Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

Semi-Supervised Question Answering with Question-Answer Pair Generation in Bengali

Authors

Md. Amimul Ehsan, 170041015

Md. Shihab Shahriar, 170041016

Ahmad Al Fayad Chowdhury, 170041041

Supervisor

Dr. Abu Raihan Mostofa Kamal

Professor, Department of CSE

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc. Engineering in CSE

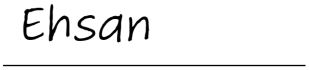
Academic Year: 2020-2021

25th April, 2022

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. Amimul Ehsan, Md.Shihab Shahriar and Ahmad Al Fayad Chowdhury under the supervision of Dr. Abu Raihan Mostofa Kamal, Professor of Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Authors:



Md. Amimul Ehsan

Student ID: 170041015

Md. Shihab Shahriar

Student ID: 170041016

Ahmad Al Fayad Chowdhury

Student ID: 170041041

Supervisor:

Dr. Abu Raihan Mostofa Kamal

Professor

Department of Computer Science and Engineering

Islamic University of Technology

Acknowledgement

We would like express our gratitude towards IUT authority for granting us the fund and providing assistance required to implement our proposed system. We are indebted to our professor, Dr. Abu Raihan Mostofa Kamal for providing us with insightful knowledge and guiding us at every stage of our journey, and would like to extend our sincere gratitude towards Mohammad Ishrak Abedin for his immense support. 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.

Abstract

Although deep learning architectures and large scale datasets have led to great performance on question answering tasks in high resource languages like English, their performance on lower resource languages, like Bengali, is considerably poorer. This is due to the scarcity of labeled data, which can be attributed to the massive amount of human effort and time required to create such datasets. We work towards a translated Stanford Question Answering Dataset (SQuAD) 1.1 in Bengali and ensure that it is of high quality by using a state-of-the-art translation model and a novel embedding based matching approach to properly align the answer spans in the target language (Bengali) in correspondence with the source language, English. We also introduce an end-to-end question answer generation (QAG) system in the Bengali language to generate question answering (QA) datasets for QA models using roundtrip consistency incorporated in a sequence-to-sequence generation task using Googles mT5 model. Additionally, we train 3 different QA models on our Bengali translated dataset achieving EM and F1 scores of 46.1 and 66.2 respectively. Finally, we demonstrate the effectiveness of our QAG model on a sample dataset of news articles in generating domain-specific QA datasets.

Keywords— Question Answering, Question Answer Generation, Low Resource Language, Translated Dataset, Synthetic Dataset

Contents

List of Tables

List of Figures

1 Introduction

1.1 Question Answering

The task of **question answering** is to build systems that can respond to questions posed by humans in a natural language. This problem can be cast as an **information retrieval problem** (finding relevant documents, extracting possible answers and ranking them) or as a **reading comprehension problem** (finding answer spans within a context provided to the model).

With the advent of complex **deep neural networks** and **transfer learning** in novel architectures like biLM [1], BERT [2], UniLM [3] that are trained on large corpora, performance on the task of question answering has now become comparable to humans.

Question answering is of particular relevance to industries, for example, looking to automate responses to customer queries via chatbots, in academics, for example, in finding quick answers to questions without having to go through entire passages.

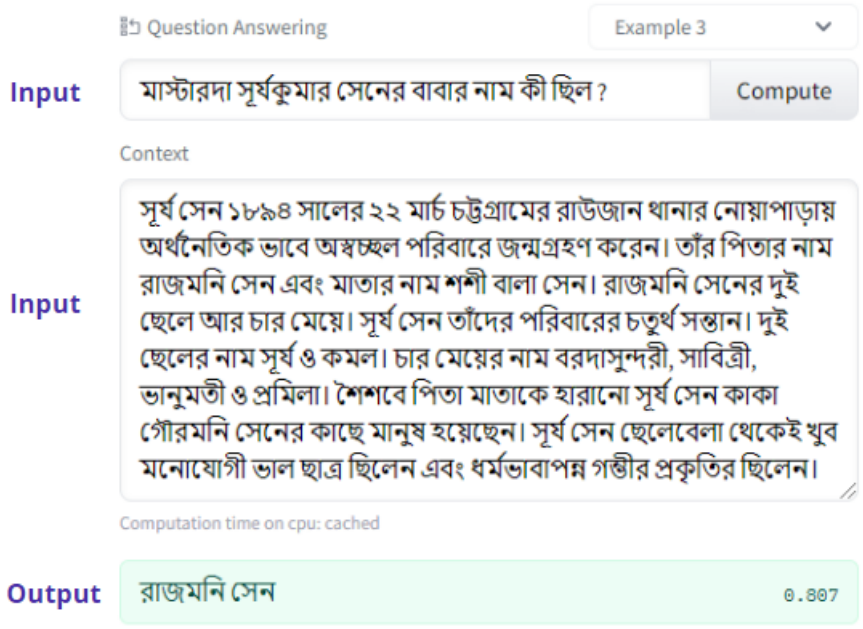


Figure 1: Example of a question answering task in Bengali

However, lower resource languages, like Bengali, have seen considerably less progress in this field than higher resource, globally spoken languages, like English. This is primarily due to the **scarcity of labeled data**, which can be attributed to the massive amount of human effort and time required to create such datasets. State-of-the-art results [4] have been obtained on a translated SQuAD [5] dataset (**Bangla-SQuAD**) using BERT, RoBERTa [6], DistilBERT [7]. But, there could be different linguistic biases present in different Bengali texts that are

not being captured in the translated SQuAD dataset; they could be present in blogs, news articles etc. which are not considered in the above evaluation and are not present as complete datasets. Furthermore, as per the literature, Bangla-SQuAD contains only a **subset of the actual SQuAD dataset**.

1.2 Translating Existing Datasets

One approach that has been explored is **machine translated datasets** from a high resource source language to a lower resource target language. However such datasets often present data quality issues. Despite being economical in terms of time and cost, a major issue in this approach is **locating the correct answer span** (the sequence of words considered to be the answer as found in the context) in the translated context. This results in discarding data samples or degrading the quality of the datasets since they cannot guarantee finding the correct answer span in the target language.

1.3 Question Generation

An alternative approach to tackle the problem is to use **automatically generated QA pairs** from a large amount of unstructured texts (e.g. Wikipedia). This task is called **question answer generation (QAG)**.
The aim of question answer generation is to take a passage and generate a good quality question-answer pair from the information provided in the passage.

Figure 2: Question answer generation from contexts

This task has garnered great interest from natural language processing communities in both industry and academia [8].
While earlier QAG models relied on **recurrent neural networks (RNNs)** and their regular and attention augmented variants [9], the core problem with RNNs is their **inability to capture semantic information in long sequences**. Recently, attention-based transformer models [10] and pre-trained language models like BERT [2] and T5 [11] are being employed for this task of

QAG [12] [13].

Again, however, as with the task of QA, QAG has seen very little progress in lower resources languages, like Bengali. To the best of our knowledge, at the time of writing, there are no experiments or models designed to tackle the task of QAG in the Bengali language.

1.4 Our work

To alleviate the issues of data scarcity in the Bengali language, we propose an **end-to-end QAG system** that takes in **contexts from different domains** and **generates QA pairs**. The system consists of two portions - an **answer extraction model** that extracts interesting answer spans from the provided context and a **question generation model** that uses the extracted answer and the context to generate QA pairs. We leverage the power of Google's mT5 language model [14] for our purposes, since this model is already pre-trained on large volumes of text in multiple languages, including Bengali, across multiple tasks.

We also encountered several issues with the existing QA dataset, Bangla-SQuAD, as a result of which, we found it unreliable to train our answer extraction and question generation models. To that end, we **translated the SQuAD1.1** using **Facebooks M2M100 model** and a novel **embedding based answer alignment** approach discussed in later sections. We then use this translated dataset to train our models and achieve respectable performance on the QAG task.

In the following sections, we examine existing works in the domain of question answering and question answer generation as well as question answering datasets and their machine translated versions. We identify limitations with one particular dataset of interest, Bangla-SQuAD, and existing tokenizers for the Bengali language. We also explain how research in this domain is stunted by a lack of accessible computational resources. Next, we present our solutions in the form of an improved translated dataset based on SQuAD1.1 as well as a QAG model that is successfully able to generate consistent QA pairs, and discuss some of our findings.

2 Related Works

2.1 Datasets

From 2015 onwards, there has been a massive boost in curating large scale question answering datasets in the English language. This has resulted in datasets such as CNN/Daily Mail Corpus [15], SQuAD 1.1 [5], RACE [16], SQuAD 2.0 [17], CoQA [18] and Natural Question Corpus [19]. Further, domain-specific question answering datasets like NewsQA [20] and TriviaQA [21] have also arisen. There are also cross-lingual reading comprehension datasets such as XQuAD [22]

and MLQA [23].

Even so, SQuAD1.0 and SQuAD2.0 remain the most widely used benchmarks, incorporating human performance baselines. These have also been included in language understanding evaluation benchmarks like GLUE and SuperGLUE.

The SQuAD dataset is divided into titles, each consisting of multiple paragraphs, or contexts, each of which has several question answer pairs (if they are answerable). The answers are marked with the start index of the span and the answer text is provided. An example is shown in figure 3:


```
{
    "title": "University_of_Notre_Dame",
    "paragraphs": [
        {
            "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is
            "qas": [
                {
                    "answers": [
                        {
                            "answer_start": 515,
                            "text": "Saint Bernadette Soubirous"
                        }
                    ],
                    "question": "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?",
                    "id": "5733be284776f41900661182"
                },
                {
                    "answers": [
                        {
                            "answer_start": 188,
                            "text": "a copper statue of Christ"
                        }
                    ],
                    "question": "What is in front of the Notre Dame Main Building?",
                    "id": "5733be284776f4190066117f"
                },
                {
                    "answers": [
                        {
                            "answer_start": 279,
                            "text": "the Main Building"
```


Figure 3: A snippet from the SQuAD dataset

There are a total of 442 titles, 18,896 paragraphs or contexts and 87,599 question answer pairs.

However, in the Bengali language, the only established datasets at the time of writing this report are the Bangla-SQuAD dataset proposed by Mayeesha et al. 2020 [4] and another factoid-based QA dataset [24]. Of these two, the former is of particular interest to us, since it is a direct translation of the SQuAD2.0 dataset into Bengali. The quality of the dataset is discussed in section 3.1.

```
{ 'paragraphs': [ { 'context': 'চার্লসটন আমেরিকা যুক্তরাষ্ট্রের দক্ষিণ '
                               'ক্যারোলাইনা রাজ্যের প্রাচীনতম এবং দ্বিতীয় '
                               'বৃহত্তম শহর, চার্লসটন কাউন্টির কাউন্টি আসন এবং '
                               'চার্লসটন – নর্থ চার্লসটন – সামারভিলে '
                               'মেট্রোপলিটন স্ট্যাটিস্টিকাল এরিয়ার প্রধান '
                               'শহর  শহরটি দক্ষিণ ক্যারোলিনার উপকূলরেখার '
                               'ভৌগলিক মিডপয়েন্টের ঠিক দক্ষিণে অবস্থিত এবং '
                               'অ্যাশলে এবং কুপার নদীর নদীর সংগম দ্বারা গঠিত '
                               'আটলান্টিক মহাসাগরের একটি খাঁটি চার্লস্টন '
                               'হারবারে অবস্থিত, অথবা স্থানীয়ভাবে প্রকাশিত '
                               'হয়েছে, "যেখানে কুপার এবং অ্যাশলে রয়েছে। '
                               'নদীগুলি একত্র হয়ে আটলান্টিক মহাসাগর গঠনে আসে। '
                               '"  ',
                   'qas': [ { 'answers': [],
                              'id': '572e8700cb0c0d14000f1252',
                              'is_impossible': True,
                              'question': 'দক্ষিণ ক্যারোলিনার প্রাচীনতম শহরটি '
                                          'কী?  '},
                            { 'answers': [ { 'answer_start': 0,
                                             'text': 'চার্লসটন আমেরিকা য'}],
                              'id': '572e8700cb0c0d14000f1253',
                              'is_impossible': False,
                              'question': 'চার্লসটন, দক্ষিণ ক্যারোলিনা কোন '
                                          'কাউন্টিতে অবস্থিত?  '},
                            { 'answers': [ { 'answer_start': 0,
                                             'text': 'চার্লসটন আমেরিকা য'}],
                              'id': '572e8700cb0c0d14000f1254',
                              'is_impossible': False,
                              'question': 'চার্লসটন কোন বন্দরে অবস্থিত?  '},
```

Figure 4: A snippet from the Bangla-SQuAD dataset

2.2 Machine Translated Datasets

Due to the impressive quality, diversity and comprehensiveness of SQuAD, recent works have replicated SQuAD in many mid-resource languages such as Korean [25], French [26], Russian [27] etc. However, the procedure to build such a large scale dataset for any language is time and labor intensive which has motivated the use of automated machine translation of SQuAD. The challenges in translating a QA dataset include **finding good Neural Machine Translation models** from English to the target language and **finding the answer span correctly** in the translated context. A recent work in this domain uses some combination of approaches to highlight the answer span in the target language such as finding the exact match of the answer text in the context, marking the answer spans in quotation while translating or discarding samples that do not contain words in the answer text in the context. The work by Carrino et al., 2020 [28] uses a **word alignment model** to find the answer span in the target language which requires a highly accurate word alignment model between the source and the target language generally available with a low accuracy for low-resource target language. In our work,

we propose a novel method using **FastText word embeddings** [29] to find the answer spans in the context of the target language.

2.3 Question Answering

There has been very scarce work in the domain of question answering in the Bengali language. Bannerjee and Bandyopadhyay, 2012 [30], attempted to build a **question classification system** in Bengali by extracting lexical, syntactic and semantic features like wh-words, other interrogative words, parts-of-speech tags etc. This work was mainly aimed at classifying questions and not answering them. It was riddled with issues mainly due to:

- The existence of far more interrogative words in Bengali than in English,
- The fact that they can appear anywhere in the sentence,
- A lack of high quality tools like parts-of-speech taggers, named entity recognition systems etc. and a benchmark corpora.

While Nirob et al., 2017 [31] built a similar system leveraging **support vector machines**, they ran into the same issues discussed.
Later, Bannerjee et al., 2014 attempted to build the first **Bengali factoid based question answering system, BFQA** [32], which was an information retrieval system that classified questions, retrieved relevant sentences, ranked them and extracted correct answers. Hoque et al., 2015 built **BQAS** [33], a bilingual question answering system that could generate and answer factoid based questions from English and Bengali documents. Nurul Huda 2019 [34] also implemented a similar question answering system but based it entirely on time-related questions.
However, none of the work before Mayeesha et al., 2020 [4] employed deep learning techniques on SQuAD-like reading comprehension datasets in Bengali. Not only do they use the **multilingual BERT model** [35] in this setting, but they also **translate a large subset of SQuAD2.0 [17] to Bengali** and use that synthetic dataset to fine-tune their model. They compare the performance of multi-lingual BERT with other variants of the BERT model like **RoBERTa** [6], and **DistilBERT** [7] and achieve the highest exact match and F1 scores with RoBERTa.

2.4 Question Answer Generation

Primarily we are focusing on the task of generating questions from a given input context and an answer. Due to the nature of this task, it has been looked at as a helpful tool especially

in academia where it can be used as a component in intelligent tutoring systems or generating assessment material for courses.

Question generation is fundamentally the task of **automatically generating questions from various inputs**. This can be used to generate assessments for course materials or as a component in intelligent tutoring systems.

Question generation is concerned with two questions - **what to ask and how to ask**. The first part, content selection, was tackled in the past by applying semantic or syntactic parsing of text sequences to obtain intermediate symbolic representations. The second part involves question construction which takes these representations and converts them to natural language questions either in a transformation-based or a template-based approach. [36]

These have all proven to be very confining, reductionist approaches. In contrast, deep learning frameworks provide end-to-end architectures jointly optimizing for both the content selection task and the question construction task. Most current models follow the **sequence-to-sequence approach** and employ **transformer-based architectures** to learn the content selection via the encoder and the question construction via the decoder. Under the sequence-to-sequence framework the task is framed as follows:

Given a context X and possibly an answer A (for answer-aware question generation), the model aims to generate a question Y that maximizes the conditional likelihood

$$\bar{Y} = arg \max_{Y} P(Y|X, A)$$

Most sequence-to-sequence question generation models differ only in certain factors like answer-encoding (for answer-aware question generation), question word generation and paragraph-level contexts. Models have solved the problem of answer encoding by either treating the answers position as an input feature (Zhao et al., 2018 [37]), by encoding the answer with a separate recurrent neural network (RNN) (Duan et al., 2017 [38], Kim et al., 2019 [39]) or a mixture of both via transformer-based (Lee et al., 2020 [40], Alberti et al., 2019 [13], Chan et al., 2019 [12]) architectures. To tackle the question word generation problem, Duan et al., 2017 proposed two sequence-to-sequence models - one to generate a "how to #", "where is #" etc. template and the other to fill in the blanks to form a complete question (for example, "where is the nearest shopping mall?"). Sun et al., 2018 [41] proposed a more flexible model by introducing an additional decoding mode specifically for generating the question word. This mode would be governed by a discrete, learned variable. Some experts argue that around 20% of the questions in SQuAD require paragraph-level information to reason around. This sparked more work with attention-based models that are incredibly effective at focusing on particular segments of even very long sequences, and gated LSTM networks [42].

BERT models have been used effectively by Alberti et al. [13] to generate synthetic QA pairs. The auxiliary tasks of **answer extraction, question generation (using the extracted answer) and question answering (on the generated question)** were each performed by separate BERT models trained on these tasks. Coupled with **roundtrip consistency** which ensures that noisy context-question-answer tuples are removed, they show that QA models that are fully pre-trained on QA datasets and synthetic QA pairs around those dataset, outperform those that are only fine-tuned on the QA datasets, reporting exact match and F1 scores that vary by only 0.1% and 0.4% from human performance on the SQuAD dataset.
Some works have also looked into **different forms of encoding the answer as an input feature**. Chan et al. [12] show that their BERT-HLSQG model, which **highlights the answer span within the context with special tokens**, improves on the previous state-of-the-art QAG models BLEU4 score of 16.85 by 5.32 points to 22.17. This model is able to outperform previously suggested RNN and LSTM based models.

3 Limitations of Existing Systems

3.1 Dataset

As mentioned previously, the most prominent dataset for QA and subsequently QAG tasks in the Bengali language, at the time of writing this report, is the Bangla-SQuAD dataset published by Mayeesha et. al. It relies on cloud-based translations as well as a few crowd-sourced questions. However, this dataset, too, is not without its issues. In exploring this dataset, we encountered several data quality problems:

1. **Questions that are marked impossible to answer but actually have answers:** Although the answers to these questions are not explicitly in the form that the question expects them to be in, there are still answers to these questions within the provided context. For example:

   Context: এফবিআইয়ের সদর দফতরটি ওয়াশিংটন, ডিসির জে এডগার হুভার বিল্ডিংয়ে অবস্থিত, আমেরিকা যুক্তরাষ্ট্রের বড় বড় শহরগুলিতে ৫ টি ফিল্ড অফিস রয়েছে। এফবিআই পুরো মার্কিন যুক্তরাষ্ট্র জুড়ে ৪০০ টিরও বেশি আবাসিক সংস্থা এবং পাশাপাশি মার্কিন যুক্তরাষ্ট্র দূতাবাস এবং কনসুলেটে ৫০ টিরও বেশি আইনী সংযুক্তি বজায় রাখে। ভার্জিনিয়ার কোয়ান্টিকোতে সুবিধাযুক্ত এফবিআইয়ের অনেকগুলি কার্যকারিতা, পাশাপাশি পশ্চিম ভার্জিনিয়ার ক্লার্কসবার্গে একটি "ডেটা ক্যাম্পাস" যেখানে আমেরিকা যুক্তরাষ্ট্র থেকে ৯৬৯ মিলিয়ন আঙুলের ছাপ "সংরক্ষণ করা হয়, আমেরিকান কর্তৃপক্ষের কাছ থেকে সংগ্রহ করা অন্যান্যদের সাথে সৌদি আরব এবং ইয়েমেন, ইরাক ও আফগানিস্তানের বন্দিদের। " এফবিআই তার রেকর্ডস ম্যানেজমেন্ট ডিভিশন, যা ফ্রিডম অফ ইনফরমেশন অ্যাক্টের এফওআইএ অনুরোধ প্রেরণ করে, ভার্জিনিয়ার উইনচেষ্টারে

স্থানান্তরিত করার প্রক্রিয়াধীন রয়েছে।

Question: এফবিআইয়ের ডেটা ক্যাম্পাসটি কোথায়?

This is marked as impossible, however the answer (পশ্চিম ভার্জিনিয়ার ক্লার্কসবার্গে) is present in the context.

2. **Questions that are possible to answer but have no answers listed**: These questions are not marked as impossible to answer and naturally one would expect the answers of these questions to be present in the context provided. For example:

Context: চার্লসটনের একটি আর্দ্র সাবট্রোপিকাল জলবায়ু কপেন জলবায়ু শ্রেণিবিন্যাস সিএফএ রয়েছে, সেখানে হালকা শীত, গরম, আর্দ্র গ্রীষ্ম এবং সারা বছর ধরে উল্লেখযোগ্য বৃষ্টিপাত রয়েছে। গ্রীষ্মটি সবচেয়ে আর্দ্র মৌসুম; বার্ষিক বৃষ্টিপাতের প্রায় অর্ধেকটি বজ্রপাতের আকারে জুন থেকে সেপ্টেম্বর পর্যন্ত ঘটে। পতনের নভেম্বর মাসের তুলনায় তুলনামূলকভাবে উষ্ণ থাকে। শীতকাল সংক্ষিপ্ত এবং হালকা এবং মাঝে মাঝে বৃষ্টিপাতের বৈশিষ্ট্যযুক্ত। পরিমাপযোগ্য তুষার ২০.১ সেমি বা ০.২৫ সেমি কেবল দশকে বেশিরভাগ সময়ে ঘটেছিল, সর্বশেষ ঘটনাটি ২৬ ডিসেম্বর, ২০১০-এ ঘটেছিল তবে, ১৯৮৮ সালের ৬ ডিসেম্বর ১৫ সেমি বিমানবন্দরে পড়েছিল, রেকর্ডে বৃহত্তম একক দিনের পতন, একক ঝড় এবং তু রেকর্ডে অবদান ৮.০ ২০ সেমি তুষারপাতের।

Question: চার্লস্টনের বার্ষিক বৃষ্টিপাতের অর্ধেকটি কোন আকারে ঘটে?

Answers: [“text”: “ ”, “answer_start”: 524]

This question is marked as possible to answer and the answer is within the context (বার্ষিক বৃষ্টিপাতের প্রায় অর্ধেকটি বজ্রপাতের আকারে জুন থেকে সেপ্টেম্বর পর্যন্ত ঘটে), but the answer is listed only as a sequence of whitespaces.

3. **Answers being incomplete especially at compound Bengali characters**: Because of the way the Bengali language makes use of compound characters (like for example), there is a misalignment in the answer start and end positions within the context. For example:

Context: রোমান ধর্মীয় বিশ্বাস খ্রিস্টপূর্ব ৮০০ আগে রোমের প্রতিষ্ঠা থেকে শুরু করে। যাইহোক, সাধারণত প্রজাতন্ত্র এবং প্রাথমিক সাম্রাজ্যের সাথে জড়িত রোমান ধর্ম গ্রীক সংস্কৃতির সংস্পর্শে এসে প্রায় গ্রীক ধর্মীয় বিশ্বাসকে অবলম্বন করার আগ পর্যন্ত প্রায় ৫০০ বছর অবধি শুরু হয় নি। ব্যক্তিগত ও ব্যক্তিগত উপাসনা ধর্মীয় রীতিগুলির একটি গুরুত্বপূর্ণ দিক ছিল। এক অর্থে প্রতিটি পরিবারই ছিল দেবতাদের মন্দির। প্রতিটি পরিবারের একটি বেদী ছিল ল্যারিরিয়াম, যেখানে পরিবারের সদস্যরা প্রার্থনা করতেন, আচার অনুষ্ঠান করতেন এবং পরিবারের দেবতাদের সাথে যোগাযোগ করতেন। রোমানদের উপাসনা করা অনেক দেবতা প্রোটো-ইন্দো-ইউরোপীয় প্যানথিয়ন থেকে এসেছিলেন, অন্যরা গ্রীক দেবদেবীদের উপর ভিত্তি করে তৈরি

করেছিলেন। দুটি বিখ্যাত দেবতা হলেন বৃহস্পতি রাজা শ্বর এবং মঙ্গল যুদ্ধের দেবতা। ভূমধ্যসাগরের বেশিরভাগ অঞ্চলে এর সাংস্কৃতিক প্রভাব ছড়িয়ে পড়ার সাথে সাথে রোমানরা বিদেশী দেবতাকে তাদের নিজস্ব সংস্কৃতিতে গ্রহণ করতে শুরু করেছিল, পাশাপাশি অন্যান্য দার্শনিক তিহ্য যেমন সিনিকিজম এবং স্টোইসিজমকে গ্রহণ করেছিল।

Question: সাধারণত প্রজাতন্ত্রের সাথে চিহ্নিত রোমান ধর্ম প্রথম প্রতিষ্ঠিত হয় কখন?

Answer: প্রায় ৫০০ বছর অ

The answer to this question is firstly marked incorrectly within the context, and further the answer ends abruptly where it should have been প্রায় ৫০০ বছর অবধি শুরু হয় নি.

4. **Generally wrongly marked answers**: Some answers, upon human inspection, appear to be marked incorrectly as exemplified above. These are attributable to poor translation schemes and human errors in annotating the data.

3.2 Computational Resources

A monumental challenge with most NLP tasks is circumventing the **huge amount of computational resources** required for training the incredibly large models. Most transformer-based models including the mT5 model that we have experimented with throughout this work come with **several million parameters** (BERT has 340 million parameters and mT5-small has approximately 350 million parameters; the larger models may contain more than 1 billion parameters). It is a herculean task to perform the matrix multiplications and the differential calculations involved in the backpropagation steps on simple CPU setups.

GPUs are more optimized for the large amounts of matrix multiplication and other linear algebraic manipulations involved in deep learning and in NLP. This is primarily because **GPUs are bandwidth-optimized**, meaning they can process a large amount of data fairly quickly, whereas CPUs are latency-optimized, meaning they can process very small amounts of data but they can do so incredibly quickly. With the advent of newer, advanced GPUs, they have caught up to CPUs in terms of processing speed and as such are very good computational resources to be used in deep learning tasks, where there is a large amount of data and a lot of computation is involved. **GPUs are able to parallelize operations** on data very effectively.

One must also rely very specifically on GPUs that come with **CUDA cores**, i.e. GPUs manufactured by NVIDIA, such as the RTX series of GPUs (RTX2080, RTX3080 etc.) and the Tesla P and K series GPUs.

Both the **number of CUDA cores** available as well as the **VRAM** of the GPU model used in the deep learning project are important factors to consider. While the CUDA cores are respon-

sible for the computations involved, the entire model with its million or billion parameters has to be loaded entirely on the VRAM of the GPU and as such commonly **VRAM requirements exceed 6GB**.

These requirements make it very difficult to train large language models on personal computers. Commonly **Google Colaboratory** and **Kaggle**, both interactive Python notebook platforms provided by Google, are used. The free tier of both platforms provide 12GB of RAM, a Haswell-family 2-core CPU running at 2.3GHz and usually a Tesla K80 GPU (although available for a maximum of 12 hours of continuous usage; there is also an indefinite cooldown period of no GPU allocation once the GPU has been overused by a user). The paid tier offers a slightly better GPU (Tesla P100 with up to 12GB of VRAM), but up to 24GB of RAM and a longer period of continuous GPU usage.

We found it best to perform several experiments on **small subsets of the data** available to us, often taking **50% to 80%** of the data for the entirety of the training-testing phase. Because of the limitations of the VRAM, we were also forced to perform the training in **smaller batch sizes of up to 8 samples per batch**; this translated to very long training times of as much as **8 to 10 hours** even on the smaller subset of the training data. We are confident that given more resources and more time to fine-tune our experimental models as well as our final model, we can converge to better results. That being said, the results produced using the provided setup are strongly representative of the potential this QAG system has.

3.3 Tokenizer Issues

When trying to align answers post translation to find the starting index of the answer span, we quickly found that the **embeddings generated were often inaccurate**, leading to **improper alignments**. Upon further investigation we noticed that this was primarily because of the way the Bengali tokenizers were splitting the tokens. We experimented with several Bengali tokenizers, including BNLP Sentence Tokenizer, BNLTK Tokenizer etc. and found this issue persistent in all our experiments.

We found that Bengali tokenizers are **not particularly adept at handling compound characters**, (like ক্ষ for example) **and punctuations** (especially when detokenizing back to the actual words in the sentence). This becomes a very prominent problem that requires immediate solutions when one realizes the extent to which punctuation and compound characters are used in the Bengali language.

To that end, we built a custom tokenizer by extending existing tokenizers with **several conditional regular expression (regex) clauses** to handle these cases. This custom tokenizer is

able to properly tokenize and detokenize contexts and tokens respectively.

4 Proposal and Contributions

4.1 Proposal

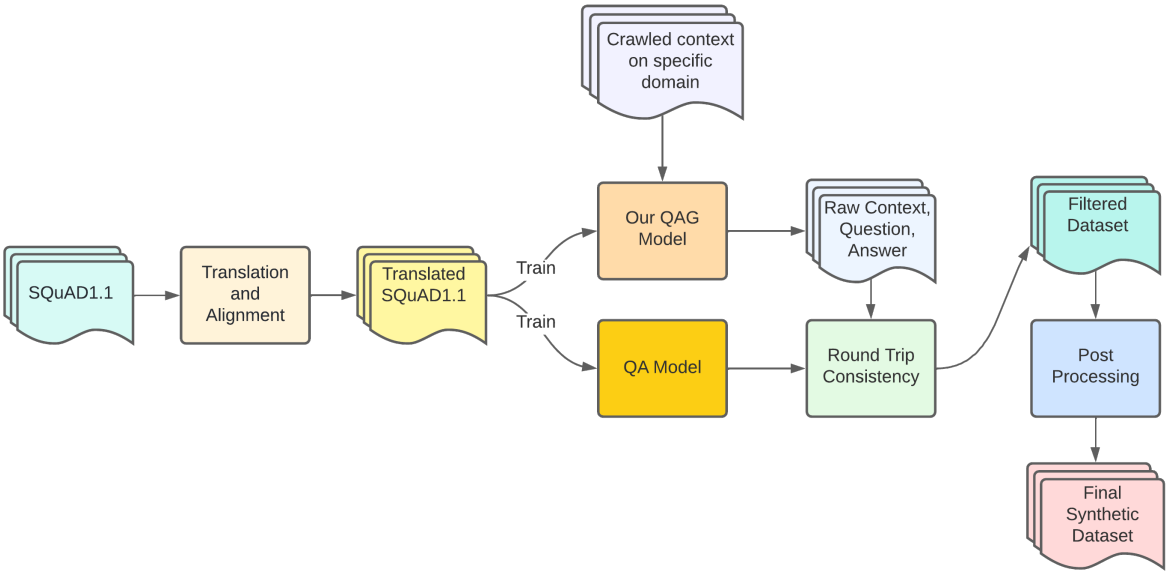


Figure 5: Overall workflow of our proposed methodology

The diagram 5 outlines our proposed workflow. We split the entire process into two stages - **training** and **synthetic data generation** (or inference).

In the **training phase**, we **translate the SQuAD1.1 dataset** via a **translation model** and a novel **embedding based alignment approach** (discussed later). We use that translated dataset to train a QAG model (discussed in section 4.1.2) as well as a **QA model for roundtrip consistency filtering** (discussed in section 4.1.6). The entire process is summarized in diagram 6:

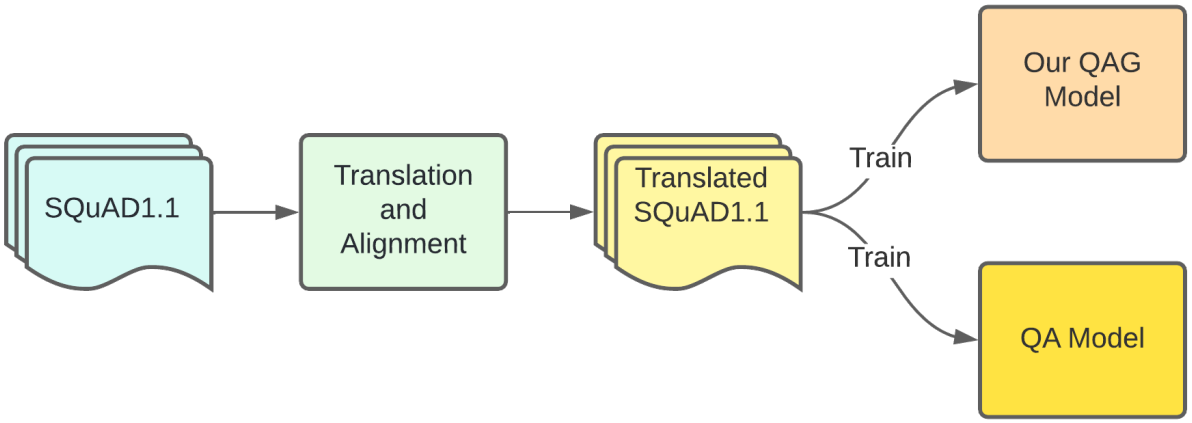


Figure 6: Workflow of the translation and training phase

In the **synthetic data generation (inference) phase** (diagram 7), our trained QAG model is fed **domain-specific contexts** (that have been scraped from the internet or obtained from other sources) as input. The model processes these and **produces questions using the question generation model** from the a**nswers provided by the answer extraction model** to form QA pairs. The trained QA model is then fed these contexts and questions to predict answers. If the answers are a match or within 0.5 F1 of each other, the **QA pair is retained as a valid QA pair**; otherwise the QA pair is discarded. Afterwards, to **filter out similar questions**, we retain only the **QA pair that shows the highest probability of generating that answer** (measured by the sum of the start and end positional logit score).

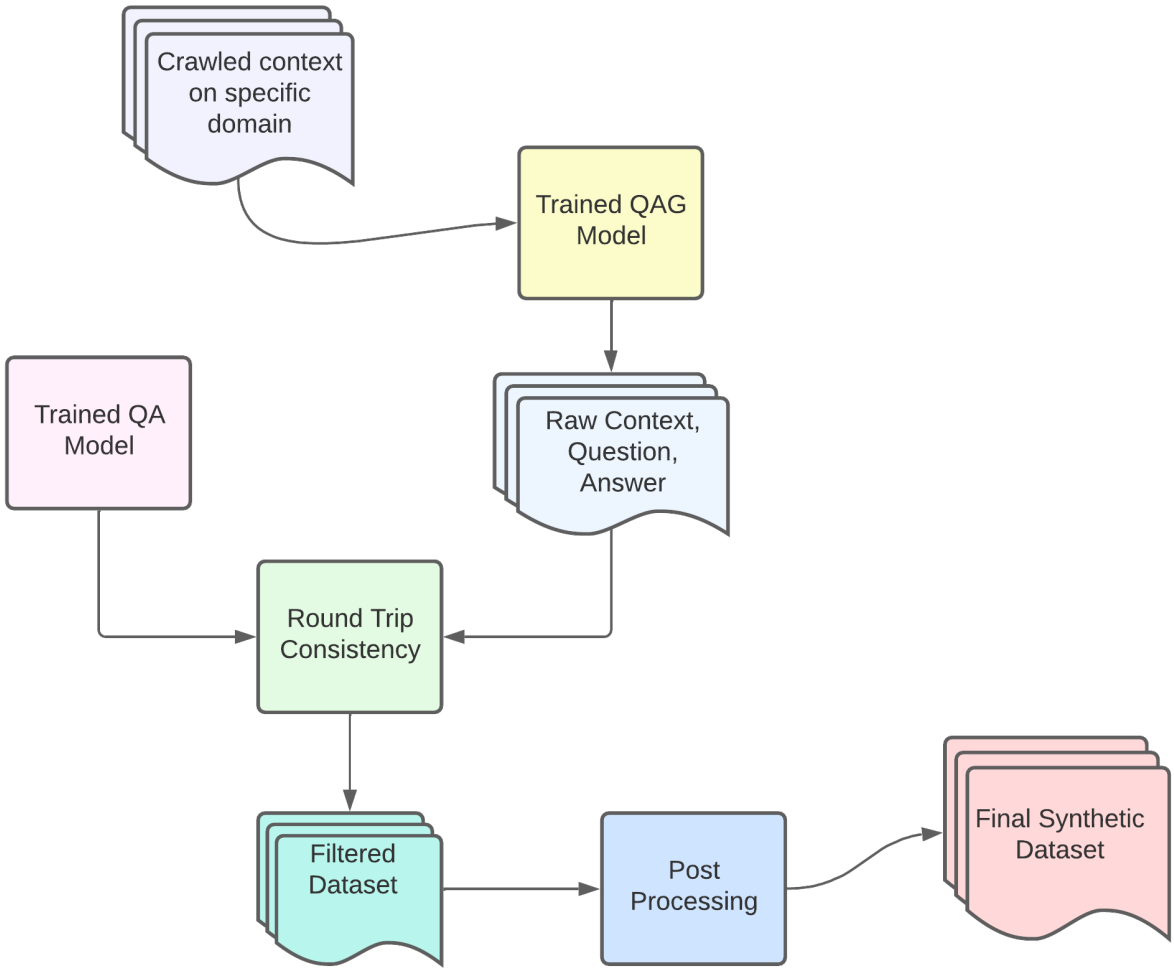


Figure 7: Workflow of the inference phase

At the end of this, the roundtrip consistency filtered dataset is post processed to bring it to a format similar to that of established datasets like SQuAD, NewsQA etc. and the final synthetic dataset is then used for augmenting the training and testing data for QA models as required.

4.1.1 Translation and Alignment

To circumvent the above issues, we translated the original **SQuAD1.1 dataset into Bengali** using the **M2M100 model released by Facebook in 2020**. This is a **1.2 billion parameter multilingual encoder-decoder translation model** that has shown significantly better performance at the task of machine translation than other models by as much as 10.2 BLEU at the time of writing this report. We choose SQuAD1.1 to translate into Bengali and not SQuAD2.0 because SQuAD2.0 has "impossible questions" (questions that cannot be answered) and these would have to be discarded since we are not generating such questions in this scope of work. We then used **FastText word embeddings** and the concept of **cosine similarity**

to locate the answer span in the translated Bengali context.

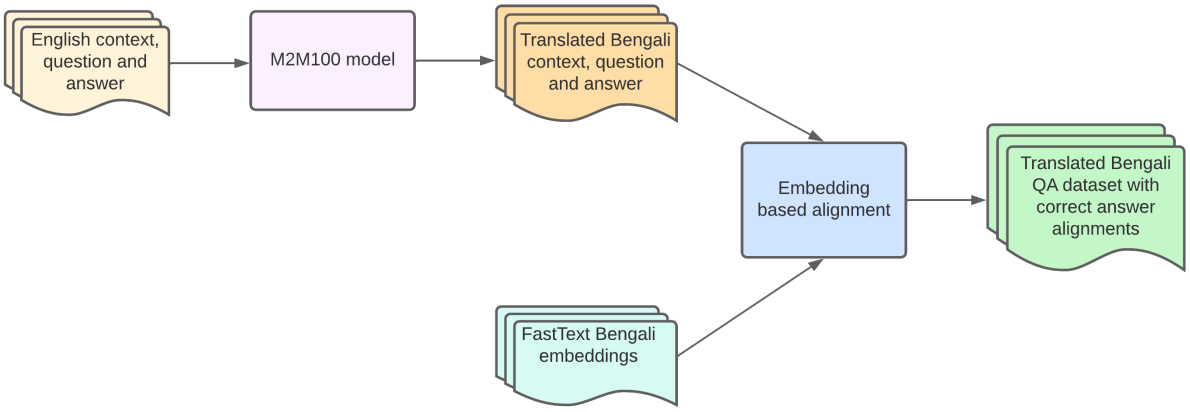


Figure 8: Workflow of the translation process

The procedure followed was:

1. **Translating English contexts, questions and answers into Bengali:** We first **tokenize** the English context into individual sentences using the sentence tokenizer from the Natural Language Toolkit (nltk) library. While doing so, we also mark the sentence with the answers that they contain.

2. **Calculating cosine similarities of word embeddings to find the answer span in Bengali context:** Before identifying and highlighting the answer span, we look at the embeddings of the words in Bengali answer text and the context sentence that contains the answer text. For this purpose we used the **FastText** Bengali word embeddings.
   We use the embeddings to find the part of the Bengali context that contains the answer. For a window of length equal to the translated answer text in the context we find the **sum of cosine similarity** between the words from that window and the permutation of answer words which gives the maximum sum. We then select the answer span from the context by selecting the window that gives the maximum alignment score, which is found by normalizing the sum of the cosine similarities.
   However, we identified that after translation, the Bengali answer **can contain more words than the English sentence it was translated from**. So, we also used windows of size

   (a) **One word more** than the length of the translated answer text

   (b) **Two words more** than the length of the translated answer text

   In these cases, when we measure the sum of cosine similarity with a permutation of answer

tokens, we do not use a word in the window of the context sentence (in case a) or we do not use two words in the window of the context sentence(in case b)

In post processing, we select only those question-answer pairs with an **alignment score of greater than or equal to 0.5 based on our methodology**. Since the SQuAD test set is not publicly available, we use the SQuAD dev set as the test set and split the SQuAD train set for our train and dev sets as shown below.
An example of the model translating an English context into Bengali is shown below:

English context: Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

Our translated Bengali context: স্কুলটিতে ক্যাথলিক চরিত্র রয়েছে। আটপ প্রধান বিল্ডিংয়ের স্বর্ণের ঘোড়াটি মেরির একটি স্বর্ণের মূর্তি। প্রধান ভবনের সামনে এবং তার মুখোমুখি, **"Venite Ad Me Omnes"** এর সাথে সাজানো অস্ত্রের সাথে খ্রীষ্টের একটি তামা মূর্তি রয়েছে। প্রধান ভবনের পাশে রয়েছে পবিত্র হৃদয়ের বেসিলিকা। বেসিলিক্সের পেছনে রয়েছে গুহা, নামায ও বিবেচনার একটি মারিয়ানা স্থান। এটি ফ্রান্সের লুরডের গুহাটির একটি প্রতিফলন, যেখানে পবিত্র বার্ন্যাডেট সুবিরুসের কাছে ১৮৫৮ সালে পবিত্র মেরি প্রদর্শিত হন। প্রধান ড্রাইভের শেষে (এবং একটি সরাসরি লাইনে যা ৩ মূর্তি এবং গোল্ড ডোমের মাধ্যমে সংযুক্ত করে), মেরির একটি সহজ, আধুনিক পাথর মূর্তি।

We find here that the model is incredibly accurate in generating a translation.
On the issue of locating answer spans, we make use of our algorithm outlined above. A few example question answer pairs for the above context in English and our translated version of it is given below:

Reference sentence from Bengali translation: আটপ প্রধান বিল্ডিংয়ের স্বর্ণের ঘোড়াটি মেরির একটি স্বর্ণের মূর্তি। প্রধান ভবনের সামনে এবং তার মুখোমুখি, **"Venite Ad Me Omnes"** এর সাথে সাজানো অস্ত্রের সাথে খ্রীষ্টের একটি তামা মূর্তি রয়েছে।

Translated Bengali answer: খ্রীষ্টের একটি কপার মূর্তি রয়েছে।

However, we find that this particular text span is not found in the translated Bengali context.

In particular the word "কপার" is nowhere to be found in the reference sentence; the translation model translated this word into "তামা. As a result of this, the **algorithm looks through the entire context, comparing word spans to find the closest match and extracts the following span**.

Extracted answer span from the context: খ্রীষ্টের একটি তামা মূর্তি

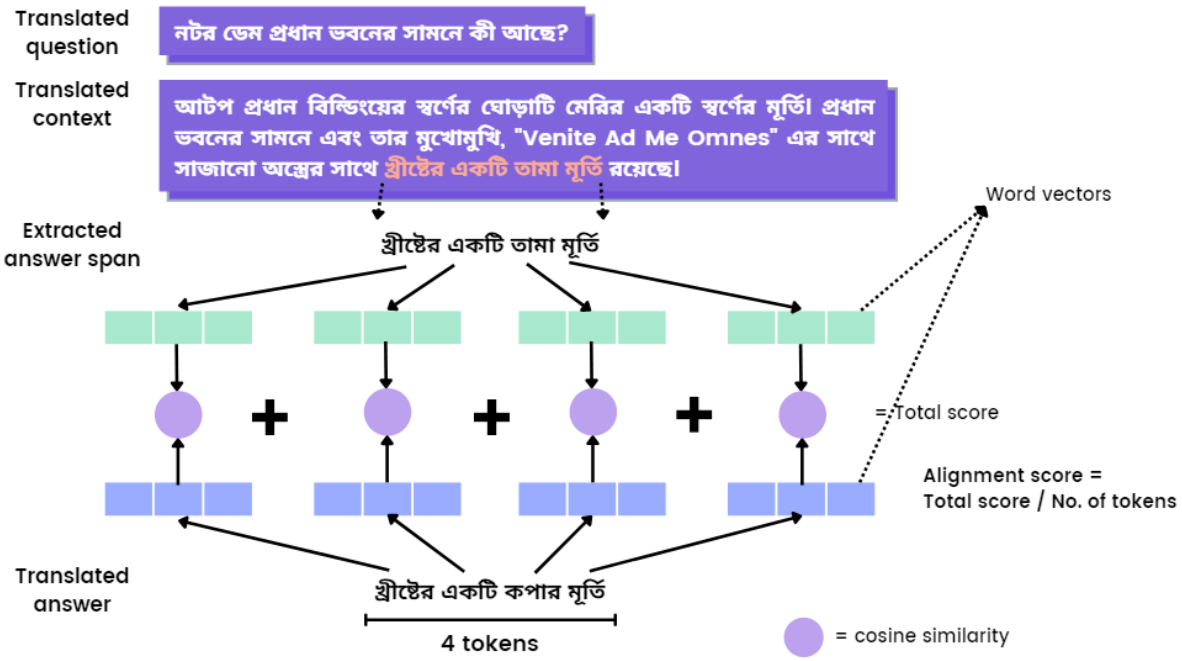


Figure 9: Extracting answer span from the translated context where similar (but not the same) word is found present

The following example outlines how the algorithm handles cases where the translated context does not have the same sequence of words found in the translated Bengali answer. Instead they are found **rearranged** in the translated context.

Reference sentence from Bengali translation: যেমনটি তার সাম্রাজ্যিক আদেশে প্রমাণিত, হংওয়ু সম্রাট তিব্বত ও চীনের মধ্যে বৌদ্ধ সম্পর্কের ব্যাপারে সচেতন ছিলেন এবং এটি গড়ে তুলতে চেয়েছিলেন।

Translated Bengali answer: চীন ও তিব্বতের মধ্যে বৌদ্ধ সম্পর্ক

The **algorithm looks through the translated context and finds that the following span is the closest match to the translated Bengali answer**.

Extracted answer span from the context: তিব্বত ও চীনের মধ্যে বৌদ্ধ সম্পর্কের

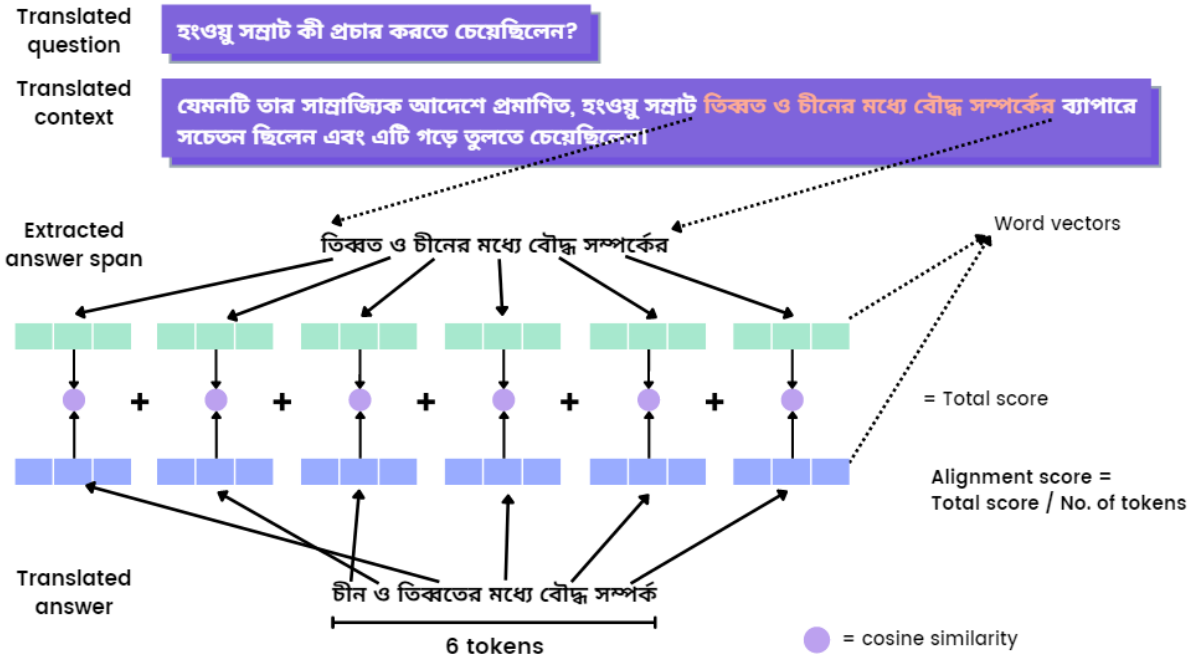


Figure 10: Extracting answer span from the translated context when the sequence of words (in the translated answer) is rearranged in the translated context

4.1.2 Question Answer Generation Model

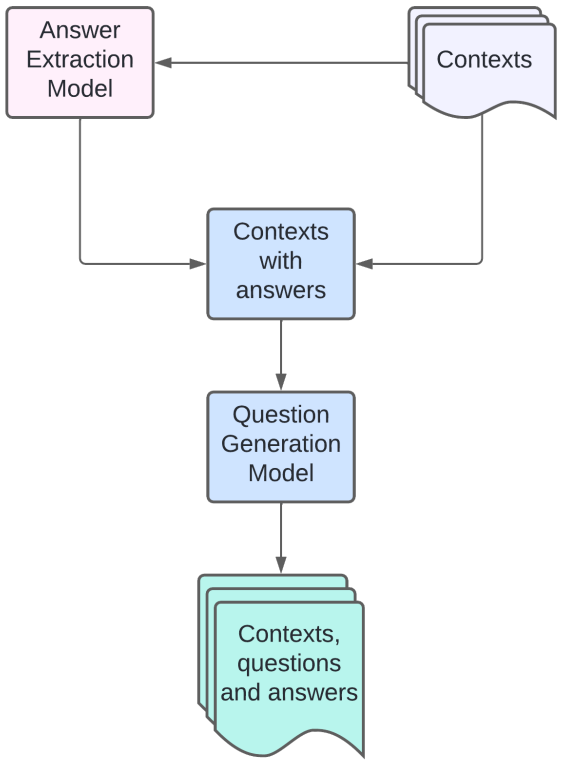


Figure 11: Workflow of our proposed Question Answer Generation model

The question answer generation model is comprised of 2 components:

1. **Answer extraction:** This component examines the context for interesting answer spans and then these are extracted. These interesting answer spans may be identified by a variety of methods, including **named entity recognition (NER)**, **extractive text summarization**, **sequence-to-sequence generation** etc. We experiment with NER and sequence-to-sequence generation using highlighting in our work.

2. **Question generation:** The question generation component takes the **answers extracted by the previous component** and the **context provided** to **generate questions**, and, subsequently, **QA pairs**. As discussed previously, multiple forms of question generation have been explored, including the use of **hierarchical variational autoencoders** [40], **regular sequence-to-sequence generation** etc. Of these, we experiment mainly with regular sequence-to-sequence generation and **different forms of highlighting** and **answer appending** as per reviewed literature.

4.1.3 Answer extraction using NER

Named Entity Recognition (NER), as the name suggests, is a technique where a pretrained model is able to identify **different named entities** from a given text input. These entities may be locations (LOC), persons (PER), organizations (ORG) etc. In Bengali, there have been recent advancements made in NER using deep learning techniques [43]. We make use of the **SUST Bangla Natural Language Toolkit (SBLTK)** published online and its **BERT-Multilingual-Uncased NER model** because of its promising accuracy of 90.5% and its huge training corpus of more than 65,000 documents.
We split the context into sentences and for each sentence we identify the named entities present using the NER model discussed above. We retain all named entities for this purpose.
An example of the NER model is shown in figure 12:

```
{'answers': [['সাউথাম্পটন', 'যুক্তরাজ্যের'],
             ['ইংল্যান্ডের', 'গাড়ীটি'],
             ['মোটরসাইকেল',
              'লন্ডন',
              'উইনচেস্টারে',
              'মিডল্যান্ডস',
              'উত্তর'],
             ['গাড়িটি', 'হাতিয়ার']],
 'context': ['সাউথাম্পটন একটি গুরুত্বপূর্ণ যুক্তরাজ্যের বন্দর যা '
             'দেশের বাকি অংশের সাথে ভাল পরিবহন সংযোগ আছে।',
             'ইংল্যান্ডের দক্ষিণাঞ্চলীয় উপকূলে স্থানগুলি সংযুক্ত '
             'করে M২৭ গাড়ীটি শহরের উত্তর দিকে চলে যায়।',
             'M৩ মোটরসাইকেল লন্ডন এবং এছাড়াও, একটি লিঙ্ক মাধ্যমে '
             'A৩৪ (ইউরোপীয় রুটের অংশ E০৫) উইনচেস্টারে, মিডল্যান্ডস '
             'এবং উত্তর সঙ্গে।',
             'এম ২৭১ গাড়িটি এম ২৭ এর একটি হাতিয়ার, যা এটি ওয়েস্ট '
             'ডকস এবং শহরের কেন্দ্রের সাথে সংযুক্ত করে।']},
```

Figure 12: Example of Answer Extraction using the NER model

4.1.4 Answer extraction using mT5 and highlighting

We also tried to **extract answers from a given context** by modeling it as a **sequence-to-sequence task**. We use the mT5-small model for this task. **mT5-small** is the smaller, multilingual variant of the T5 [11] model and has around **300 million parameters**, trained on a **huge mC4 corpus** and thus **supporting 101 languages including Bengali**. We chose this over BERT since other sequence-to-sequence tasks in the Bengali language that we experimented with outside of this work showed better, more **human readable output** with mT5 than with BERT.

We take contexts from the our translated dataset and split them into individual sentences using **bnltk**. We highlight each sentence that has an answer span within it with a special <hl> tag, and the answer spans are joined with the special <sep> token. An example is shown below:

Context:

The Islamic University of Technology is situated in Boardbazar, Gazipur and currently has 6 departments CSE, MPE, EEE, BTM, CEE and TVE.

Input sequence:

<hl>The Islamic University of Technology is situated in Boardbazar, Gazipur and currently has 6 departments CSE, MPE, EEE, BTM, CEE and TVE.<hl>

Target sequence (answer span):

CSE, MPE, EEE, BTM, CEE and TVE<sep>

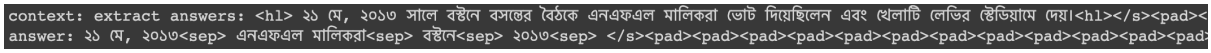


Figure 13: Highlighted inputs to the answer extraction mT5 model

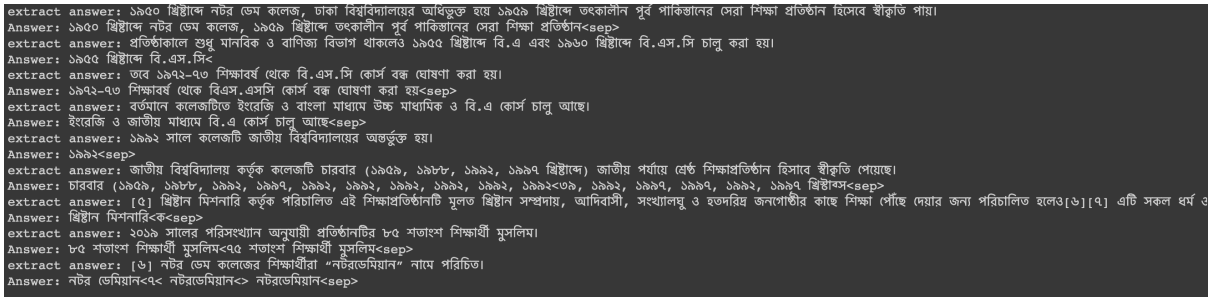


Figure 14: Examples of Answers generated by the mT5 model

We train the model on the training set of our translated dataset over:

- **50% of the training data** - We are forced to use a small subset of the actual translated dataset due to computational resource constraints.
- **3 epochs** - The literature reviewed specifies 3 to 4 epochs. We choose 3 epochs due to computational resource constraints and to prevent overfitting.
- **Batch size 8** - Given the GPU and the amount of VRAM, we can only fit 8 samples of the data at any given instance, and so we pick a batch size of 8. This allows us to perform mini batch gradient descent which is a middle ground between stochastic gradient (which performs gradient descent after every sample encountered) and batch gradient descent (which performs gradient descent after all the samples are encountered).
- **Learning rate of 3e-4** - Although the literature reviewed specifies a learning rate of 1e-4, we find in our experiments that the training loss converges neatly with slight fluctuations at a learning rate of 3e-4 while also learning the parameter values sufficiently quickly.

4.1.5 Question Generation via appending

We use the **mT5-small** model for conditionally generating questions based on the context and the answer. The rationale behind using this model is the same as in the case of the answer extraction model experiments - mT5 is trained on a huge mC4 corpus and supports and provides excellent human readable results in the Bengali language.

We use the **conditional generation head for mT5**. The model takes a **source text as input and generates a target text** based on the source text. We format the source text as follows:

context: c answer: a, where c is the context and a is an answer found in that context.

For the target text, we feed the model the question corresponding to that answer during training time. However, the model performed poorly and **generated undecipherable questions** when we fed the entire context as source text. To circumvent that, we opted to pass as input to the model **the answer and only the sentence from the context that contains the answer**. Although this method of training restricts the model to generate questions based on single sentence inputs, the model performs considerably well and the questions generated were much more readable and understandable.
Example of the generated questions are shown below:

context: প্রথম বছরের অধ্যয়ন প্রোগ্রামটি ১৯৬২ সালে প্রতিষ্ঠিত হয়েছিল যাতে তারা একটি প্রধান ঘোষণা করার আগে স্কুলে তাদের প্রথম বছরে প্রবেশকারী নবজাতকদের নেতৃত্ব দেয়। **answer:** প্রথম বছরের অধ্যয়ন প্রোগ্রামটি

১৯৬২ সালে কোন প্রোগ্রামটি প্রতিষ্ঠিত হয়েছিল**?**

To ensure the generated question is semantically correct, we generate **five questions for each answer using beam search** and later **filter out incorrect questions during roundtrip consistency** described later. An example is shown below with all the five questions:

context: প্রথম বছরের অধ্যয়ন প্রোগ্রামটি ১৯৬২ সালে প্রতিষ্ঠিত হয়েছিল যাতে তারা একটি প্রধান ঘোষণা করার আগে স্কুলে তাদের প্রথম বছরে প্রবেশকারী নবজাতকদের নেতৃত্ব দেয়। **answer:** প্রথম বছরের অধ্যয়ন প্রোগ্রামটি

১৯৬২ সালে কোন প্রোগ্রামটি প্রতিষ্ঠিত হয়েছিল**?**
১৯৬২ সালে প্রতিষ্ঠিত প্রথম বছরের অধ্যয়ন প্রোগ্রামটি কী ছিল**?**
১৯৬২ সালে প্রথম বছরের অধ্যয়ন প্রোগ্রামটি কী ছিল**?**
১৯৬২ সালে প্রতিষ্ঠিত প্রথম বছরের অধ্যয়ন প্রোগ্রাম কোনটি ছিল**?**
১৯৬২ সালে প্রতিষ্ঠিত প্রথম বছরের অধ্যয়ন প্রোগ্রাম কোনটি**?**

To train the QG model we use our translated dataset based on SQuAD1.1. We only use answers that have an alignment score of greater than or equal to 0.5. We train the model over

- **2 epochs** - In accordance with computational resource limitations and reviewed literature to **prevent overfitting**.
- **Learning rate of 1e-3** - Here we find that a learning rate of 1e-3 allows the loss to converge sufficiently quickly while also allowing the model parameters to be learned fairly consistently.

- **Batch size of 8** - Coping with the same computational resource limitations mentioned in the answer extraction experiment section.

4.1.6 Roundtrip Consistency

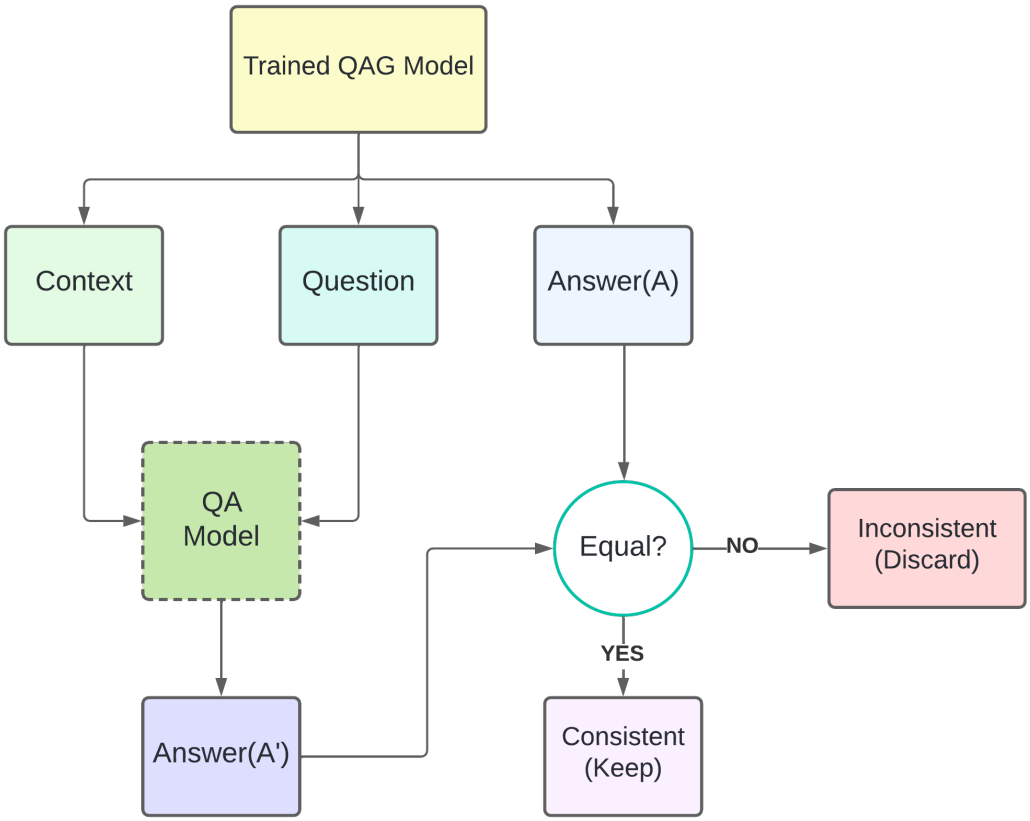


Figure 15: Workflow of the Roundtrip Consistency phase

This phase makes use of a **pre-trained QA model** to **independently predict answers to the questions generated by the previous component**. The QAG model generates multiple questions for the same answer. The trained QA model checks each of the questions and predicts answers for each of them. If the QA models predicted answers and the QAG models provided answers are an **exact match or within a 0.5 F1 threshold of each other**, the **QA pair is retained** in the filtered dataset; otherwise it is discarded.

Even so, there may be similar questions for the same answer, as exemplified in figure 16:

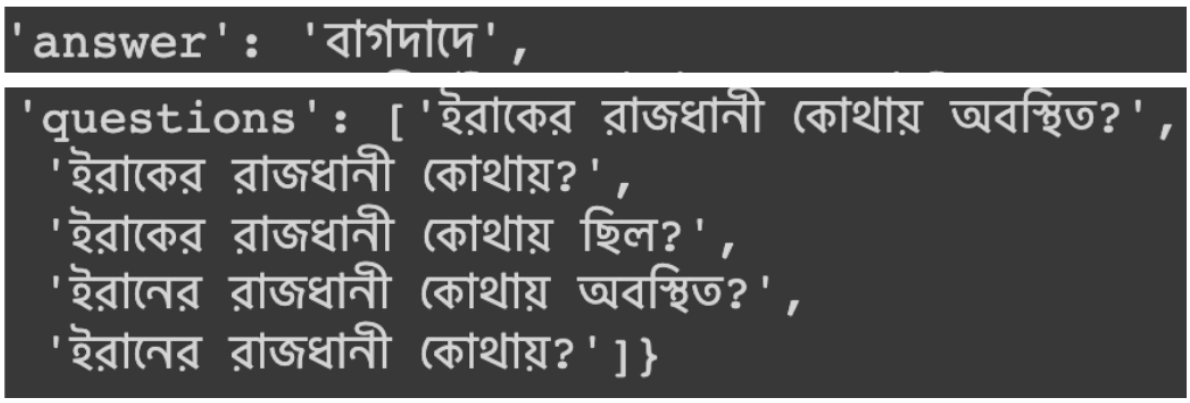


Figure 16: Similar questions for the same answer

To filter out similar questions, we retain only the question that **produces the answer with the highest probability**. This is obtained by taking the sum of the **answer span's start and end positional logit scores**.

4.1.7 Evaluation Metrics

Exact Match (EM) is a **measure of the proportion of predicted answers that match any of the ground truth answers exactly**. That is to say, if one of the ground truth answers is "Islamic University of Technology", the only predicted answer that would be an exact match is "Islamic University of Technology", and not any of "Islamic University", "University of Technology" etc. As such, it is an **incredibly strict measure** of a QA models performance. It is to be noted that for null ground truth answers (blank answers particularly found in the case of "impossible questions"), if the model predicts any answer at all, the EM score for that sample is taken to be 0.
It is normally calculated as per the following formula

$$EM = \frac{\sum_{i=1}^{N} F(x_i)}{N}$$

$$Where\ f(x_i) = \begin{cases} 1, & \text{if predicted answer = correct answer} \\ 0, & \text{otherwise} \end{cases}$$

F1 score is a **less strict measure of the performance of a QA model**. It is calculated as the **harmonic mean of the precision and recall**. In the case of sequence-to-sequence generation NLP tasks, the ground truth tokens and the predicted output tokens are treated as bags of words. As such precision, recall and F1 scores are calculated as per the following formulas.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

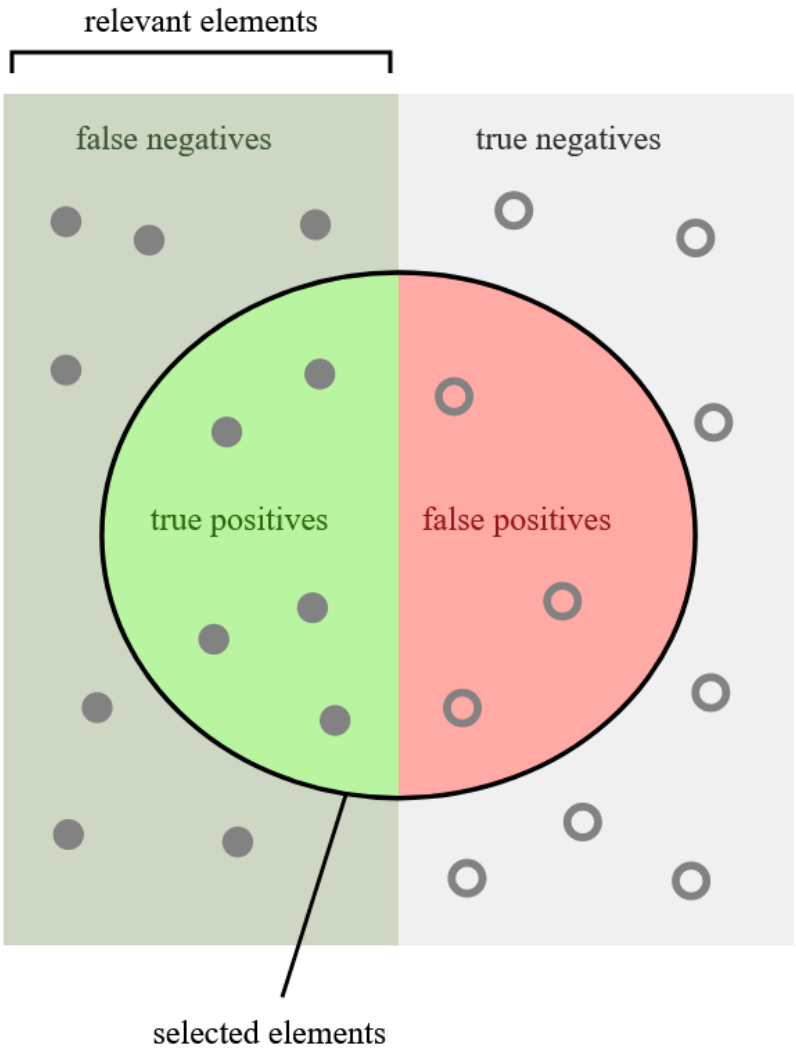
$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

This formula can also be equivalently written as,

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

F1 scores are important metrics when we want to maximize both precision and recall (approaching an exact match in the ground truth answer and the predicted answer). Precision is a measure of how many of the predicted answers were correctly predicted as per the ground truth, whereas recall is a measure of how many correct ground truth answers were present in the models predictions.

Put simply, if there are **g tokens in the ground truth** answers, **p tokens in the predicted answers** and **c tokens that are common to both ground answers and predicted answers**, the model should aim to maximize both the ratio $\frac{\mathbf{c}}{\mathbf{p}}$ (most of the models predicted output tokens overlaps with the ground truth answers tokens) and $\frac{\mathbf{c}}{\mathbf{g}}$ (most of the ground truth tokens are present in the models predicted output).



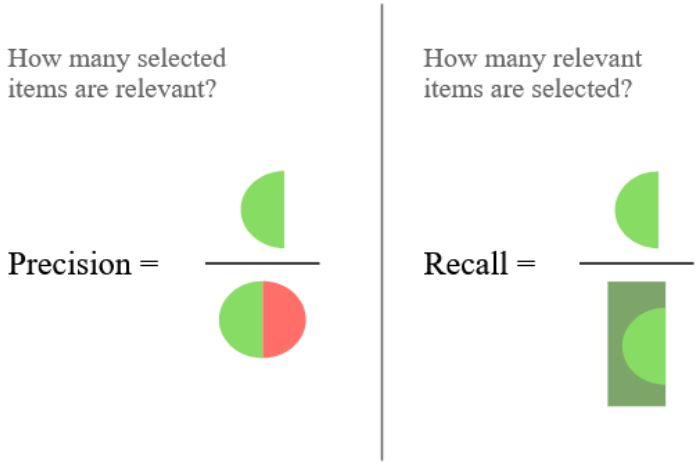


Figure 17: Precision, Recall and F1 explained graphically

4.2 Contributions

Major contributions of our work include:

1. **QA dataset (75,000+ QA pairs) based on translated SQuAD1.1** - This is the result of our translation methodology applied on the existing SQuAD1.1 dataset. The resulting dataset has been used extensively throughout our work.

2. **QA model benchmarks trained on the translated dataset** - To establish naive benchmarks, we also train and test basic QA models on our dataset and post the results here for comparisons.

3. **Low-resource adapted QAG model in Bengali** - Our QAG model shows great promise even in low computational resource scenarios, requiring training over comparatively small subsets of data over a respectable period of time to generate domain-specific synthetic QA datasets.

4. **Synthetic QA dataset in Bengali** - We test our approach on a sampling of news articles and present a few examples of the generated dataset, to judge the quality of the output.

5 Result Analysis and Discussion

5.1 Translation and QA model training

As mentioned previously, for our dataset, we select only those QA pairs with an **alignment score of greater than or equal to 0.5** based on our methodology. Since the SQuAD test set is not publicly available, we use the **SQuAD dev set as the test set** and **split the SQuAD train set for our train and dev sets** as shown in table 1 to release as our final dataset based on the SQuAD1.1.

Train		**Dev**		**Test**	
Paragraphs 400	Contexts 17126	Paragraphs 42	Contexts 1770	Paragraphs 48	Contexts 2067
	QA 63628		QA 6569		QA 8233

Table 1: Train-test splits on SQuAD dataset for our translated dataset

Model Name	Exact Match	F1
bert-base-multilingual-cased	44.67	64.02
bert-base-multilingual-uncased	44.52	64.10
XLM-RoBERTa	46.70	66.37

Table 2: EM and F1 scores of 3 models on our translated dataset

The dataset consists of **75000+ QA pairs**. We train 3 QA models on the dataset and evaluate on our test set. The results are shown in the following table:

Due to the scarcity of models pre-trained in the Bengali language, we show the score for only 3 models **pre-trained on multilingual datasets**. The BERT models both cased and uncased, perform similarly in Bengali with F1 scores 64.02 and 64.10 while the XLM-RoBERTa performs better with the highest F1 score of 66.37. While these scores may seem poor in the scale or leaderboard of SQuAD1.1, we need to realize that **this is a Bengali dataset** and **Bengali transformer-based models are scarce and not very well-trained**. Further, there are **issues with tokenization** as discussed previously, and of course, there may be a **failure to incorporate certain linguistic biases in a direct translated dataset** as opposed to an organically generated Bengali dataset. Moreover, these are very **simple models** and **no degree of ensembling** or **intensive training over multiple epochs** has been used yet. We believe these techniques can **improve the EM and F1 scores** further.
We also **fine-tuned the XLM-RoBERTa model on our translated dataset** and **tested it on the Bangla-SQuAD test dataset** achieving **much higher EM and F1 scores of 70 and 81 respectively**. This is a testament to the **consistency and superior quality of our version of the dataset**.

5.2 Synthetic Dataset Generated by QAG model

To show the effectiveness of our overall QAG model, we scrape **314 articles from a Bengali newspaper** and generate a total of **4582 QA pairs**. A portion of the dataset is shown in figure 18:

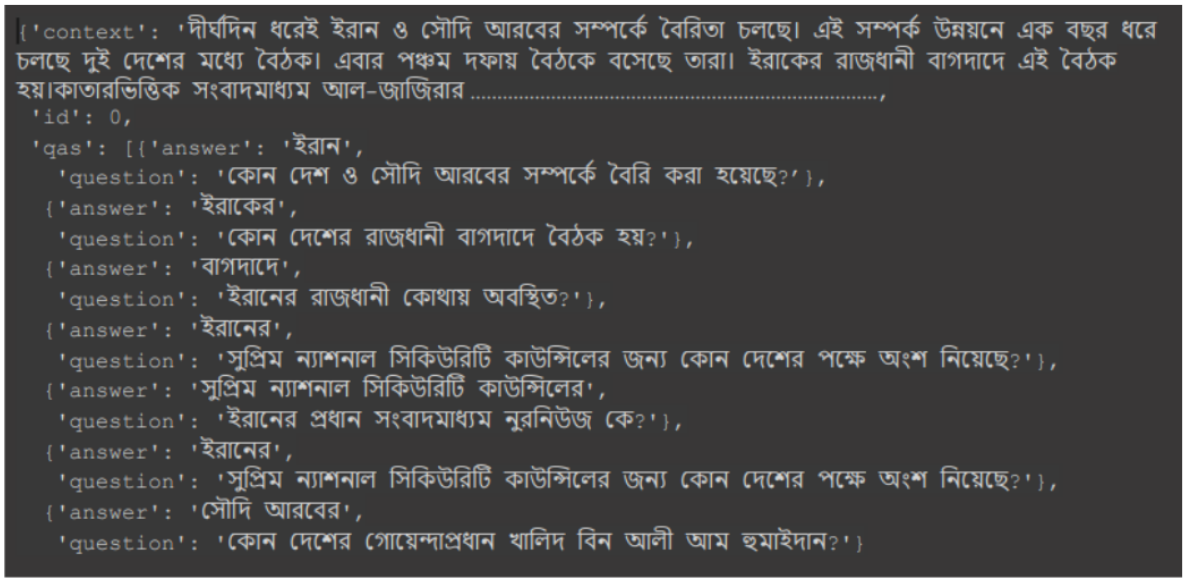


Figure 18: Generated synthetic dataset on a sample sentence from a news article

The generated question answer pairs from the new articles are quite consistent with a small percentage of errors. However a major issue with this dataset is the fact that the **answers are all named entities**. Interesting text spans that do not contain any named entities are ignored by the model. For example an interesting answer span could be **পঞ্চম দফা** could be an interesting answer to the question **কত দফায় বৈঠকে বসেছে তারা?** However, this will be overlooked by the NER model.

To circumvent this issue, we experimented with the **highlighting based answer extraction model outlined above on both the Bangla-SQuAD dataset and our translated dataset**. The results generated are shown in figure 19 and 20:

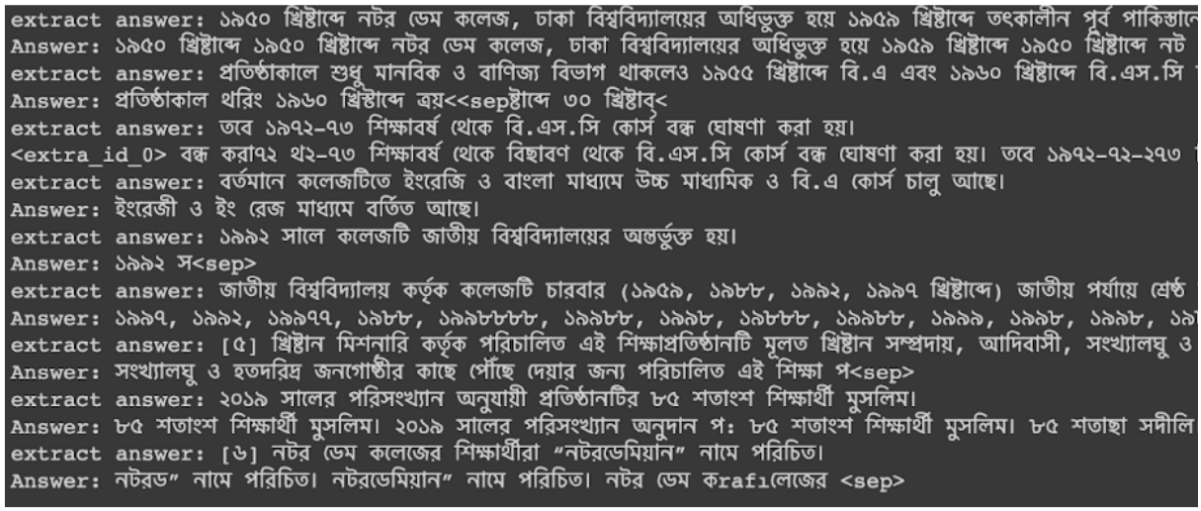


Figure 19: Answer Extraction model trained on Bangla-SQuAD

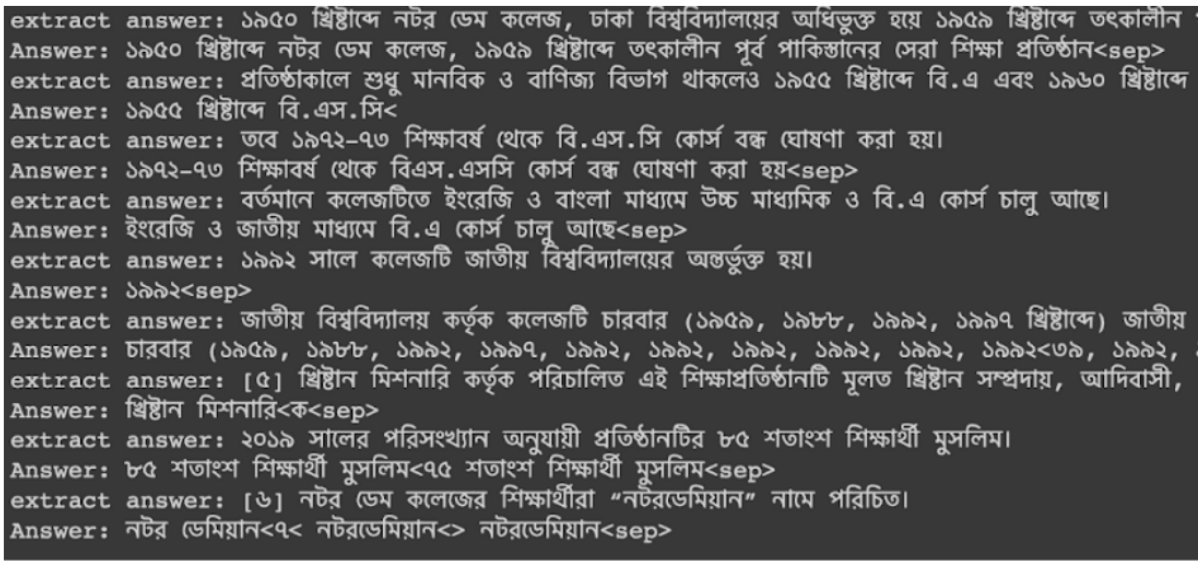


Figure 20: Answer Extraction model trained on our translated dataset

As can be clearly seen in figure 20, the **answer extraction model trained on the Bangla-SQuAD dataset generates very undecipherable answers** whereas that trained on **our translated dataset generates more readable answers**. This is primarily because of the way **the answer starting and ending positions are aligned** in our dataset and **also because of the superior quality of our translation using the M2M100 model** as opposed to the one used to produce Bangla-SQuAD.

However, this model still produces **repeated answers** as can be seen in the last example. We believe that training this model over a few more epochs and experimenting with the **encoding truncation settings**, we can reach a very optimum solution where the model does not generate repetitive answers. We can also **experiment with other forms of highlighting mentioned** in the reviewed literature, for example, **highlighting the answer span within the sentence** as well.

6 Conclusion and Future Work

In this work, we present **a translated version of SQuAD1.1 with answer starting and ending positions being aligned based on a novel embedding matching approach**. We also propose an **end-to-end QAG system** to **automate the generation of QA pairs** and, subsequently, synthetic QA datasets for the Bengali language. We show some samples of our model's output and also discuss the limitations of the work as it exists right now. The limitations are **focused mainly in the answer extraction portion of the QAG system since the answers extracted are solely based on named entity recognition**. Another

obvious constraint is the one posed by the **lack of computational resources** available to us at the time of writing this report.
To build on the work further and to produce better results, we leave the following scope for future work:

1. **Translate SQuAD2.0** - SQuAD2.0 builds on SQuAD1.1 by adding 'impossible' questions. The authors of the paper show that it is also important to teach QA models which questions they cannot answer. The SQuAD2.0 dataset translated via our methods could prove to be a better resource than the current translated dataset and Bangla-SQuAD as well.

2. **Search for a better Answer Extraction model**, preferably one using a sequence-to-sequence conditional generational setting as opposed to NER or other extractive methods. This will allow us to encompass a wider range of answer types, possibly True/False and/or Yes/No answers as well.

3. **Publish synthetic datasets on specific domains** - Due to the lack of quality datasets specific to particular domains such as medicinal science, news articles etc. there are no QA models in Bengali that can tackle these domains' questions easily. Via our work, we hope to spearhead dataset generation in these specific domains aided by our QAG system.

4. **Fine-tune QA models on synthetic datasets** to perform comparative analysis.

5. **Improve the answer alignment technique** to reduce noisy answer spans in our translated datasets, preferably by looking at other similarity metrics.

References

[1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 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), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[3] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] 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.

[5] 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), pp. 2383–2392, Association for Computational Linguistics, Nov. 2016.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019.

[8] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910, 2018.

[9] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natu-*

ral Language Processing, (Lisbon, Portugal), pp. 1412–1421, Association for Computational Linguistics, Sept. 2015.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019.

[12] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, (Hong Kong, China), pp. 154–162, Association for Computational Linguistics, Nov. 2019.

[13] C. Alberti, D. Andor, E. Pitler, J. Devlin, and M. Collins, "Synthetic QA corpora generation with roundtrip consistency," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 6168–6173, Association for Computational Linguistics, July 2019.

[14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 483–498, Association for Computational Linguistics, June 2021.

[15] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (Cambridge, MA, USA), p. 16931701, MIT Press, 2015.

[16] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), Association for Computational Linguistics, Sept. 2017.

[17] 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), pp. 784–789, Association for Computational Linguistics, July 2018.

[18] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, Mar. 2019.

[19] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, Mar. 2019.

[20] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "NewsQA: A machine comprehension dataset," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, (Vancouver, Canada), pp. 191–200, Association for Computational Linguistics, Aug. 2017.

[21] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 1601–1611, Association for Computational Linguistics, July 2017.

[22] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 4623–4637, Association for Computational Linguistics, July 2020.

[23] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 7315–7330, Association for Computational Linguistics, July 2020.

[24] 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)*, pp. 1–8, IEEE, 2020.

[25] K. Lee, K. Yoon, S. Park, and S.-w. Hwang, "Semi-supervised training data generation for multilingual question answering," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[26] A. Asai, A. Eriguchi, K. Hashimoto, and Y. Tsuruoka, "Multilingual extractive reading comprehension by runtime machine translation," 2018.

[27] P. Efimov, A. Chertok, L. Boytsov, and P. Braslavski, "Sberquad–russian reading comprehension dataset: Description and analysis," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 3–15, Springer, 2020.

[28] C. P. Carrino, M. R. Costa-jussà, and J. A. R. Fonollosa, "Automatic Spanish translation of SQuAD dataset for multi-lingual question answering," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 5515–5523, European Language Resources Association, May 2020.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[30] S. Banerjee and S. Bandyopadhyay, "Bengali question classification: Towards developing QA system," in *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, WSSANLP@COLING 2012, Mumbai, India, December 8, 2012* (V. Sornlertlamvanich and A. Malik, eds.), pp. 25–40, The COLING 2012 Organizing Committee, 2012.

[31] S. M. H. Nirob, M. K. Nayeem, and M. S. Islam, "Question classification using support vector machine with hybrid feature extraction method," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6, 2017.

[32] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "BFQA: A bengali factoid question answering system," in *Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings* (P. Sojka, A. Horák, I. Kopecek, and K. Pala, eds.), vol. 8655 of *Lecture Notes in Computer Science*, pp. 217–224, Springer, 2014.

[33] S. Hoque, M. S. Arefin, and M. M. Hoque, "Bqas: A bilingual question answering system," in *2015 2nd International Conference on Electrical Information and Communication Technologies (EICT)*, pp. 586–591, 2015.

[34] S. T. Islam and M. N. Huda, "Design and development of question answering system in bangla language from multiple documents," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1–4, IEEE, 2019.

[35] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4996–5001, Association for Computational Linguistics, July 2019.

[36] L. Pan, W. Lei, T.-S. Chua, and M.-Y. Kan, "Recent advances in neural question generation," 2019.

[37] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 3901–3910, Association for Computational Linguistics, Oct.-Nov. 2018.

[38] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 866–874, Association for Computational Linguistics, Sept. 2017.

[39] Y. Kim, H. Lee, J. Shin, and K. Jung, "Improving neural question generation using answer separation," *CoRR*, vol. abs/1809.02393, 2018.

[40] D. B. Lee, S. Lee, W. T. Jeong, D. Kim, and S. J. Hwang, "Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 208–224, Association for Computational Linguistics, July 2020.

[41] K. Zi, X. Sun, Y. Cao, S. Wang, X. Feng, Z. Ma, and C. Cao, "Answer-focused and position-aware neural network for transfer learning in question generation," in *Knowledge Science, Engineering and Management - 12th International Conference, KSEM 2019, Athens, Greece, August 28-30, 2019, Proceedings, Part II* (C. Douligeris, D. Karagiannis, and D. Apostolou, eds.), vol. 11776 of *Lecture Notes in Computer Science*, pp. 339–352, Springer, 2019.

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 17351780, nov 1997.

[43] R. Karim, M. Islam, S. R. Simanto, S. A. Chowdhury, K. Roy, A. Al Neon, M. Hasan, A. Firoze, R. M. Rahman, *et al.*, "A step towards information extraction: Named entity recognition in bangla using deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 6, pp. 7401–7413, 2019.