BnText2Table – Dataset and Text-to-Table generation in Bangla

by

Tahreema Rahman Zariyat
20101433
Fahim Irfan Ahmed
20101508
Tahsina Tajrim Oishi
20101394
Maruf Morshed
20101299

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Full Name & Signature:

Tahreema Rahman Zariyat
20101433

Fahim Irfan Ahmed
20101508

Oishi

Tahsina Tajrim Oishi
20101394

Maruf

Maruf Morshed
20101299

Approval

The thesis titled "BnText2Table – Dataset and Text-to-Table generation in Bangla" submitted by

1. Tahreema Rahman Zariyat(20101433)
2. Fahim Irfan Ahmed(20101508)
3. Tahsina Tajrim Oishi(20101394)
4. Maruf Morshed(20101299)

of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January, 2024.

Examining Committee:

Supervisor:
(Member)

Saiful

Md Saiful Islam
Senior Lecturer
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Ethics Statement

We hereby declare that this research is based on our own research findings. Things which are taken from external sources have been acknowledged and cited properly. Furthermore, we confirm that this thesis has not been submitted or presented, either in its entirety or partially for the purpose of receiving a degree from any other educational institution or university.

Abstract

In this fast-paced world, everyone relies on technology to get their work done quickly and efficiently, since using technology greatly simplifies every task that needs to be done. The majority of the publications are lengthy and packed with crucial data. However, in many instances, extra words are also added to boost the word count, which causes a number of difficulties when trying to get the desired information. For the English language, numerous tools are available to summarize the text and present it in tabular form. However, it is not the same for our mother tongue, Bangla. Despite being the 5th most-spoken native language in the world, there is no tool available to ease the workload in Bengali language. Our research will assist in such circumstances by summarizing the given information in tabular form within the shortest possible time. Since there is no dataset available that will be suitable for our research, we have prepared the dataset ourselves. Then, we have used the mBART-50-large, mT5-base, mT5-m2m-CrossSum and BanglaT5 models for the implementation. Finding the appropriate table headers in light of the context and order of the data is the most important task in this study. To sum up, our main goal is to develop a benchmark dataset for a text-to-table model for the betterment of the NLP research community.

Keywords: Bangla NLP; Text2Text; Summarizer; mBART; Transformer; Information Extraction; T5; mT5

Acknowledgement

We extend our gratitude to the Great Almighty Allah for guiding us through the completion of our thesis without any major setbacks. Our appreciation also goes to our supervisor, Mr. Md Saiful Islam sir for his kind support and guidance throughout the process. We are grateful for the unwavering support of our parents as well, without whom this achievement would not have been possible.

Table of Contents

List of Figures

Chapter 1

1.1 Introduction

Ever since mankind has become acquainted with the art of storing information in the form of writing, the possibilities of forgetting crucial knowledge have been reduced. Gradually, with the development of civilization and the discovery of technology, instead of storing data by writing on paper, we started typing it into electronic devices. This method is both eco-friendly and much more convenient than writing on paper. Since there is no restriction on storage, most of the stored data becomes lengthy, with excess words being added to enhance the quality of the writing. However, when it is required to extract accurate information from the bulk of text data, it is a troublesome task to go through each word. whereas a preferable way to present the data is in a table, where it is arranged systematically in particular rows and columns. So, it is easier to find the necessary information in the tabular format of the text data.

We draw inspiration for our research from the widely researched “Table to Text” conversion [1] [2], whose main task is to give a description as output of the input table. However, the inverse of this problem, “Text to Table”, which is a new problem setting of Information Extraction (IE), is not been much researched yet. [3] Traditionally, IE tasks include named entity recognition and relation extraction, among others. [4][5] However, “Text to Table” is unique compared to the traditional IE approaches.

Even though it is possible to find different tools for the English language with the advancement of technology, there is hardly any tool that can convert the long Bangla texts into precise tabular information. Despite being a widely spoken language in the world, there are hardly any resources to help with the research. As a result, the researchers and people working with Bangla need to go the extra mile to find out the needed information.

Moreover, most of the paragraphs are stuffed with strong Bengali vocabulary and proverbs that are difficult for the majority of the native people to figure out their core meaning. This makes the information extraction process more difficult. So with the help of our research methods, apart from the commonly used English texts, it will take into account Bangla information and provide the labelled data in order to identify the required data. This will make the work of Bengali researchers a lot easier and more feasible, and so provide the world with a more precise collection of data.

While “Table to Text” is required when the information provided in a table is needed to be described in natural language, the data extracted by “Text to Table” is needed for document summarization and text mining. Since the topic is still new and not much researched, our main focus is to broaden the horizons on this and formulate a model that will be able to analyse text documentation and produce the same information in a tabular format. For the information extraction of Bengali language, it is very difficult to look for required datasets to clarify the results. As a result, preparing datasets from scratch had been needed and so we can verify our results. The goal is to make Bangla information extraction as easier as the English language.

1.2 Problem Statement

The amount of digitized data is increasing rapidly with each passing day. If we searched for a particular piece of information, we would see several results found in documents. However, extracting desirable information from such documents is very time-consuming and requires too much effort. It is very challenging to go through enormous documents and extract information while maintaining both accuracy and robustness. This is where tabulation comes in handy.

There has been remarkable research for such tasks in other languages but not in Bangla because it is yet considered as a low-resource language. Therefore, we thought of proposing a system for our mother language. The system will be extremely helpful for reading any documents or texts as it will be extracting every crucial piece of information from the documents or texts and sorting them in a table format, grabbing and fetching key information from any document will be much easier. In addition to saving time and effort, arranging the data in tables also makes the data easily understandable. This process will make the unprocessed documents or text data more classified as well. Eventually, the reader should be able to retrieve an entire document's gist and crucial information within a very short time and through an easy process.

Due to the unavailability of the appropriate dataset for our research, we had to develop a benchmarking dataset by ourselves. This dataset will not only be helpful for our research but also contribute greatly in the NLP research community of Bangla language in the future.

Our intent is to reduce the hassle of creating the table and inserting data manually through the implementation of various models and techniques. We want to find a suitable method to analyse complex data with ease and process it to assemble the ideal table. The proposed final system would be able to extract the significant information from any text or document and arrange it in a table format. We will build an efficient method using a summarizer [6], named entity recognition [7] and transformer [8]. In order to test our model and compare its accuracy, we will use the available datasets so that we can compare our results with the existing methods and maintain a higher accuracy level than others.

1.3 Research Objectives

The main goal is to develop a dataset which will be suitable for tasks related to Text to Table generation in Bangla language. In addition to that, our proposed model would be able to accomplish a few objectives. These are mentioned as follows:

1. There is scarcity of Bangla dataset. So we will prepare a benchmarking dataset which will be useful for future NLP task in Bangla language.
2. It will be able to analyze the whole input text and produce the information in a table.
3. The generated table will be free of redundant information and contain only the necessary items.
4. The main motivation to find the best model for text to table conversion much more efficiently.

With these proposed objectives, it would be possible to implement the model, making it a useful tool for people who work with thousands and thousands of words of Bangla every day. Not only will the valuable time of people be saved, but also the higher efficiency of texts will play a significant role in different workings.

Chapter 2

Literature Review

2.1 Background Study

The only study done on this exact topic was conducted by [9]. They developed a new method employing the vanilla sequence-to-sequence model fine tuned from pre-trained language model BART along with the techniques of table constraint and table relation embeddings. The results showed that their customised method shows significant improvement in accuracy of information extraction and is able to boost the performance of the vanilla sequence-to-sequence model. They used four existing datasets which are traditionally utilised for "Table to Text". In the first dataset, their method turned out to be the best performer with the vanilla sequence-to-sequence being a close second in terms of most of the measures. Especially in case of the F1 score of the non-header cells the model scored 90.8 for the teams and 88.97 for the players without any error rate. However, in the remaining three datasets, the results of the customised method and vanilla sequence-to-sequence were comparable.

Furthermore, they conducted an additional study on their method. They accomplished it by excluding the pre-trained language model (BART), table constraint (TC) and table relation embedding (TRE) from their newly created method. The resultant method turned out to be similar to the vanilla sequence-to-sequence model. Thus, the results of the last three datasets turned out to be the same for the vanilla seq2seq method and their customised method. So it was concluded that the usage of TC and TRE plays an important role in elevating the effectiveness of tables in crucial tasks, such as – in case of tables which are huge and contain innumerable rows and columns just like the first dataset that was used.

While substantial research has been done on the concept of "Table to Text", there has not been an extensive study on "Text to Table" till now. There has been only one considerable study on our desired topic. Due to the insufficiency of prior research, we have split our topic of interest into three parts and evaluated them independently. The three distinct sectors of our interest are summarization, named entity recognition, and sequence-to-sequence transformer.

2.2 Transformer

In [10], the authors used TRANSEQ with the combined efficiency of a transformer and a sequence-to-sequence model. They used an encode-decode paradigm where they implemented two layers of encoders (Transformer Encoder & GRU-RNN Encoder). For the pretraining module of the input sentence, the Transformer Encoder with 6 stacked identical layers, helps to reach an adequately structured representation to achieve the desired output.

[11] proposed a model which employs an architecture that is predicated on a Standard Transformer, and it is made up of an auto-regressive decoder and a multilayer bidirectional encoder. The Standard Transformer serves as the foundation for the architecture. Enhanced sequence-to-sequence Autoencoder using a Contrastive Learning framework is used by them in order to increase the sequence-to-sequence model's potential for denoising and to extend the model's flexibility by means of fine-tuning. During the fine-tuning phase, ESACL will optimize the model. This will result in a significant reduction in the total amount of time spent training and will make better use of the computational resources that are available.

From [12], we can see that the authors implemented a sequence-to-sequence model that consists of a left-to-right autoregressive decoder and a bidirectional encoder over distorted text. Denoising autoencoders, BART uses the sequence-to-sequence Transformer as the base architecture. The encoders and decoders for the small model each have six layers, but in this case, the encoders and decoders for the large model each have twelve layers. Backpropagating the cross-entropy loss that occurs at the output of the BART model is the method that is used to train the source encoder in two separate phases.

[13] proposed a new Seq2Seq model with a formation based on the encoder-decoder architecture of the Transformer model. They introduced N-gram Prediction, a multilayer Transformer encoder, and for better self-attention prediction, they included the N-stream self-attention mechanism. Rather than the one step ahead optimization of the conventional Seq2Seq models, the ProphetNet is optimized by N-step ahead prediction, which helps to predict the upcoming tokens continuously depending on the context tokens that were received before at different times.

2.3 Reverse Problem

According to [14], text production from non-linguistic input can be done with the help of data-to-text generation. The use of extensive datasets and neural network models that are trained end-to-end without explicitly modelling what to say and in what order has been made feasible by recent developments in data-to-text generation. This paper portrays a neural network architecture that combines planning and content selection without jeopardising end-to-end training. The generation task is split into two sections. Here, a content plan is created outlining what information should be provided and in what sequence given an archive of data records (combined with descriptive papers), and then we build the document taking the content plan into account. On the recently released RotoWIRE dataset, studies demonstrating automatic and human-based evaluation reveal that this model1 outperforms strong baselines, advancing the forefront of the field.

[15] talked about neural pipelines. The goal was to use neural pipelines for data-to-text generation that can help provide a natural language description of structured data. Data-to-text (D2T) generation is affected by overloading in the data encoding and recurring training data noise as a result of training on in-domain data. In the paper, it looked at methods to make pretrained language models (PLMs) less reliant on D2T generation datasets while still utilising their surface realisation abilities. Therefore, creating text by modifying single-item descriptions using a series of modules trained on three text-based operations from the general domain: ordering, aggregation, and paragraph compression, was suggested. On a synthetic corpus called WikiFluent that was constructed from the English Wikipedia, we train PLMs to do these tasks. The evaluations conducted on the WebNLG and E2E datasets, two of the most important triple-to-text datasets, demonstrate that the method enables D2T generation from RDF triples in zero-shot scenarios.

[16] portrayed the dual tasks of data-to-text and text-to-data. Transforming structured data, like graphs or tables, into conversational text and the reverse are known as data-to-text (D2T) and text-to-data (T2D) activities. Typically, the aforementioned tasks are completed independently and with the use of data taken from one repository. The present systems make use of computational models of language that have already been trained and are optimised for D2T or T2D tasks. This strategy has two key drawbacks: first, learning is constrained by the lack of available data; second, every task and origin requires a different system that must be customised. In this research, a more general situation where data are available from multiple separate sources is considered. Each source provides an independent archive of text and structured data, each with a distinct data format and semantic domain. A variational auto-encoder model with separated style and content parameters that enables us to capture the variety resulting from various text and data sources is developed. The obligations of D2T and T2D are handled simultaneously by this model. The model is tested on multiple datasets and illustrates that, by learning from many sources, it can sometimes exceed its supervised single-source equivalent in terms of performance.

2.4 Bangla NLP

For the Bangla information extraction, different researchers had different ideas. According to [17], a Text-to-Text Transfer Transformer (T5) Language Model with the small variant of BanglaT5 is used to detect grammatical faults in Bangla. The model is fine-tuned using a dataset of 9385 sentences, where errors were bracketed by a dedicated demarcation sign. The T5 model required significant post-processing to be tailored to the error detection task because it was primarily intended for translation and not for this particular use. According to our tests, the T5 model can identify grammatical faults in Bangla with a low Levenshtein Distance; nevertheless, post-processing is necessary for best results. After post-processing the output of the refined model, the final average Levenshtein Distance on a test set of 5000 sentences was 1.0394. This paper highlights the difficulties in modifying a translation model for grammar and provides a thorough examination of the mistakes identified by the model. We show that T5 models can identify grammatical problems in a variety of languages by expanding our method to different languages.

There is another Bangla paper that has the implementation of NER. [18] portrayed the Bengali complex. According to them, one of the core tasks in natural language processing is named entity recognition (NER), which is the act of identifying and organising named things in text. Despite being the seventh most spoken language in the world, not much work has been done on complicated named entity identification in Bangla. Since CNER requires the identification and classification of complex and compound entities—which are uncommon in Bangla language—it is a more difficult task than regular NER. The Bangla Complex Named Entity Recognition Challenge winning solution is presented in this work. It addresses the CNER task on the BanglaCoNER dataset by employing two distinct methods: fine-tuning transformer-based Deep Learning models, like BanglaBERT, and using Conditional Random Fields (CRF). The dataset included 800 sentences in the.conll format for validation and 15300 sentences for training. The dataset's seven distinct NER tags and the noticeable frequency of English words in them, as shown by exploratory data analysis (EDA) on the dataset, indicate that it is most likely synthetic and the result of translation. In addition to optimising the BanglaBERT (big) model for NER, we experimented with a range of feature combinations, including as Part of Speech (POS) tags, word suffixes, Gazetteers, and cluster information from embeddings. We discovered that not all linguistic patterns are instantly clear to humans or even intuitive. For this reason, deep learning-based models have shown to be more successful in natural language processing (NLP), including the CNER examination. On the validation set, fine-tuned BanglaBERT (big) model obtains an F1 Score of 0.79. All things considered, our research underscores the significance of Bangla Complex Named Entity Recognition, especially when dealing with artificial datasets. Additionally, our results show how efficient Deep Learning models like BanglaBERT are for NER in Bangla.

Chapter 3

Dataset

3.1 Main Dataset

For our preliminary study, we have used a dataset of biographies extracted from Wikipedia. From that, we have randomly sampled some of them to carry forward our work. Then, we have converted the English biographies into Bangla in two ways, translation and transliteration. These two are distinct ways of conveying the meaning of words of one language to another.

3.2 Work Flow

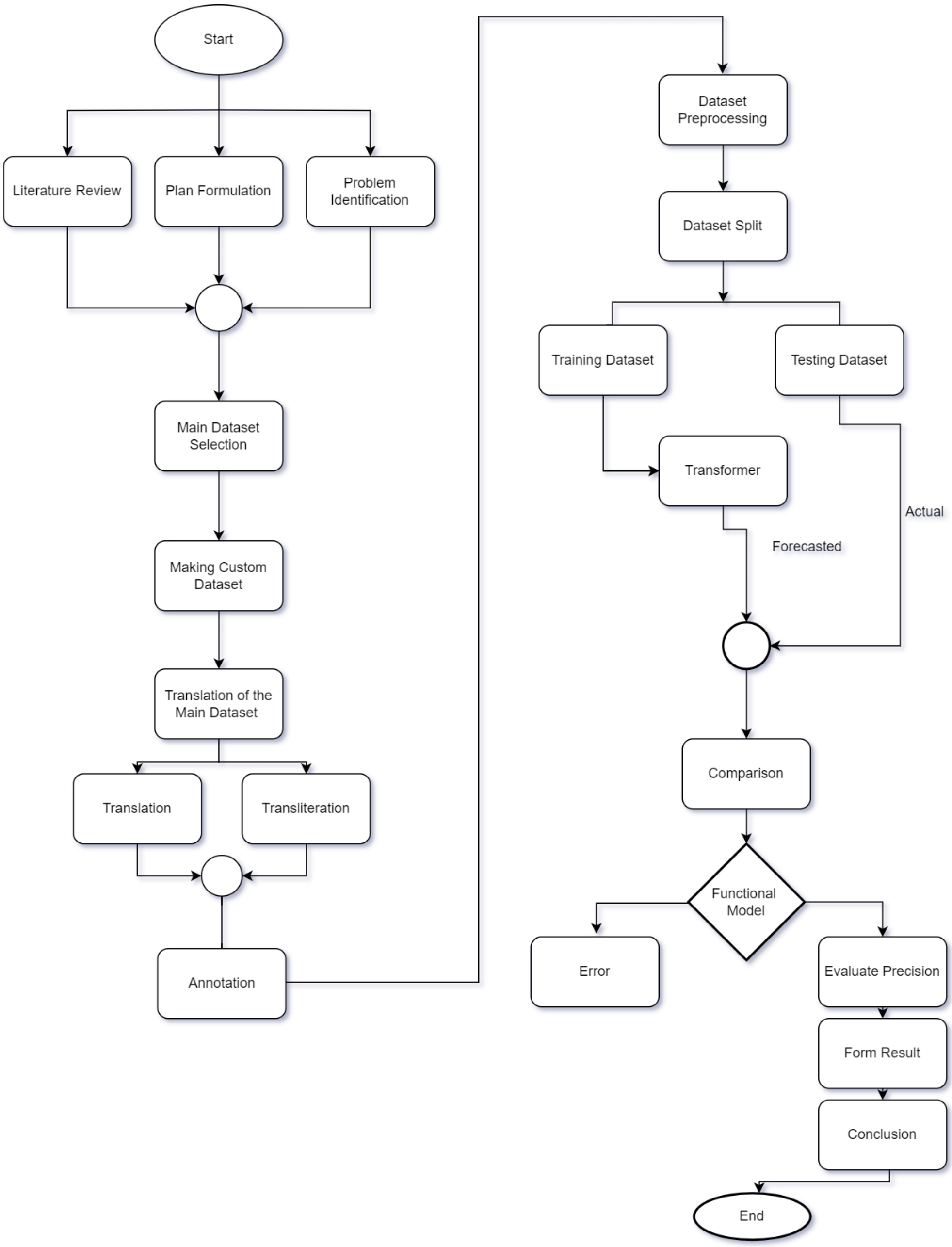
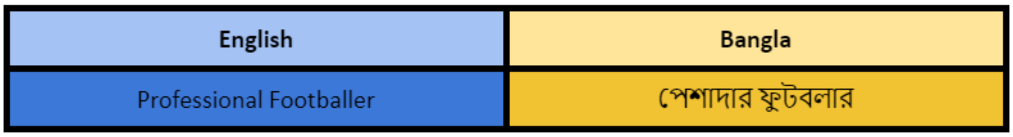


Figure 3.1: Processing the Dataset

3.2.1 Translation

This means to convert the meaning of a word from one language to another. In this process, the words are changed but the context remains the same.

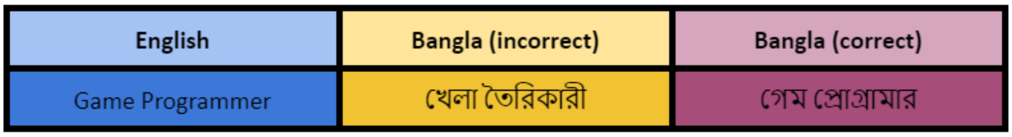


English	Bangla
Professional Footballer	পেশাদার ফুটবলার

Figure 3.2: Translation

3.2.2 Transliteration

This means to transfer each word of one language to another without changing the meaning of it. This method of mapping also helps people to pronounce words which are originally written in foreign languages.

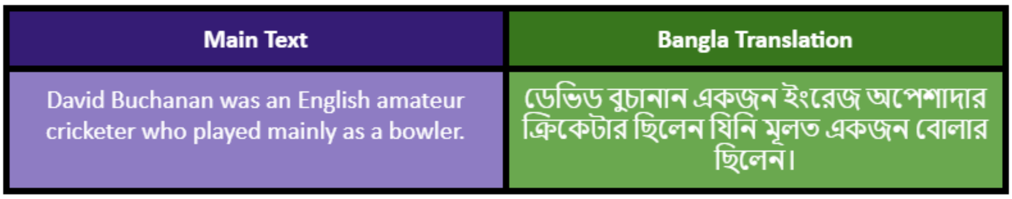


English	Bangla (incorrect)	Bangla (correct)
Game Programmer	খেলা তৈরিকারী	গেম প্রোগ্রামার

Figure 3.3: Transliteration

3.3 Annotation Guidelines

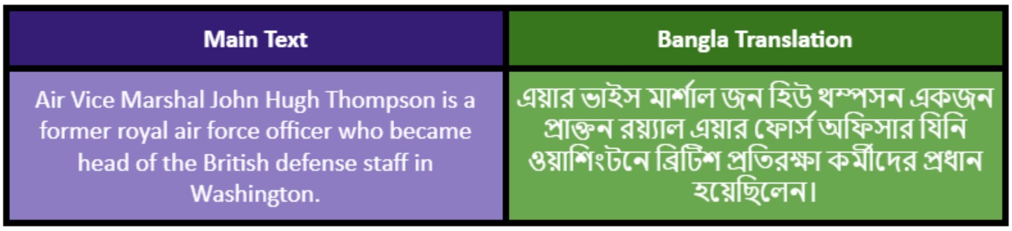
1. In the cases, where translation can provide the expected output we just literally translated the data.



Main Text	Bangla Translation
David Buchanan was an English amateur cricketer who played mainly as a bowler.	ডেভিড বুচানান একজন ইংরেজ অপেশাদার ক্রিকেটার ছিলেন যিনি মূলত একজন বোলার ছিলেন।

Figure 3.4: Guideline 1

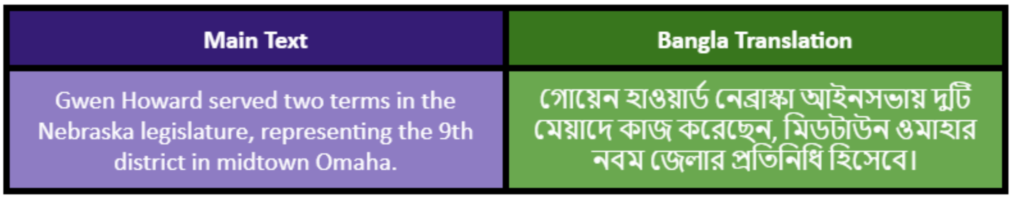
2. In the cases, where translation does not provide any proper output we chose to transliterate the data.



Main Text	Bangla Translation
Air Vice Marshal John Hugh Thompson is a former royal air force officer who became head of the British defense staff in Washington.	এয়ার ভাইস মার্শাল জন হিউ থম্পসন একজন প্রাক্তন রয়্যাল এয়ার ফোর্স অফিসার যিনি ওয়াশিংটনে ব্রিটিশ প্রতিরক্ষা কর্মীদের প্রধান হয়েছিলেন।

Figure 3.5: Guideline 2

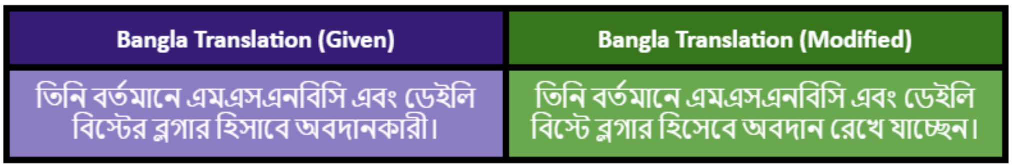
3. To make the data more accurate we had to change some words and rearrange in order to make the meaning of the sentence according to the original text.



Main Text	Bangla Translation
Gwen Howard served two terms in the Nebraska legislature, representing the 9th district in midtown Omaha.	গোয়েন হাওয়ার্ড নেব্রাস্কা আইনসভায় দুটি মেয়াদে কাজ করেছেন, মিডটাউন ওমাহার নবম জেলার প্রতিনিধি হিসেবে।

Figure 3.6: Guideline 3

4. To align with the exact meaning of the English translation, we had to restructure some of the sentences to make it more proper and meaningful.



Bangla Translation (Given)	Bangla Translation (Modified)
তিনি বর্তমানে এমএসএনবিসি এবং ডেইলি বিস্টের ব্লগার হিসাবে অবদানকারী।	তিনি বর্তমানে এমএসএনবিসি এবং ডেইলি বিস্টে ব্লগার হিসেবে অবদান রেখে যাচ্ছেন।

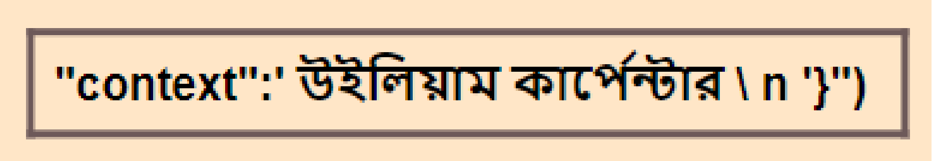
Figure 3.7: Guideline 4

3.4 Cleaning of the Dataset

1. We found a lot of irrelevant signs and symbols while translating the data.

   For example: “&”, “\n” etc.

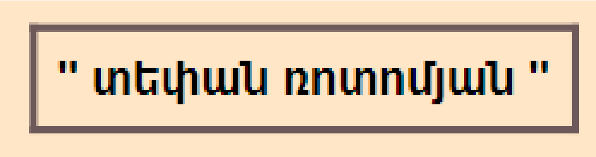
   These signs and symbols had to be removed for the dataset to be more precise.



```
"context":' উইলিয়াম কার্পেন্টার \ n '}")
```

Figure 3.8: Cleaning Process 1

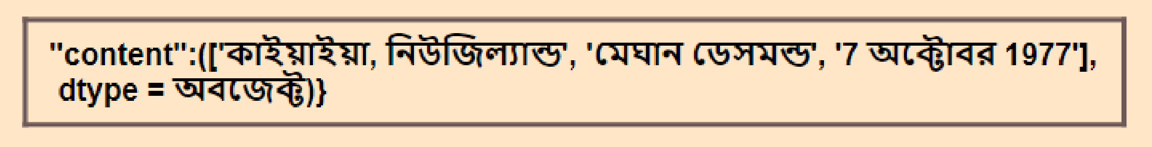
2. There were some foreign words which were found in the dataset which were not relevant and we had to remove those to avoid any difficulty.



```
" ստեփան ռոստոմյան "
```

Figure 3.9: Cleaning Process 2

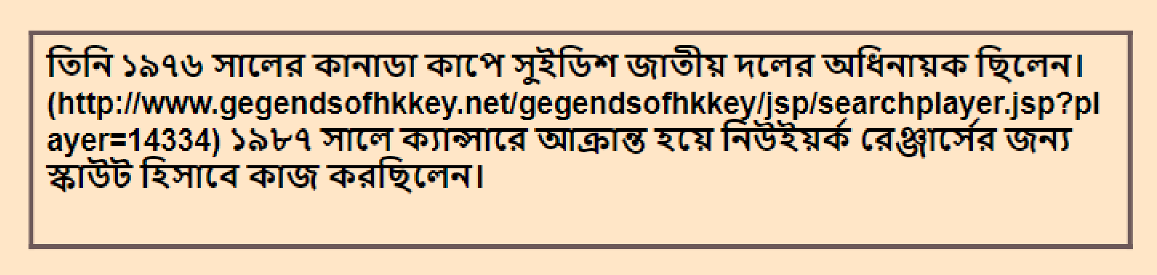
3. In the table header files we found dtype in almost every header file. We had to remove them.



```
"content":(['কাইয়াইয়া, নিউজিল্যান্ড', 'মেঘান ডেসমন্ড', '7 অক্টোবর 1977'],
 dtype = অবজেক্ট)}
```

Figure 3.10: Cleaning Process 3

4. There were some some URLs found in the dataset, we had to remove them for data relevance.



তিনি ১৯৭৬ সালের কানাডা কাপে সুইডিশ জাতীয় দলের অধিনায়ক ছিলেন। (http://www.gegendsofhkkey.net/gegendsofhkkey/jsp/searchplayer.jsp?player=14334) ১৯৮৭ সালে ক্যান্সারে আক্রান্ত হয়ে নিউইয়র্ক রেঞ্জার্সের জন্য স্কাউট হিসাবে কাজ করছিলেন।

Figure 3.11: Cleaning Process 4

3.5 Tokenization

These are the token counts we found before and after implementing the models:

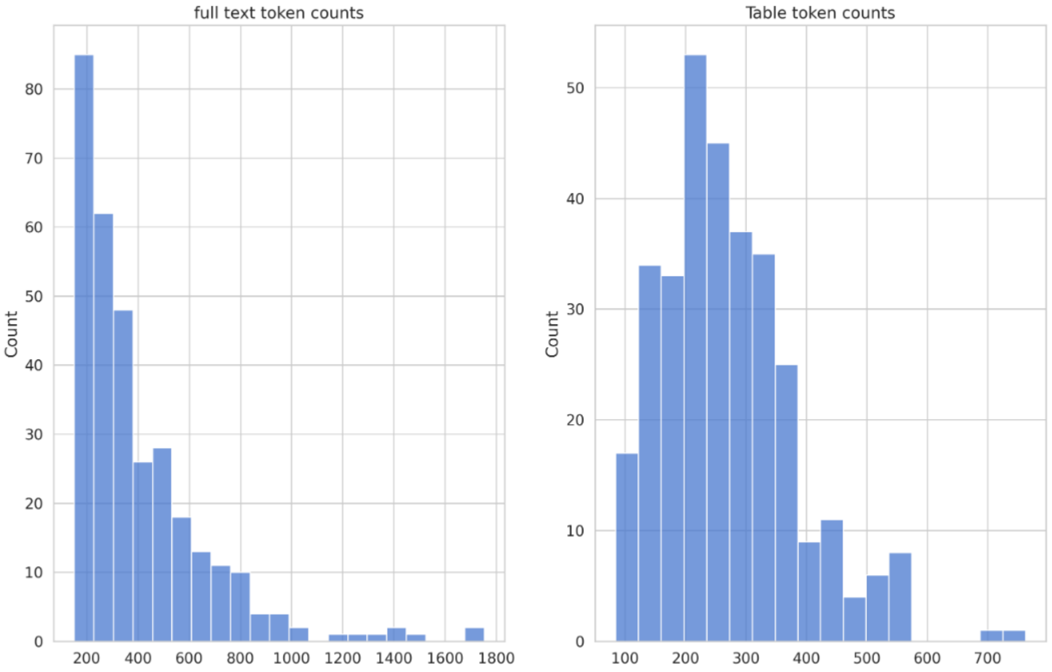


Figure 3.12: Number of Tokens

3.6 Limitations

Implementing the dataset into these two models proved to be a difficult task for us. The initial dataset being the WikiBio dataset in English language could perform pretty well in these models. But after the custom dataset which is the translation of the WikiBio dataset was made, this did not perform very well compared to that of the English dataset of WikiBio. The translated data being not having high accuracy created a lot of issues with the implementation of the dataset. The implementation of the Bangla dataset had less accuracy in terms of results also for a lot of factors. The factors being the Translation and Transliteration difference. So, the translation being the literal translated language of the other language and the transliteration being kind of the phonetic translation of the other language.

Chapter 4

Description of Models

4.1 Transformer

4.1.1 Definition

A transformer model is a deep learning architecture outlined for sequence-to-sequence chores, such as natural language processing and machine translation. It makes use of self-attention processes to examine input data simultaneously, effectively capturing long-range relationships. Transformers allow concurrent computation across input sequences, which improves scalability and training speed in contrast to standard sequential models.

The key components of the model are feedforward layers and multi-head self-attention, which enable the retention of intricate patterns. Transformers are an essential breakthrough in the deep learning space because they are now standards in many applications and exhibit cutting-edge performance in tasks requiring contextual knowledge.

4.1.2 Working Principle

The transformer model's ability to efficiently gather contextual links within input sequences is contingent on its self-attention mechanism. To represent the sequence order, relative encodings are added after the input sequence has been embedded into vectors. The multi-head self-attention mechanism, which determines the attention scores of each element in the sequence in relation to all others, is the central component of the model. The model can assess contributions from different positions owing to these scores, which establish the relative value of various sequence segments for each element. This is followed by feedforward neural network processing and accumulating the attention-weighted values.

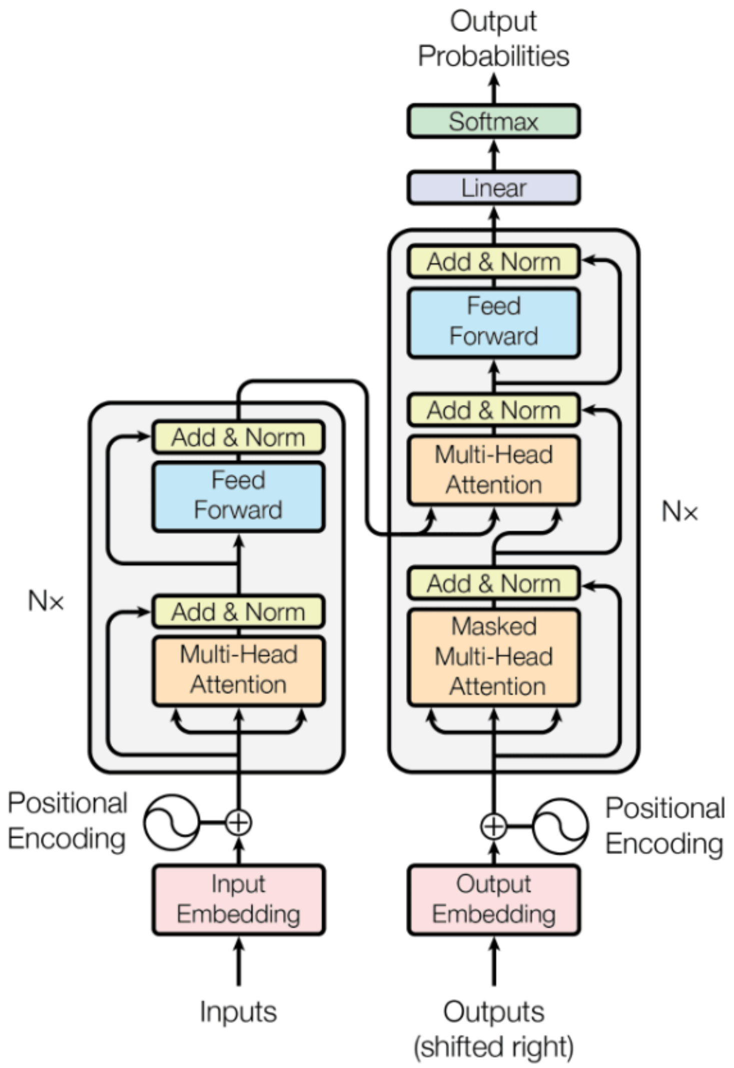


Figure 4.1: Transformer Model

The model is able to learn a variety of interactions in parallel by repeating this procedure across several attention heads. Training stability is boosted by residual connections and layer normalisation. The transformer thrives at natural language processing and other sequence-based tasks because of its architecture, which makes parallelization a breeze and allows for faster training. It also encodes long-range dependencies in sequential data effectively.

Both encoder and decoder modules usually have multiple layers. Depending on the specific specifications of a job or dataset, the number of encoder and decoder layers can be modified. To help the model properly capture complicated and hierarchical patterns, it is typical to have plenty of stacked layers. A hyperparameter that can be altered throughout the model-building and training stages is the precise number of layers a Transformer model uses. Depending on the application and available computational resources, different numbers of layers may be utilised.

In a Transformer model, positional encoding serves as a guide for the model about the relative placements of tokens inside a particular sequence. Positional encoding is introduced to inject positional information because the Transformer architecture fails to comprehend the order of tokens in a sequence by definition.

Typically, partial encoding is carried out on the tokens' input embeddings. It gives every token a distinctive positional signature according to where it is in the ordered sequence. This allows the model to take into account the logical progression of the input data by allowing it to differentiate between tokens at various places. For endeavours requiring sequential information, like language translation and text production in natural language interpreting, positional encoding is necessary.

4.1.3 Implementation

Encoder and decoder components must be constructed in order to implement a transformer model. Sequences entered are processed by the encoder, and sequences exited are generated by the decoder. Each layer of the encoder and decoder is comprised of a feedforward neural network and a multi-head self-attention mechanism. Sequence order is integrated into the model via positional encodings. Using reverse propagation to update its parameters, the model minimises a suitable loss function during training. During training, attention masks can be used to stop paying attention to upcoming tokens.

Transformers are frequently tuned for particular tasks and pre-trained on huge datasets. Pre-built transformer layers are provided by well-known deep learning frameworks like TensorFlow and PyTorch, which makes configuration easier. Tools for attention visualisation make the model's concentration easier to comprehend. Effective implementation ensures optimal performance on workloads such as machine translation, picture captioning, and natural language comprehension, involving the proper selection of optimizers, regularisation, and hyperparameter tuning.

4.2 T5

4.2.1 Definition

The T5 or Text-To-Text Transfer Transformer model is a new addition to the wide range of transfer learning techniques in NLP. As mentioned in the name, this has a Transformer based architecture and a text to text framework, which means it takes a text as input and produces another text as the output. This model was developed by Google's AI Team.

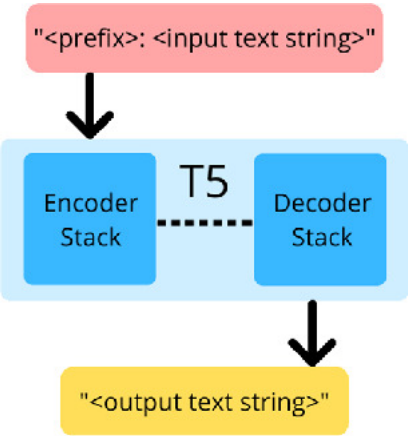


Figure 4.2: Structure of T5 Model

4.2.2 Working Principle

A number of NLP tasks such as text summarization, sentiment analysis, language translation, text classification, question answering, text generation, etc. can be accomplished using the T5 model. To differentiate between the tasks, a prefix is used before the input to indicate which task the model is going to be used for. [19]

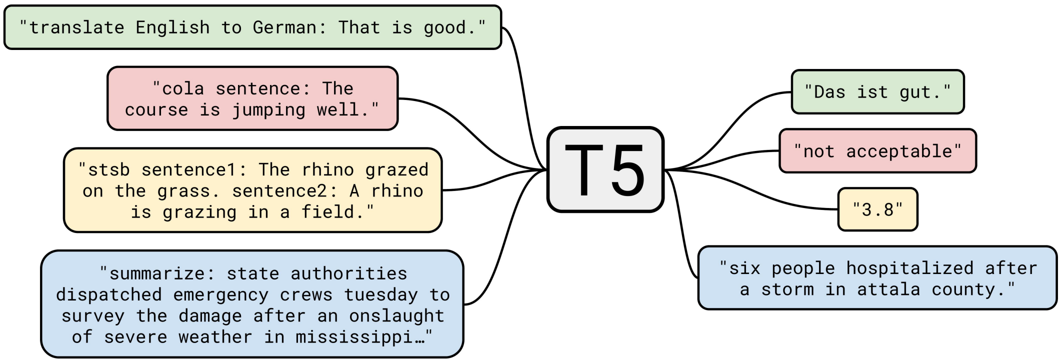

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 4.3: Working of T5 Model

Apart from being applicable to multiple tasks, T5 is also available in different sizes, such as - t5 small, t5 base, t5 large, etc. In our research, we have used the small version of the model for text summarization only. T5 Small has almost 60 million parameters, and for summarization, we will use the prefix “summarize” before the input.

4.2.3 Implementation

At first, the dataset is prepared by making sure all the input text has “summarize” as their prefix. This prefix will help the model locate which text to work on. Then the pre-trained T5 small model is used. After the completion of the task, it is evaluated on the basis of ROUGE score which checks the overlap between the produced text and the given input text.

4.3 mT5-base

4.3.1 Definition

The T5 model works on a wide variety of NLP tasks in English language. However, a significant portion of the world population does not speak English. Thus, such language models limit the usage of other languages in NLP tasks. So, [20] introduces a multilingual variant of the T5 model which has been named mT5. This abbreviation stands for Multilingual Text-To-Text Transfer Transformer.

4.3.2 Working Principle

This version of T5 was pre-trained on a new Common Crawl-based dataset which covers 101 languages including Bengali. This multilingual variant has been trained following a similar recipe as the original T5 language model.

mT5 is pre-trained on the mC4 corpus. Like the multilingual version of the T5 model, a multilingual variant of the C4 dataset was developed specifically and it is called mC4. This contains natural text in 101 languages and it has been drawn from the public Common Crawl web scrape. This variant model requires to be fine-tuned first before using it on any task unlike the original T5 model. The reason for this is that mT5 was pre-trained only unsupervised. Yet, mT5 is able to achieve state-of-the-art performance on many cross-lingual NLP tasks such as– classification, named entity recognition, question-answering, etc.

4.3.3 Implementation

Due to the lack of supervised training while pre-training on mC4, a prefix needs to be used while multi-task fine-tuning. The available prefixes are base, small, large, xl, xxl. mT5 has parameters from 300 million to 13 billion but for our work, we are using the “base” prefix which has 580 million parameters. But such task prefix is not needed during single-task fine-tuning. Lastly, the results found from mT5 are almost similar to those found from T5. So, now NLP tasks can also be accomplished in non-English languages as well.

4.4 BERT

4.4.1 Definition

BERT, or bidirectional encoder representations from transformers is a new natural language processing model that has changed the landscape of language. This model was developed by Google. This model is a pre-trained neural network architecture that is very efficient at capturing the contextual relationships among the words in a sentence. This implies that this model was pre-trained on the raw text only, where human interaction was absent while labeling the data.

NER, or named entity recognition, is a natural language processing technique with the help of which we can extract information from text. Here, we used bert-base-uncased to extract the named entities from the text of the datasets. Here, uncased implies that it will not find the difference between "Python" and "python". BERT is a transformer model that is pre-trained on a large amount of English text data in a self-supervised fashion. BERT has two variations, namely, base and large, for both cased and uncased input text. Uncased models get rid of accent markers. This model has almost 110 million English parameters.

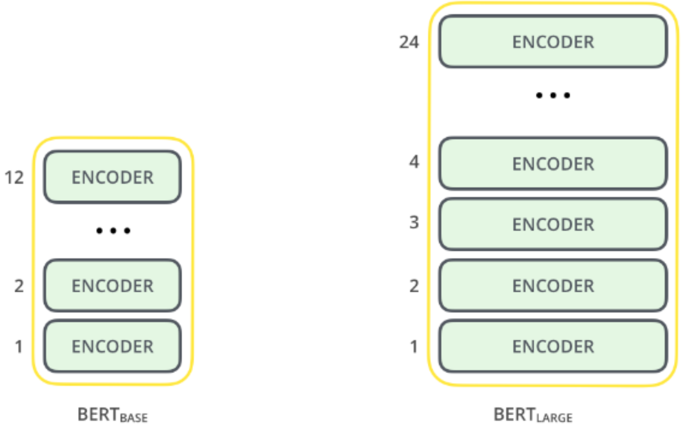


Figure 4.4: Structure of BERT Models

4.4.2 Working Principle

The BERT model relies on the Transformer architecture. Transformer is a neural network architecture where there are mechanisms, They are encoder and decoder, respectively. Here, BERT is a language model, and that is why it uses the encoder mechanism. [21] Here, the directional model reads the text in a directional way, which is left-to-right or right-to-left.

On the other hand, the transformer encoder reads the whole text input or whole sequence at once. This is why it is considered a bidirectional model. These traits of the transformer allow it to learn the whole context of the sentence or words based on the whole thing. In this model, the input will have a sequence of tokens. This sequence of tokens is embedded into vectors and then processed in the neural network. The output of this model will have a vector of size H, where each vector will represent an input token having the same index.

4.4.3 Implementation

Named Entity Recognition(NER), the application receives a sequence of text, and then it will identify the named entities or types of entities that are present in the text (person, date, organization). By using BERT, we can train an efficient NER model that will predict the NER label by training the output vector of each token on a classification layer.

4.5 mBART-50-large

4.5.1 Definition

The pre-trained model mbart-large-50 is primarily intended to be adapted for translation workloads. Further multilingual sequence-to-sequence tasks can be used to optimise it. By extending the encapsulation layers of the original mbart-large-cc25 checkpoint with randomly initiated vectors for an additional set of 25 language tokens and pretrained on 50 languages, MBart-50 is developed.

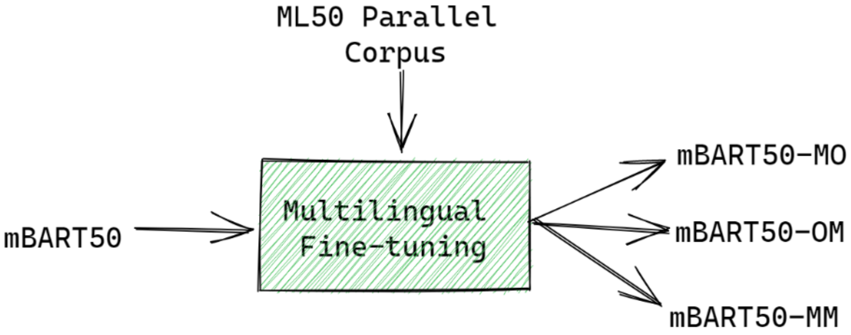


Figure 4.5: Structure of mBART50

4.5.2 Working Principle

Since the model is a transformer architecture variant of mBART, it most likely operates on the same foundations. It is probably pre-trained with an noise elimination autoencoder objective, which enables it to function well on a range of NLP tasks, such as summarization and translation.

4.5.3 Implementation

After installing the transformers library, the input texts need to be tokenized. Then the output is generated from the pre-trained model. The tokenizer's decode function is then taken to decode the output into human-readable text. Lastly, it is fine tuned for specific tasks.

4.6 Bangla T5 Transformer

BanglaT5[22] is a sequence-to-sequence transformer model. For more accurate analysis of data and insightful output, the pre-trained language model has been updated. First, we installed the transformer library in preparation for the implementation procedure. Next, the Transformers library's T5 model is implemented. Our data frame was split into three sets: training, cross-validation, and testing. The three sets had 78, 9, and 4 samples, respectively. With cross-validation, hyperparameter tuning and model evaluation are done after the model has been fine-tuned using the training data set. BanglaT5 is a text-to-text transformer that uses raw text as input. The AutoTokenizer package of Transformers automatically tokenizes the content. T5 can only receive 512 tokens at a time because of the token input limitation. A list of the fine-tuning hyperparameters is provided. Additionally, it contains a list of the fine tuning outcomes. To fine-tune BanglaT5, 50 epochs were executed in total. The table includes the last five sets of epoch results alongside the first ten sets.

4.7 mT5 m2m CrossSum

Cross-lingual summarizing is the process of employing a source text written in one language to create a summary in another. 1.68 million article-summary samples in more than 1,500 language pairs compose its dataset. Using a multilingual abstractive summary dataset, CrossSum[23] is constructed by cross-lingually extracting parallel articles authored in multiple languages and validating its quality through a controlled human evaluation. We suggest a multi-phase data sampling technique that can effectively train a cross-lingual summarization model that can summarize a piece of content in any language that the target audience communicates. LaSE, an embedding-based metric for automatically assessing summaries generated by models, is also introduced. LaSE and ROUGE have a strong correlation, but LaSE can be measured with accuracy even when there are no citations in the target language, unlike ROUGE. The model we propose consistently performs better than initial models on ROUGE and LaSE. As far as we are aware, CrossSum is the biggest cross-lingual summarizing dataset available, in addition to the first one that isn't English-centric. To stimulate more study on cross-lingual summarization, we are making the dataset, models, and scripts for training and evaluation accessible.

Chapter 5

Analysis of Result

5.1 Current Result

For our translated dataset we chose two models to test and run. The models are the mBART-large-50 and the mT5-base. Below is a comparative table of the results of the models.

Model	Precision	Recall	F1
mBART__large__50	0.56	0.57	0.53
mt5__m2m__crossSum	0.48	0.49	0.47
bangla__t5	0.43	0.42	0.45
mt5__base	0.45	0.44	0.43

Table 5.1: Model Performance Metrics

The results show that after implementing all the models, the mBART-large-50 is slightly better than the other models. The mBART-large-50 model has a F1 score of 0.53, whereas the mt5-base model has the F1 score of 0.43. Meanwhile, the mt5-m2m-corssSum, bangla-t5 and the mt5-base have the F1 Score of 0.47, 0.45, 0.43 consecutively. The precision values for mBART-large-50, mt5-m2m-crossSum, bangla-t5 and mt5-base are 0.56, 0.48, 0.43, and 0.45 consecutively. The precision values for mBART-large-50, mt5-m2m-crossSum, bangla-t5 and mt5-base are 0.56, 0.48, 0.43, 0.45 consecutively. The recall values for mBART-large-50, mt5-m2m-crossSum, bangla-t5 and mt5-base are 0.57, 0.49, 0.42, 0.44 consecutively.

The translated dataset labeling was an extremely difficult task to perform and the labeling might not be very accurate. The labeling being the most important factor caused a lot of reduction in the accuracy of the datasets. When the labeling is not perfect, the headers, the contents, the contexts may not be very relevant to the models. In the case of different language training, there is less precision in the results.

Keeping in mind all the factors, we used the mBART-large-50 and mt5-base model considering the fact that these are used for translations tasks and multilingual text related tasks. As these models are familiar to multilingual tasks, we decided to choose to run these on our translated Bangla dataset. The mBART-large-50, has an accuracy of 0.71, whereas the mt5-base model has an accuracy of 0.78.

5.2 Future Work

We have worked on the custom Bangla dataset, a translated dataset of the WikiBio dataset. As translation of a well known dataset like the WikiBio dataset is a troublesome work, we tried our best to make the dataset proper and run it with the chosen models. But considering the fact that, it has a big margin of data and there is a scope to work on it in the future, we can do a lot of work with it.

Firstly, the perfect adaptation and translation of the data will be a very good work for the Bangla dataset. As there are not enough and robust Bangla datasets, this WikiBio dataset of biography can be a very authentic source of information and can be made into a very well built dataset of Bangla.

Besides, there were a lot of issues with the labeling of the data. It takes a lot of time and effort to do the data labeling and still can not be highly precise. So, in terms of data labeling, a lot of effort is needed to make the dataset more accurate and robust.

In addition to that, there is a lot of scope to work on this data and give generative works through it. A major step of making a model were there inputs will be given in English and then there will be outputs in different languages defining the classifications of the data.

Chapter 6

Conclusion

Even though writing on electronic media has been a blessing for the environment, it has also increased our work to find the accurate information that people are looking for. This text-to-table extraction will be a revolutionary innovation that will be beneficial for the present world. Moreover, it was quite an impossible task few years back to consider this research on Bengali language. As a result, it will be a very useful one and one of the most demanding tool to use for the researchers. While table-to-text is a common topic to talk about, its inverse problem will be equally significant. Not only will it make our lives easier, but it will also prevent duplicity and repetitive information from being used. Moreover, as it the very first initiative for Bengali language, it will earn a huge popularity that will lead to using this tool even out of curiosity. Afterwards, when people get to understand the significance of it, this innovation will turn out into an extraordinary one. We have formalised different pre-trained models for this information extraction, such as the summarize model, the sequence-to-sequence model, and so on. Different experiments have been conducted with different sets of datasets. The results were noteworthy and outperformed our approaches. The frameworks used in this research have been effective, and in most of the cases, they showed maximum accuracy. The challenges faced by this process have also been taken into consideration for further study. As a result, we are expecting a fruitful outcome from this research.

Bibliography

[1] Sam Wiseman, Stuart Shieber, and Alexander Rush. "Challenges in Data-to-Document Generation." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2253–2263. DOI: 10.18653/v1/D17-1239. URL: https://aclanthology.org/D17-1239.

[2] Ratish Puduppully, Li Dong, and Mirella Lapata. "Data-to-Text Generation with Content Selection and Planning." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 6908–6915. DOI: 10.1609/aaai.v33i01.33016908. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4668.

[3] David Wadden et al. "Entity, Relation, and Event Extraction with Contextualized Span Representations." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5784–5789. DOI: 10.18653/v1/D19-1585. URL: https://aclanthology.org/D19-1585.

[4] Tapas Nayak and Hwee Tou Ng. "Effective Modeling of Encoder-Decoder Architecture for Joint Entity and Relation Extraction." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 8528–8535. DOI: 10.1609/aaai.v34i05.6374. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6374.

[5] Suncong Zheng et al. "Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1227–1236. DOI: 10.18653/v1/P17-1113. URL: https://aclanthology.org/P17-1113.

[6] Tian Shi et al. "Neural Abstractive Text Summarization with Sequence-to-Sequence Models." In: *CoRR* abs/1812.02303 (2018). arXiv: 1812.02303. URL: http://arxiv.org/abs/1812.02303.

[7] Emma Strubell et al. "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2670–2680. DOI: 10.18653/v1/D17-1283. URL: https://aclanthology.org/D17-1283.

[8] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

[9] Xueqing Wu, Jiacheng Zhang, and Hang Li. “Text-to-Table: A New Way of Information Extraction.” In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2518–2533. DOI: 10.18653/v1/2022.acl-long.180. URL: https://aclanthology.org/2022.acl-long.180.

[10] Elozino Egonmwan and Yllias Chali. “Transformer and seq2seq model for Paraphrase Generation.” In: *Proceedings of the 3rd Workshop on Neural Generation and Translation.* Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 249–255. DOI: 10.18653/v1/D19-5627. URL: https://aclanthology.org/D19-5627.

[11] Chujie Zheng et al. “Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization.” In: *CoRR* abs/2108.11992 (2021). arXiv: 2108.11992. URL: https://arxiv.org/abs/2108.11992.

[12] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In: *CoRR* abs/1910.13461 (2019). arXiv: 1910.13461. URL: http://arxiv.org/abs/1910.13461.

[13] Yu Yan et al. “ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training.” In: *CoRR* abs/2001.04063 (2020). arXiv: 2001.04063. URL: https://arxiv.org/abs/2001.04063.

[14] Ratish Puduppully, Li Dong, and Mirella Lapata. “Data-to-Text Generation with Content Selection and Planning.” In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 6908–6915. DOI: 10.1609/aaai.v33i01.33016908. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4668.

[15] Zdeněk Kasner and Ondřej Dušek. *Neural Pipeline for Zero-Shot Data-to-Text Generation.* 2022. arXiv: 2203.16279 [cs.CL].

[16] Song Duong et al. *Learning from Multiple Sources for Data-to-Text and Text-to-Data.* 2023. arXiv: 2302.11269 [cs.LG].

[17] H. A. Z. Sameen Shahgir and Khondker Salman Sayeed. *Bangla Grammatical Error Detection Using T5 Transformer Model.* 2023. arXiv: 2303.10612 [cs.CL].

[18] HAZ Sameen Shahgir, Ramisa Alam, and Md. Zarif Ul Alam. *BanglaCoNER: Towards Robust Bangla Complex Named Entity Recognition.* 2023. arXiv: 2303.09306 [cs.CL].

[19] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* 2020. arXiv: 1910.10683 [cs.LG].

[20] Linting Xue et al. *mT5: A massively multilingual pre-trained text-to-text transformer.* 2021. arXiv: 2010.11934 [cs.CL].

[21] Ankit Agrawal et al. "BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling." In: *Applied Sciences* 12.3 (2022). ISSN: 2076-3417. DOI: 10.3390/app12030976. URL: https://www.mdpi.com/2076-3417/12/3/976.

[22] Abhik Bhattacharjee et al. *BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla*. 2023. arXiv: 2205.11081 `[cs.CL]`.

[23] Abhik Bhattacharjee et al. *CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs*. 2023. arXiv: 2112.08804 `[cs.CL]`.