ্লে ডিভাইস, সার্কিট, আবরণ, এবং পাওয়ার সাপ্লাই দিয়ে গঠিত। 3 কম্পিউটারের প্রধান আউটপট ডিভাইস হিসাবেই বেশি ব্যবহার করা				
	BNWIKI0 20004	1 ল্যাপটপ হল বহনযোগ্য ব্যক্তিগত কম্পিউটার যা দেখতে ঝিনুক আকৃতির এবং ভ্রমণ উপযোগী। 2 ল্যাপটপ এবং নোটবুক উভয়কে পূর্বে ভিন্ন ধরা হত কিন্তু বর্তমানে তা মানা হয় না। 3 ল্যাপটপ বিভিন্ন কাজে ব্যবহার করা হয় যেমন কর্মক্ষেত্রে, শিক্ষায়	1 মাইক্রোফোন এক ধরনের যন্ত্র, যাকে কথ্য ভাষায় মাইক ও বলা হয়ে থাকে। 2 এটি এক ধরনের সেন্সর হিসেবে কাজ করে, যা শব্দ শক্তিকে তড়িৎশক্তিতে রূপান্তর করে। 3 ফলে তারের মাধ্যমে সংবাহন সম্ভব হয়।				
BNWIKI0 20005	BNWIKI0 20006	1 মোবাইল ফোন এক ধরণের যোগাযোগ ব্যবস্থা যাতে বেতার তরঙ্গ ব্যবহৃত হয়ে থাকে। 2 "মোবাইল ফোন" শব্দদ্বয় দ্বারা একই সঙ্গে মোবাইল ফোন বা সেলুলার ফোন ব্যবস্থা এবং গ্রাহকের ব্যবহার্য হ্যাণ্ডসেট বোঝানো হয়ে থাকে।	1 বেতার তরঙ্গ বা রেডিও তরঙ্গ এক প্রকারের তড়িৎ-চৌম্বকীয় বিকিরণ। 2 এটি সর্বাপেক্ষা বৃহত্তম তরঙ্গদৈর্ঘ্য বিশিষ্ট তড়িৎ চৌম্বকীয় বিকিরণ যার তরঙ্গদৈর্ঘ্যের সীমা ১ মিলিমিটার থেকে ১০,০০০ কিলোমিটার পর্যন্ত বিস্তৃত হয়। 3 এই তরঙ্গ খালি চোখে দেখা যায় না।				
	BNWIKI0 20007	1 মোবাইল ফোন এক ধরণের যোগাযোগ ব্যবস্থা যাতে বেতার তরঙ্গ ব্যবহৃত হয়ে থাকে। 2 "মোবাইল ফোন" শব্দদ্বয় দ্বারা একই সঙ্গে মোবাইল ফোন বা সেলুলার ফোন ব্যবস্থা এবং গ্রাহকের ব্যবহার্য হ্যাণ্ডসেট বোঝানো হয়ে থাকে।	1 মটোরোলা ইনকর্পোরেটেড ইলিনয়ভিত্তিক মার্কিন বহুজাতিক টেলিকমিউনিকেশন কোম্পানি। 2 মোটোরোলা ইলিনয়ের শিকাগোতে ১৯২৮ সালে যাত্রা শুরু করে। 3 ২০১১ সালের ৪ঠা জানুয়ারি, এই কোম্পানিটি মোটোরোলা মোবিলিটি ও মোটোরোলা সলিউশনস দুটি কোম্পানিতে বিভক্ত হয়।				

Fig 01: The Structure of Provided Excel sheet

Figure 3.2: Crowd Worker Guideline Page 01



Question Answer

In the sheet you have to fill in 04 following columns:

a. **Question**
b. **Main Passage Sentence Number**
c. **Link Passage Sentence Number**
d. **Answer**

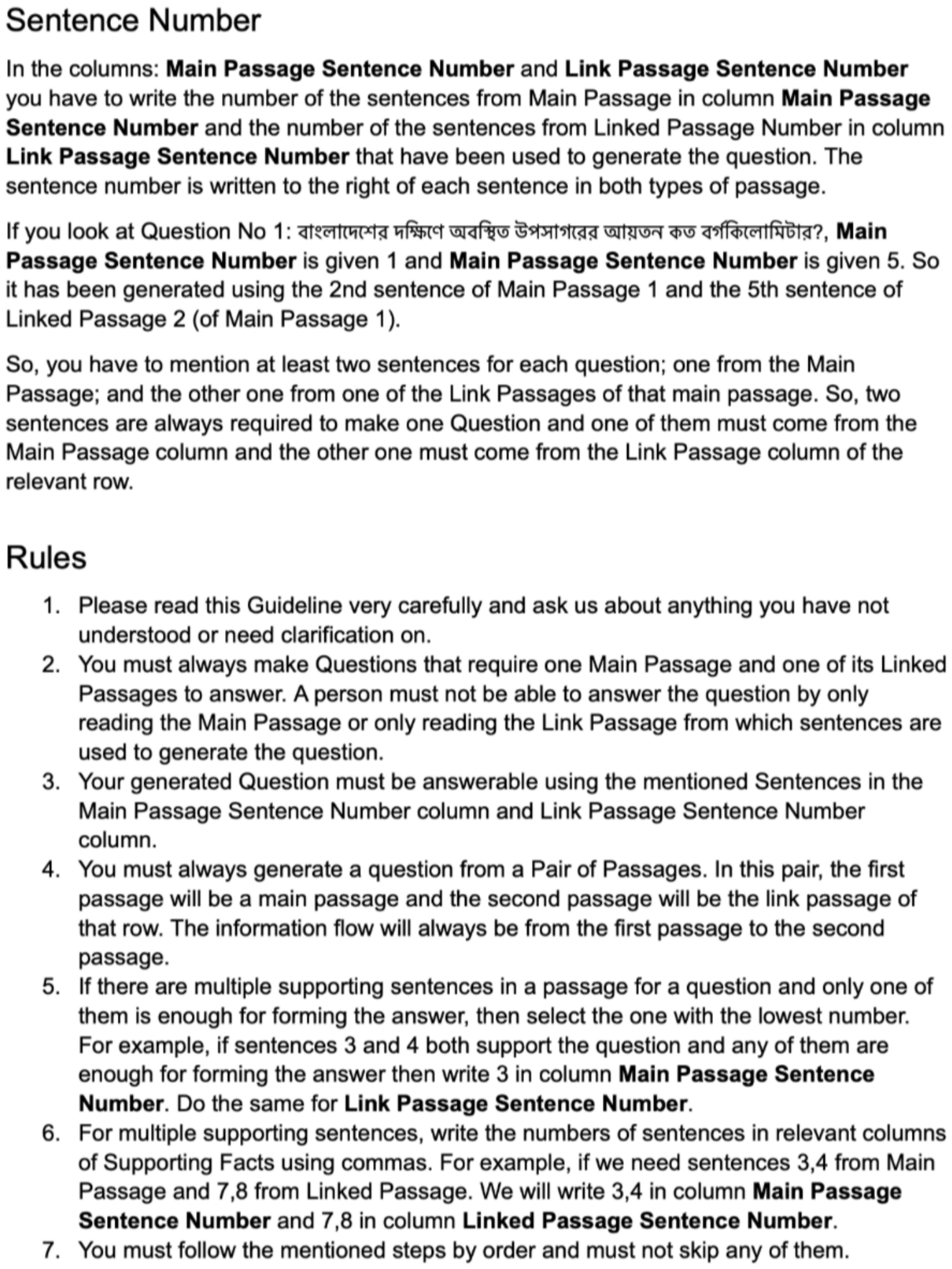
For a Generated Question from a Main Passage - Link Passage pair, if you have written your question in the **Question** column by indexing it as 1 (i.e. 1. বাংলাদেশের সবচেয়ে বড় নদীর উৎপত্তি কোন পর্বতমালায়?) then write the answer giving the same index (i.e. 1. হিলাময়) in the **Answer** column and give the same index in both **Main Passage Sentence Number** (i.e 1. 3) column and **Link Passage Sentence Number** (i.e 1. 5) column in the relevant row of the pair.

Steps

You have been given 15 Main Passage - Link Passage pairs in the Excel sheet:

1. First, you have to read a Main Passage and its Linked Passage very carefully.
2. Then, you have to make Questions that require one or more sentences from the Main Passage and one of its Linked Passages to answer. Remember, a person must not be able to answer the question if s/he only reads the whole Main Passage or only the whole Link Passage.
3. Try to generate at least 3 Questions for one Main Passage - Link Passage Pair. But this is not a hard rule. We just want you to try this. In many cases, it may not be feasible. If such, you don't have to worry about it.
4. For each of the Questions write the Answer, Main Passage Sentence Number, and Link Passage Sentence Number in relevant columns.
5. Try to make the make Questions as diverse as possible. We are giving you some types just to give you an idea. But you are free to make any types of question you want, we expect you won't be limited to these types only.
    a. Factoid Type: (Answers are usually short phrases)
        i. What: বাংলাদেশের দক্ষিণে অবস্থিত উপসাগরের আয়তন কত বর্গকিলোমিটার?
        ii. Who: পূর্ব বাংলার ১৯৭০ সালের সাধারণ নির্বাচনের সংখ্যাগরিষ্ঠ দলের নেতা কে ছিলেন?
        iii. When: কোন যুগে ভৌগলিকভাবে একটি উর্বর বদ্বীপের উপরে অবস্থিত রাষ্ট্রটি একটি প্রদেশ ছিল?
        iv. Where: বাংলাদেশের দক্ষিণে অবস্থিত উপসাগরটি ভারত মহাসাগরের কোনদিকে অবস্থিত?
        v. Which: প্রাচীন ও ধ্রুপদী যুগে পূর্ববাংলায় কোন জনপদগুলো গড়ে উঠেছিল?
    b. Casual Type: (Answers are usually descriptive)
        i. Why: দক্ষিণ এশিয়ার সার্বভৌম রাষ্ট্রে জনযুদ্ধ কেন সংঘটিত হয়?
        ii. How: কীভাবে দক্ষিণ এশিয়ার স্বাধীন ও সার্বভৌম রাষ্ট্রের অভ্যুদয় ঘটে?
    c. Confirmation Type: (Answer: Yes or No)
        i. বারাক ওবামা কি যুক্তরাষ্ট্রের ৪৪-তম প্রসিডেন্ট ছিলেন?
    d. Comparison Type:
        i. বাংলাদেশের হয়ে ওডিআই ফরম্যাটে লিটন দাস ও আশরাফুলের মধ্যে কে বেশি রান করেছে?
6. Lastly check your questions for grammatical and spelling mistakes.
7. Repeat the above six steps for 15 pairs.

Figure 3.3: Crowd Worker Guideline Page 02



Sentence Number

In the columns: **Main Passage Sentence Number** and **Link Passage Sentence Number** you have to write the number of the sentences from Main Passage in column **Main Passage Sentence Number** and the number of the sentences from Linked Passage Number in column **Link Passage Sentence Number** that have been used to generate the question. The sentence number is written to the right of each sentence in both types of passage.

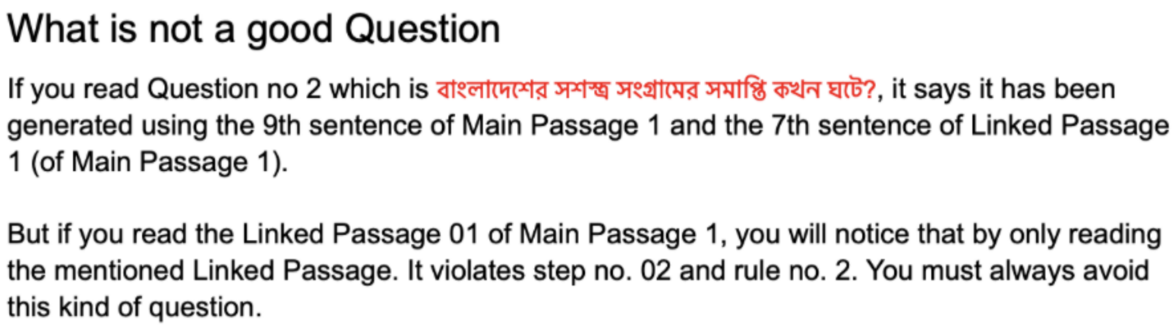
If you look at Question No 1: বাংলাদেশের দক্ষিণে অবস্থিত উপসাগরের আয়তন কত বর্গকিলোমিটার?, **Main Passage Sentence Number** is given 1 and **Main Passage Sentence Number** is given 5. So it has been generated using the 2nd sentence of Main Passage 1 and the 5th sentence of Linked Passage 2 (of Main Passage 1).

So, you have to mention at least two sentences for each question; one from the Main Passage; and the other one from one of the Link Passages of that main passage. So, two sentences are always required to make one Question and one of them must come from the Main Passage column and the other one must come from the Link Passage column of the relevant row.

Rules

1. Please read this Guideline very carefully and ask us about anything you have not understood or need clarification on.
2. You must always make Questions that require one Main Passage and one of its Linked Passages to answer. A person must not be able to answer the question by only reading the Main Passage or only reading the Link Passage from which sentences are used to generate the question.
3. Your generated Question must be answerable using the mentioned Sentences in the Main Passage Sentence Number column and Link Passage Sentence Number column.
4. You must always generate a question from a Pair of Passages. In this pair, the first passage will be a main passage and the second passage will be the link passage of that row. The information flow will always be from the first passage to the second passage.
5. If there are multiple supporting sentences in a passage for a question and only one of them is enough for forming the answer, then select the one with the lowest number. For example, if sentences 3 and 4 both support the question and any of them are enough for forming the answer then write 3 in column **Main Passage Sentence Number**. Do the same for **Link Passage Sentence Number**.
6. For multiple supporting sentences, write the numbers of sentences in relevant columns of Supporting Facts using commas. For example, if we need sentences 3,4 from Main Passage and 7,8 from Linked Passage. We will write 3,4 in column **Main Passage Sentence Number** and 7,8 in column **Linked Passage Sentence Number**.
7. You must follow the mentioned steps by order and must not skip any of them.

Figure 3.4: Crowd Worker Guideline Page 03



What is not a good Question

If you read Question no 2 which is বাংলাদেশের সশস্ত্র সংগ্রামের সমাপ্তি কখন ঘটে?, it says it has been generated using the 9th sentence of Main Passage 1 and the 7th sentence of Linked Passage 1 (of Main Passage 1).

But if you read the Linked Passage 01 of Main Passage 1, you will notice that by only reading the mentioned Linked Passage. It violates step no. 02 and rule no. 2. You must always avoid this kind of question.

Figure 3.5: Crowd Worker Guideline Page 04

3.1.3 Multi-hopness Verification

In this stage, we have verified the multi-hopness of each QA pair. Multi-hopness indicates that each question needs sentences at least two passages to generate the answer. For this, we have manually checked each of the generated questions from the step above.

3.1.4 Quality Verification

In this stage, we have verified the quality of the dataset.

1. Looked for syntactic and other errors and corrected them.
2. Verified that the questions are answerable from the passages.
3. Verified the diversity of the questions and the answers (different types of reasoning, questions, and answers).
4. Format and finalize the dataset.

3.2 Dataset Evaluation

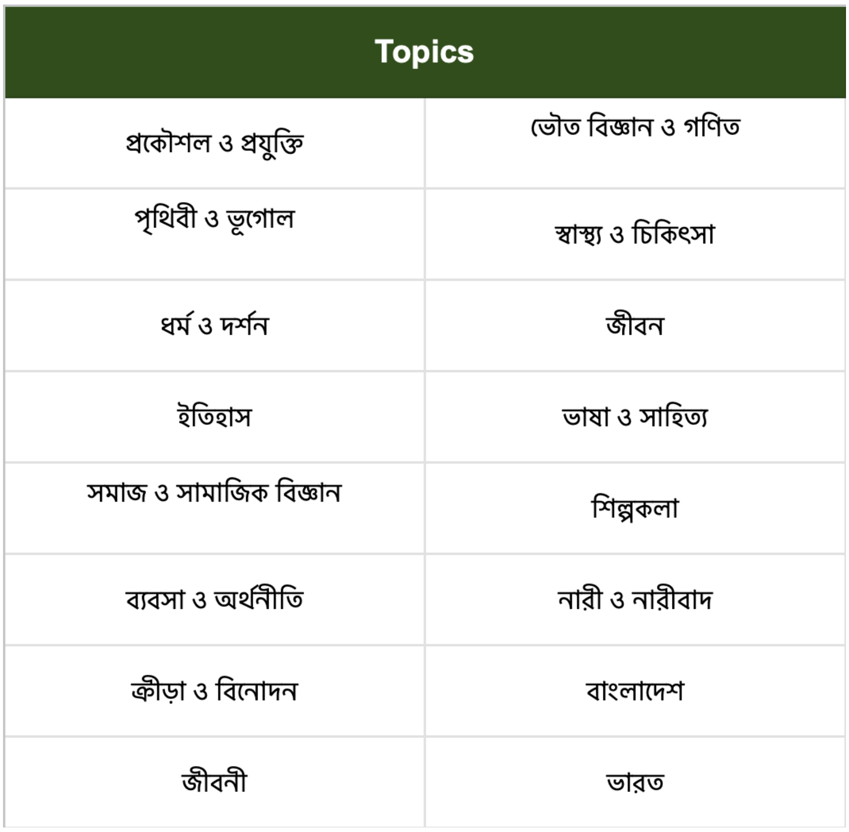
We have used a pre-trained language model trained in the Bengali language to evaluate the accuracy of our dataset.

Chapter 4

Dataset Analysis and Evaluation

4.1 Dataset Analysis

Analysis of Passages. We have collected passages from the following 16 different main topics of Bengali Wikipedia.



Topics	
প্রকৌশল ও প্রযুক্তি	ভৌত বিজ্ঞান ও গণিত
পৃথিবী ও ভূগোল	স্বাস্থ্য ও চিকিৎসা
ধর্ম ও দর্শন	জীবন
ইতিহাস	ভাষা ও সাহিত্য
সমাজ ও সামাজিক বিজ্ঞান	শিল্পকলা
ব্যবসা ও অর্থনীতি	নারী ও নারীবাদ
ক্রীড়া ও বিনোদন	বাংলাদেশ
জীবনী	ভারত

Figure 4.1: Bengali Wikipedia Topics

We have curated a total of 1,351 unique passages. 414 of these are main passages and 937 of them are linked passages. In many cases, the main passages have been used as linked passages as well.

Main Passage	Linked Passage	Total
414	937	1351

Table 4.1: Number of passages in each type

We have analyzed the number of passages generated from each topic. The following line graph shows the number of main passages, linked passages, and total passages correlate positively.

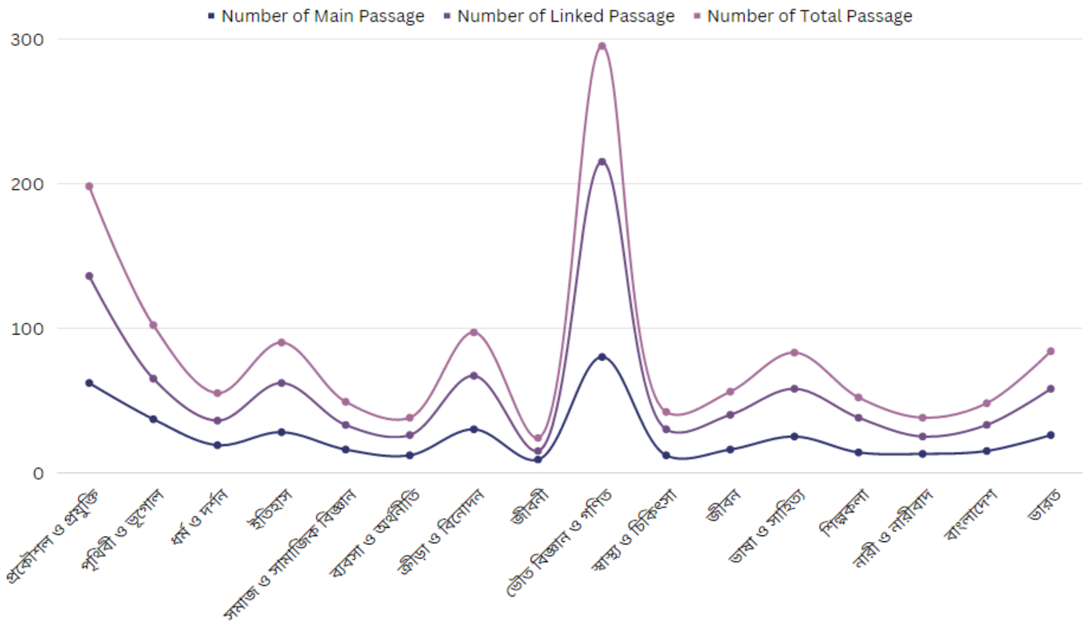


Figure 4.2: Number of Passages Per Topic

Variation in QA Pairs.

We have analyzed the number of QA pairs in each topic. The following bar chart depicts the number of QA pairs per topic and their comparison. We can see it correlates with the number of generated passages (whether it is main, liked, or total) in each topic.

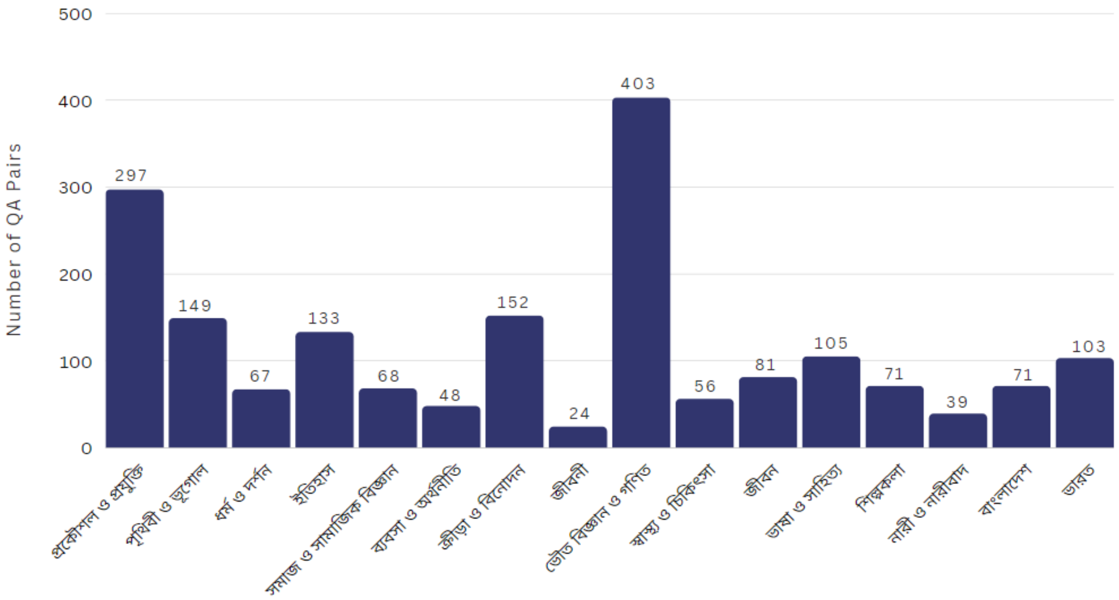


Figure 4.3: QA Pairs Per Topic

Question Types. The crowd workers have mainly generated four types of questions. They have mostly generated factoid-type questions which contribute 73% to the total number of questions, followed by causal-type questions with 18%. They made confirmation-type questions the least with 3.5% and comparison-type questions slightly higher than confirmation-type questions with 4.5%. Nonetheless, with a combined total percentage of 8, they really have a very small presence in the dataset.

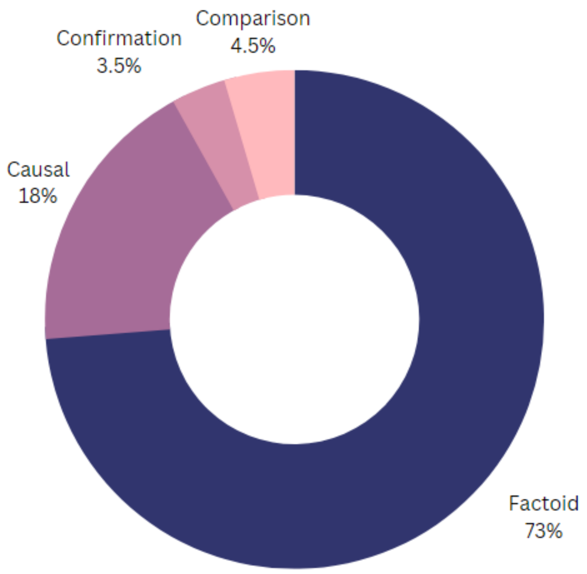


Figure 4.4: QA Pairs generated per Question Type

- **Factoid Type**: Answers to these types of questions are usually short phrases and a fact is wanted in the question. These types of questions include keywords like What, who, when, where, and which. The following graph shows the percentage of QA pairs generated in each type of Factoid Question. The percentage is shown relative to the total number of Factoid Questions.

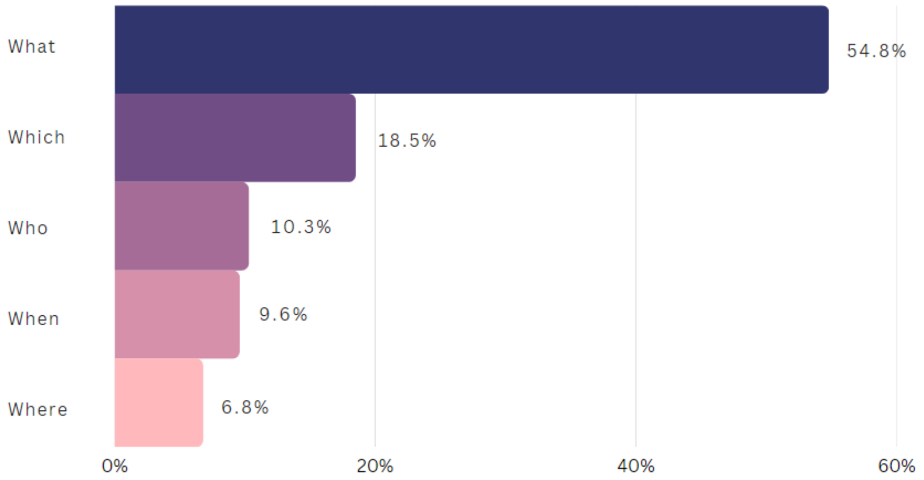


Figure 4.5: Percentage of QA Paris in each Factoid Question Type

- **Casual Type**: Answers to these types of questions are usually descriptive. These types of questions include keywords like why and how. The following graph shows the percentage of QA pairs generated in each type of Causal Question. The percentage is shown relative to the total number of Causal Questions.

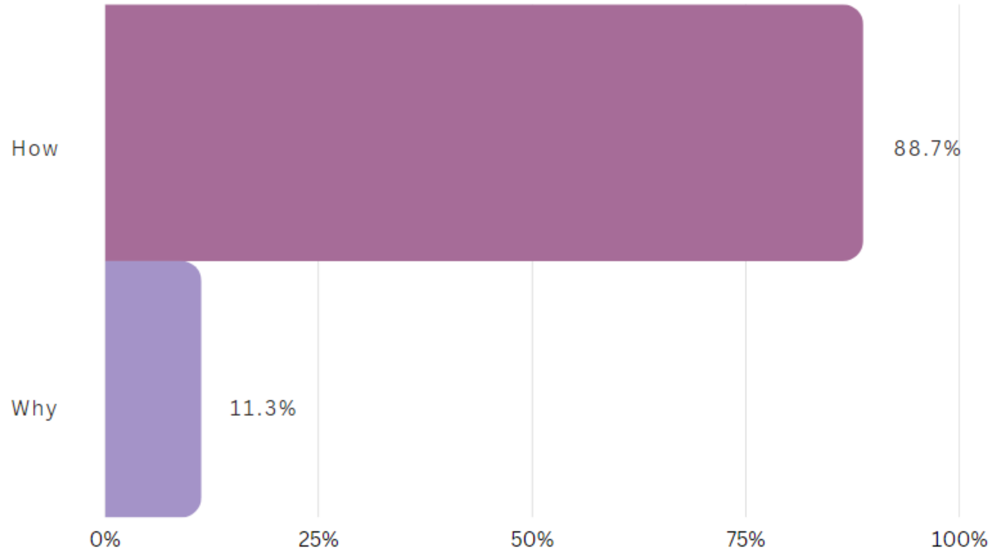


Figure 4.6: Percentage of QA Paris in each Causal Question Type

- **Confirmation Type**: Answers to these types of questions is either yes or no.
- **Comparison Type**: These questions ask to compare two or more persons, places, items, etc. The answer to this type of question is either true or false.

In terms of QA pair generation, the information flow was always from the main passage to the linked passage. And the generated questions need at least one sentence from the main passage and one from the linked passage to generate the answer. But in some cases, more than one sentence from a main passage or linked passage or both have been used to generate a question.

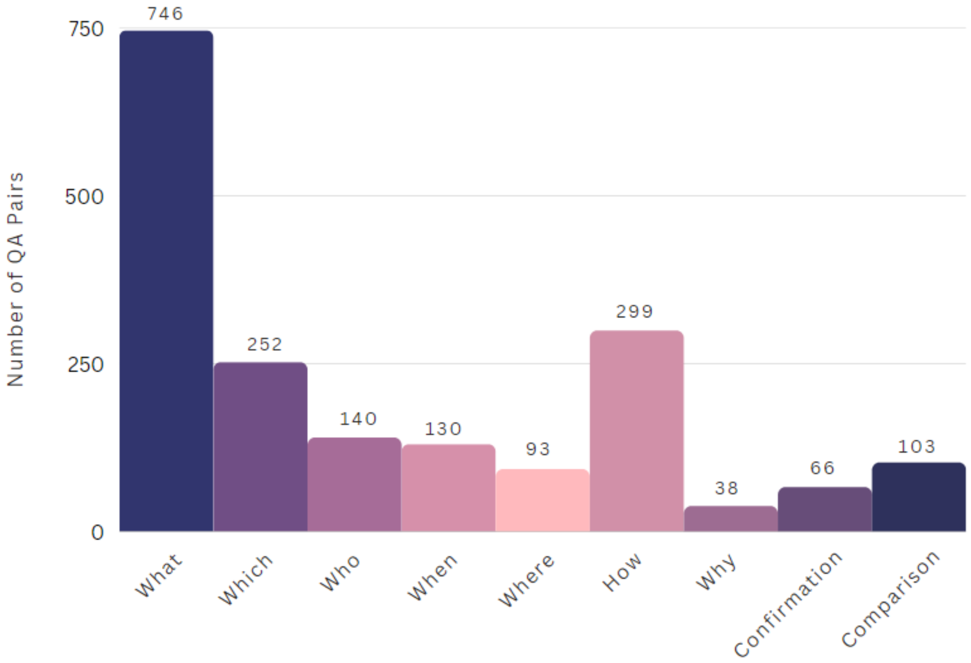


Figure 4.7: QA Pairs Per Question Type

The crowd workers have generated a total of 1874 QA pairs. A detailed analysis of the types of questions is shown in the table below.

Question Type		Percentage	Total	Number of QA Pairs
Factoid	What	40%	73%	746
	Which	13.5%		252
	Who	7.5%		140
	When	7%		130
	Where	5%		93
Causal	How	16%	18%	299
	Why	2%		38
Confirmation		3.5%	3.5%	66
Comparison		5.5%	5.5%	103

Table 4.2: Variation in Question Types

Our analysis of question types in terms of hop reveals that our dataset consists of 413 single-hop question-answer (QA) pairs and 1454 multi-hop QA pairs. These findings indicate the presence of both straightforward, single-step questions and more complex questions requiring multiple steps or sources of information to answer. This diversity in question types enhances the richness and depth of our dataset, allowing for a comprehensive exploration of different knowledge and reasoning levels.

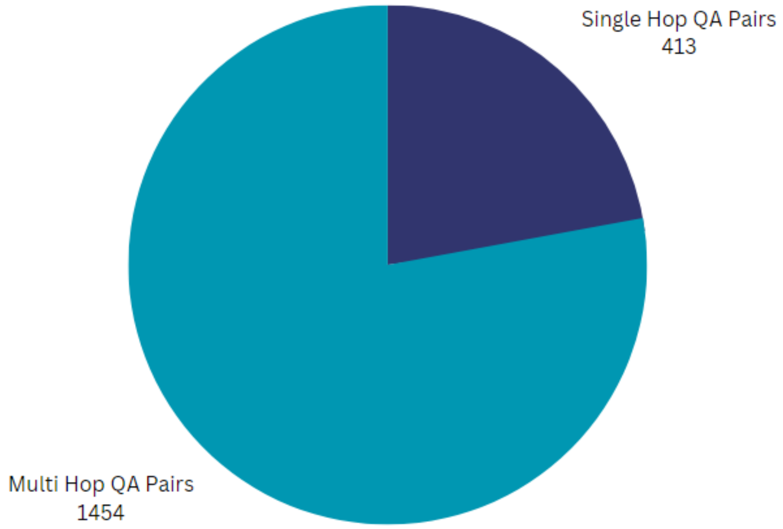


Figure 4.8: Question Types in Terms of Hop

4.2 Dataset Evaluation

We have used chat GPT manually for evaluating our dataset. Due to difficulties faced to evaluate our model using Language Models and not having enough time to solve those problems, we ultimately decided to use ChatGPT. There is also a reason for not using it earlier as we didn't have the knowledge prior to using it in dataset evaluation.

Why use ChatGPT. The reasons why we have used chatGPT:

- ChatGPT is a benchmark right now, and it outperforms regular models like BERT.
- All NLP datasets are evaluated using ChatGPT API now.

We have taken a subset of 150 QA Pairs to evaluate our dataset. We manually inputted the data into ChatGPT and asked for the answer in one line and also the supporting sentences used for generating the answer.

EM Score. In our dataset evaluation using ChatGPT, we found 63 exact matches and 87 non-matches. The exact match (EM) score, representing perfectly matching responses, was determined to be 42%. These results provide insights into the model's performance and highlight areas for improvement.

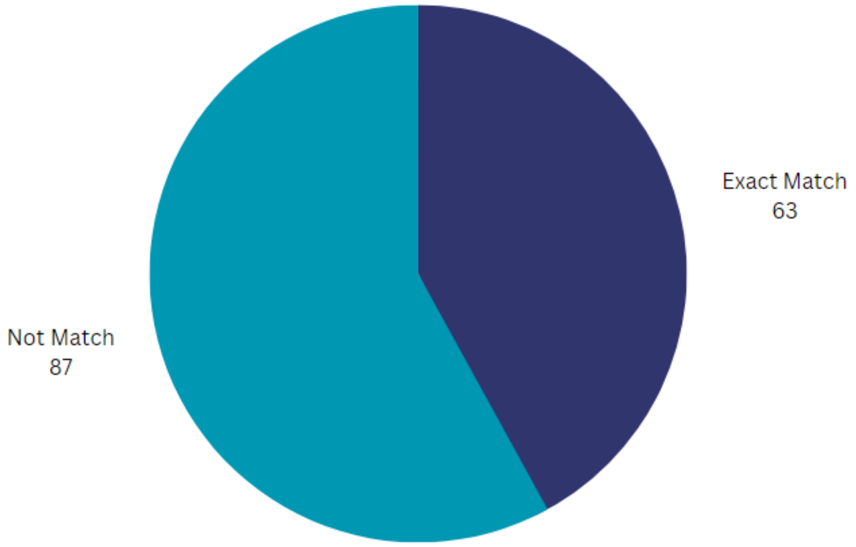


Figure 4.9: Exact Match and Not Match

F1 Score. After evaluating 150 pairs, we have got a precision of 51.5%, recall of 47%, and an F1 Score of 49.15%.

Comparison with HotPotQA. We have compared the evaluation of our dataset with the best score obtained by the HotPotQA dataset. As the first multi-hop dataset of the Bengali Language, it has done quite well compare to HotPotQA.

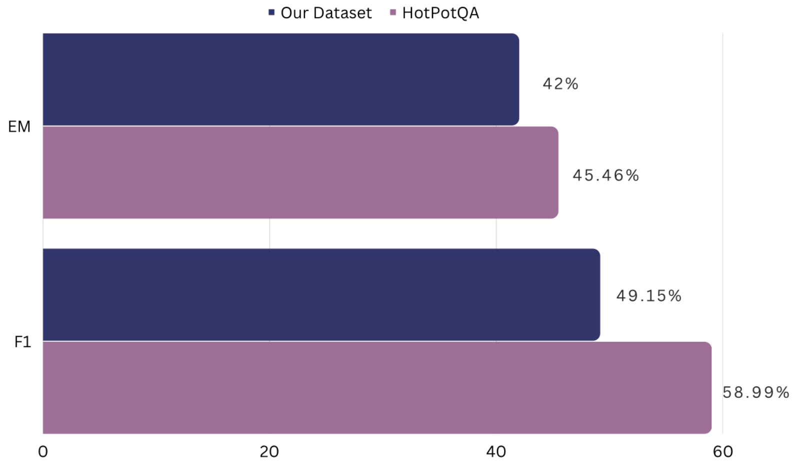


Figure 4.10: Our Dataset Vs HotPotQA

Next Steps. We are currently working on integrating the API of ChatGPT to evaluate our dataset. After properly integrating the API, we will get a proper evaluation of our dataset.

Chapter 5

Conclusion

This is the first-ever attempt at curating a multi-hop dataset in Bengali Language. In this study, we have shown how Bengali Language is syntactically and semantically different than English language and proved it is necessary to build a multi-hop dataset in Bengali Language. Our study is conducted based on the proposals and findings addressed by [3] and we also have studied other prominent papers, but the main papers [2] and [6] along with the [3] that we have used to generate a solid methodology for making a multi-hop dataset in the Bengali Language. We have taken the best parts from the mentioned methodologies of these papers and modified them according to our needs. This way, we have been able to make a very good methodology for curating our dataset.

We have studied many papers on Bengali datasets as well. This includes [7], [9], [10], and [11]. As there is still no multi-hop dataset has been made in the Bengali Language all of the datasets are single-hop datasets. In all of the studies, the researchers have focused on generating a model rather than the dataset. They have generated the datasets in order to train and evaluate the model as it is a dependent thing that needs to be done. In [7] authors have translated the SQuAD 2.0 dataset in [8]. We have analyzed the dataset and found out it is very poor and in some cases, the contexts are missing. The translation can't also reflect the behavior of the real humans, it is also a minus point of this dataset. In [9] authors have collected their contexts from newspapers, books, and sonnets. This is a single-hop dataset with 1,676 paragraphs and 8,027 questions. This dataset is best in quality than other ones that we have studied. In [10] authors have collected their contexts from the internet. They have not mentioned the domains or what type of writings they have collected, just mentioned the phrase "famous Bengali writings". Their dataset is single-hop and contains 3636 QA Pairs. In [11] authors have built a big dataset of 27.5 GB but most of the part of the dataset is for Natural Language Inference (NLI) as they have built a pre-trained model of Bangla. Their QA dataset is a translation of SQuAD 2.0 and TyDi QA (Typologically Diverse Question Answering) in [12] datasets with a total of 354K QA Pairs of 150k in SQuAD 2.0 and 204K in TyDi QA. It has been translated from English. It is a single-hop dataset with generative answering. The dataset is not up to the mark as passages are missing for some QA Pairs, and not cleaned and organized properly. So, from our study, we have found even the existing datasets are not up to the mark and they also have not mentioned the methodology they have used to create these datasets. Overall, no credible methodology, and no verification of the quality led to poor quality.

We have generated a total of 1351 passages with 414 main passages and 937 linked passages from 16 main topics of Bengali Wikipedia. Most of the passages are related to science, engineering, technology, and math as Bengali Wikipedia holds more passages regarding them.

We have analyzed the QA pairs generated from the crowd workers and have got several findings. They have generated a total of 1867 QA pairs. Among them, 413 were single-hop QA Pairs and 1454 were multi-hop QA pairs. They have mainly generated four types of questions namely factoid, causal, conformation, and comparison. The factoid-type questions dominate the dataset with 73% followed by the causal-type questions with 18%. The confirmation-type and comparison-type questions only hold 8% of the dataset which is really low. We have also analyzed the types of factoid questions that have been asked. Here 54.8% 'what' type questions dominate followed by which and who and others later on. In causal-type questions 'how' type questions dominate with 88.7% share and the rest of them are owned by 'why' type questions. We have also analyzed the number of questions generated per topic. This also shows most of the QA pairs have been generated from science, engineering, technology, and math topics.

We have evaluated our dataset over 150 QA Pairs of our dataset. These pairs are selected randomly. We have got an EM score of 42% and an F1 score of 49.15% with a precision of 51.5% and recall of 47%. Compare to our base paper HotPotQA, we were slightly behind as they had the best EM score of 45.46% and the best F1 score of 58.99%. As the first attempt, this is not a bad score at all.

Due to time constraints, we couldn't generate more QA pairs. And due to the voluntary nature of the participants, the QA Pair generation is also not up to the mark. In the future, we want to increase the number of contexts and QA Pairs and also want to make our dataset more robust and better quality of QA Pairs. We also increase the diversity of our questions and answers. We will conduct a human accuracy test on our dataset. We will also check the accuracy of the dataset using different models. For these models, we want to use existing ones and also will build some. We will try to find out the points whose modifications lead to better results. In each case, we will compare our results with human accuracy and observe the progress how the models.

References

[1] B. F. Green, A. K. Wolf, C. L. Chomsky, and K. Laughery, "Baseball: An automatic question-answerer," in *IRE-AIEE-ACM '61 (Western)*, 1961.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: `10.18653/v1/D16-1264`. [Online]. Available: `https://aclanthology.org/D16-1264`.

[3] Z. Yang, P. Qi, S. Zhang, *et al.*, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 2369–2380. DOI: `10.18653/v1/D18-1259`. [Online]. Available: `https://aclanthology.org/D18-1259`.

[4] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang, "HybridQA: A dataset of multi-hop question answering over tabular and textual data," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1026–1036. DOI: `10.18653/v1/2020.findings-emnlp.91`. [Online]. Available: `https://aclanthology.org/2020.findings-emnlp.91`.

[5] T. Kočiský, J. Schwarz, P. Blunsom, *et al.*, "The NarrativeQA reading comprehension challenge," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 317–328, 2018. DOI: `10.1162/tacl_a_00023`. [Online]. Available: `https://aclanthology.org/Q18-1023`.

[6] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking beyond the surface: A challenge set for reading comprehension over multiple sentences," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 252–262. DOI: `10.18653/v1/N18-1023`. [Online]. Available: `https://aclanthology.org/N18-1023`.

[7] 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, 2021.

[8] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 784–789. DOI: `10.18653/v1/P18-2124`. [Online]. Available: `https://aclanthology.org/P18-2124`.

[9] M. A. Haque, S. Sultana, M. J. Islam, M. A. Islam, and J. A. Ovi, "Factoid question answering over bangla comprehension," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1–8. DOI: `10.1109/ISMSIT50672.2020.9254680`.

[10] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, M. M. Anwar, and A. B. M. S. Ali, "Reading comprehension based question answering system in bangla language with transformer-based learning," *Heliyon*, vol. 8, no. 10, e11052, 2022, ISSN: 2405-8440. DOI: `https://doi.org/10.1016/j.heliyon.2022.e11052`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S2405844022023404`.

[11] A. Bhattacharjee, T. Hasan, W. Ahmad, *et al.*, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. DOI: `10.18653/v1/2022.findings-naacl.98`. [Online]. Available: `https://aclanthology.org/2022.findings-naacl.98`.

[12] J. H. Clark, E. Choi, M. Collins, *et al.*, "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, 2020. DOI: `10.1162/tacl_a_00317`. [Online]. Available: `https://aclanthology.org/2020.tacl-1.30`.

[13] S. Wang, "Machine comprehension using match-lstm and answer pointer," Aug. 2016.

[14] T. Kočiský, J. Schwarz, P. Blunsom, *et al.*, "The NarrativeQA Reading Comprehension Challenge," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 317–328, May 2018, ISSN: 2307-387X. DOI: `10.1162/tacl_a_00023`. eprint: `https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl\_a\_00023/1567652/tacl\_a\_00023.pdf`. [Online]. Available: `https://doi.org/10.1162/tacl%5C_a%5C_00023`.

[15] Y. Chang, M. Narang, H. Suzuki, G. Cao, J. Gao, and Y. Bisk, "Webqa: Multihop and multimodal qa," Sep. 2021.

[16] J. Chen and G. Durrett, "Understanding dataset design choices for multi-hop reasoning," *arXiv preprint arXiv:1904.12106*, 2019.

[17] M. Keya, A. K. M. Masum, S. Abujar, B. Majumdar, and S. Hossain, "Bengali question answering system using seq2seq learning based on general knowledge dataset," Jul. 2020. DOI: `10.1109/ICCCNT49239.2020.9225605`.

[18] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang, "HybridQA: A dataset of multi-hop question answering over tabular and textual data," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1026–1036. DOI: `10.18653/v1/2020.findings-emnlp.91`. [Online]. Available: `https://aclanthology.org/2020.findings-emnlp.91`.

[19] F. Zhu, W. Lei, Y. Huang, *et al.*, "TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 3277–3287. DOI: `10.18653/v1/2021.acl-long.254`. [Online]. Available: `https://aclanthology.org/2021.acl-long.254`.