

An Extensive Approach of Converting Chittagonian to Standard Bangla

BY

Sinthia Chowdhury

ID: 181-15-10599

AND

Deawan Rakin Ahamed Remal

ID: 181-15-10600

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori

Associate Professor and Associate Head

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Sheikh Abujar

Senior Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 202

APPROVAL

This Project titled “**An Extensive Approach of Converting Chittagonian to Standard Bangla**”, submitted by ***Sinthia Chowdhury*** and ***Deawan Rakin Ahamed Remal*** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on ***2nd January 2022***.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

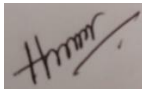
Chairman



Moushumi Zaman Bonny
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Mahfujur Rahman
Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md Arshad Ali
Associate Professor

Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

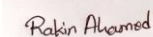


Mr. Sheikh Abujar
Lecturer (Senior Scale)
Department of CSE
Independent University (Study Leave)

Submitted by:



(Sinthia Chowdhury)
ID: 181-15-10599
Department of CSE
Daffodil International University



(Deawan Rakin Ahamed Remal)
ID: 181-15-10600
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Natural Language Processing*” to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Mr. Sheikh Abujar, Ms. Sharun Akter Khushbu, and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The demand of language translation is growing rapidly. Modern technology is offering the translation service for several languages. But the translation of different Bengali dialects is not very well explored filled in NLP. The goal of our work is to develop a proper conversion system which will convert the local form of Bangla spoken by the people of Chittagong to the standard Bangla language. The dialect of Chittagong district is one of the most different dialect in Bangladesh. We will use Natural Language Processing and Machine Learning to build an efficient system to convert the dialect to the standard form of Bangla language. So that a person who is not from Chittagong can easily communicate with the locals of Chittagong. To build this system we have used an LSTM based encoder decoder model. This model consists of LSTM, Embedding, Repeat Vector, Time Distributed dense layers. We have achieved 99.3% model accuracy for training and 72.5% model accuracy for testing sets. The accuracy we have achieved from our system is quite good. But the accuracy can rise if we can reduce our data loss more. As the data loss and the accuracy of our work is totally depending on the volume of the dataset. With time we will increase the volume of our existing dataset.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Rationale of the Study	2-3
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Report Layout	4
CHAPTER 2: BACKGROUND	5-8
2.1 Introduction	5
2.2 Related Works	5-7
2.3 Research Summary	7-8
2.4 Scope of the Problem	8
2.5 Challenges	8

CHAPTER 3: RESEARCH METHODOLOGY	9-18
3.1 Introduction	9-11
3.2 Research Subject and Instrumentation	11-12
3.3 Data Collection and Data Preprocessing	12-14
3.4 Statistical Analysis	14-15
3.5 Implementation Requirements	16-18
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	19-21
4.1 Introduction	19
4.2 Experimental Results	19-20
4.3 Descriptive Analysis	20-21
4.4 Summary	21
CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH	22-28
5.1 Summary of the Study	22
5.2 Conclusions	22
5.3 Recommendations	22-23
5.4 Implication for Further Study	23
REFERENCES	24-25

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: Workflow for Chittagonian to Standard Bangla Conversion	10
Figure 3.3.a.1: Preprocessing Steps	14
Figure 3.4.1: Dataset Overview	15
Figure 3.5.c: Architecture of a LSTM unit	17

LIST OF TABLES

TABLES	PAGE NO
Table 1: BLEU Scores and Model Configuration	20
Table 2: Training and testing results	20-21

Chapter 1

Introduction

1.1 Introduction

The conception of language has been born from the requirement to communicate with one another in the human community. Before the creation of language, people used to intercommunicate using body language. But it was very incommodious to express the genuine commentary of the speaker. Even after the creation of various rich languages, it was earlier difficult to communicate in a foreign land. We used to need a human translator for that purpose. Employing a translator was overpriced. To cut down the suffering modern science has developed the concept of machine translation. Where we confide in machines as an alternative to human translators. As a result, machine translation is now one of the most admired fields among researchers. Google is one of the most popular search engines around the world and it proclaims that around 500 million users were making use of their facilitation to translate more than 100 billion words daily by April 2016. The actual development in any linguistic area depends on historical events.

1.2 Motivation

Dialects of a language are the reflection of it. Different dialects are almost never entirely the same. A dialect is predominantly a diversity of a language that is distinctive of a precise territory [1]. Speakers of nearly connected languages can communicate to a definite boundary each using their mother tongue. In the progressed communities, the dissimilarity between dialects and related standard languages is effortless to make. The elementary cause of the dialectal difference is basically linguistic variation. Every living language is constantly enduring changes in numerous portions. Because languages are exceptionally

complex. Linguistic evolution can influence some parts and change them in a similar way in all areas where one language is spoken. Standard languages arise when a particular dialect is used in written form. It can be entirely in a particular area. Bangla is a very rich language with some very popular dialectal forms.

Chittagongian dialect is one of the richest dialects the Bangla language has but the speakers are now migrating towards using more standard Bangla than this dialect. As a result, we are in danger of losing this amazing dialect soon. So we decided to build a system to help outsiders to communicate using this dialect and keep this dialect alive. Dialects are the beauty of a language and our aim is to keep the dialect safe from disappearance. There are numerous works on the basis of Bangla language translation and the need for such assignment is sublime [2]. But the obstruction is users can not transfigure Bangla dialect using this service. As a result, the most admired and exceedingly used dialects are dissolving such as the dialect of Chittagong province. As well as for the migration of speakers of different dialects, these dialects are being adulterated [3]. The dialect of Chittagong is ranked 67th in ubiquity and it is estimated to have 14 million speakers. So we are trying to uphold the magnitude of the Chittagonian dialect here. This dialect is immensely affluent and a concoction of words from poles apart languages but does not have any written appearance to date [4]. So we want to build a written form of this dialect in our work. Again parents are encouraging the new generation not to use dialect which is causing dialect to fade away [5]. As a result, it is nearly unattainable for an outlander to acknowledge the dialect. So we are building a system to translate the Chittagonian dialect into the Standard Bangla language. There is no significant work based on this approach.

1.3 Rational of the study

The dialects standard Bangla language have is very large in number. Among these dialects Chittagongian dialect is one of the toughest to understand as an outsider having a huge number of speakers globally. But the speakers are now decreasing the use of this dialect in

their daily lives. So to make the speakers understand the priority of their mother tongue we decided to introduce a system to reduce the language barrier. So we are using the Machine Learning approach to build a system to translate standard Bangla to Chittagongian dialect. Machine Learning is a huge field with many models and methods. But for our system we are using the LSTM model for translation purposes.

1.4 Research Questions

- ❖ What is Machine Translation?
- ❖ What is the concept of dialect?
- ❖ How does Machine Translation work?
- ❖ Why is dialectal translation so important?
- ❖ In Machine Learning, how do you preprocess Bangla text data?
- ❖ What are the plans for Bangla Machine Translation in the future?

1.5 Expected Outcome

Chittagong is a port area in southeast Bangladesh. This area consists of the character of people from a different culture. The culture of the Buddhist people from Arakan is very different from the culture of the Muslims who migrated from Mughal. This area was ruled by rulers from Pre-Mughals to Portuguese and European to British. The mixture of different cultures has made the dialect so difficult that it is almost impossible for an outsider to understand their dialect and communicate. That is why we will discuss a way to convert Chittagongian to standard Bangla to minimize this kind of communication gap. Again Chittagong language has been ranked 88 as a dialect with 13 million speakers. People are now inspired to use standard Bangla instead of Chittagongian dialect. We will make people understand the value of their dialect so that they keep using that and the outsider who does not have any knowledge about their dialect can use our system.

1.6 Report Layout

This report has a total of 5 chapters.

- Chapter 1 reflects an overview of our entire research work. It is divided in a few sections like, 1.1 Introduction 1.2 Motivation 1.3 Rational of the Study 1.4 Research Questions 1.5 Expected Output 1.6 Report Layout.
- Chapter 2 reflects about the Background Studies with some subsections like, 2.1 Introduction 2.2 Related Works 2.3 Research Summary 2.4 Scope of the Problem 2.5 Challenges.
- In Chapter 3 we have represented our Research Methodology with subsections like, 3.1 Introduction 3.2 Research Subject and Instrumentation 3.3 Data Collection and Data Preprocessing 3.4 Statistical Analysis 3.5 Implementation Requirements.
- In Chapter 4 we have discussed the Experimental Results and some Research Discussion with the subsections like, 4.1 Introduction 4.2 Experimental Results 4.3 Descriptive Analysis 4.4 Summary.
- Chapter 5 reflects the Conclusion and the Future Works of our research with the subsections like, 5.1 Summary of the Study 5.2 Conclusion 5.3 Recommendations 5.4 Implication for Further Study.

At the end we share all the references which is related, encouraged and helps us to explore and complete the research.

Chapter 2

Background

2.1 Introduction

Here we will discuss all the related work, research summary, and challenges we faced in building this system in this section. In the related work section, we will discuss previous research work related r to our work and also discuss their models and methods. We will also discuss the previous accuracy rate. In the summary part, We'll talk about a summary of relevant work. In the challenging part, we will focus on the challenges we faced while working and the ways to solve those.

2.2 Related work

The intention of a literature review is to attain an understanding of the remaining work. It provides elementary knowledge on the topic of research and helps to confine reiteration. Here we are executing a literature review about Machine Translation to find out the gaps and disputes in previous studies.

2.2.a Machine Translation of Language Assortment Synthesis

English language translation is one of the most practiced fields in NLP. Different researchers have explored this language for translation purposes. Research work by Himanshu Choudhary et al. [6] from 2018 represents English to Tamil translation based on Neural Machine Translation. Neural Machine Translation technique, word embedding along with Byte Pair Encoding, Sequence to Sequence Architecture, Attention Model are the used methods for this translation. EnTam V2.0 and Opu were the sources for data collection. They received a BLEU score of 4.58.

Preslav Nakov et al. [7] worked on English to Spanish Machine Translation. The dataset used was WMT'07 News Commentary test data. They achieved an accuracy of 35.78 and 35.17 BLEU scores respectively.

The attention Based Translation method was used by Shivkaran Singh et al. [8]. for English, to Panjabi translation. TDIL corpus, EMILLE corpus, Open source parallel corpus (OPUS) were the source for data collection. They attain a BLEU score of 26.07.

Dialect conversion is not a very well explored sector in Machine Translation. In 2015 a dialects conversion for Punjabi Malwai and Doabi Dialects was done by Arvinder Singh et al. [3]. For the research, they built their own dataset. The accuracy was 95% and 94% for Malwai and Doabi respectively. Their work was limited by the amount of the training data. In 2017 a Machine Translation work for Hindi to English was done by Omkar Dhariya et al. [9] Hybrid Machine Translation Model was used in this research. CFILT, HindiEnCorp version 0.5 was the source for the required data. They got an accuracy of 94.74%.

In 2020 Alexandre Lopes et al. [10] did work on Portuguese to English and vice versa. They used the Lite Training Strategy and T5 which is a pre-trained model. ParaCrawl, EMEA, CAPES Parallel Dataset, Scielo Dataset, JRC-Acquis, Biomedical Domain Parallel Corpora were used for data collection.

Francis M. Tyers et al. [11] in 2010 used a Rule-Based Approach in Apertium machine translation platform for Breton to French translation. the.

2.2.b Delineate Analogous Work For Bangla

Some principal work on the Bangla language will now be presented. Hafizur Rahman Milon et al. [4] in 2020 worked on the translation depending on the Rule-Based Negation Handling approach. To generate Bangla words Word-to-word mapping and Morphological Translation were used. They have developed a dataset. They got an average accuracy of 94.75%. But the obstacle was the quantity of the dataset.

In 2010 Bangla to English machine translation using morphological analyzer was done by Md. Sadequr Rahman et al. [12]. This was a very new but not sufficient tactic. Another work by Sk. Borhan Uddin et al. [13] represents Bangla to English text conversion depending on the OpenNLP Tools. This system contains five major functions: (1) Bangla Grammar Detection (2) POS Tagging (3) Bangla Parse Tree Generation (4) Bangla Parse Tree to English Parse Tree Matching, and (5) Bangla to English Text Translation. This system translates simple sentences systematically but translating complex sentences is difficult. They got an average of 40% accuracy. In 2019 Md. Arid Hasan et al. [14] worked on Bangla to English language pairs using Neural Machine Translation (NMT). The data was tokenized NMT architecture as BiLSTM, Attention model, Transformer, Self-Attention, Multi-Head Attention, Position wise Feed-Forward Networks, Positional Encoding. Data was collected from online sources for this research. They got an accuracy of 14.63% and 32.18% the SUPara and ILMPC test sets. A paper by Shaykh Siddique et al. [15] from 2020 represents the English to Bangla Machine Translation using RNN (Recurrent Neural Network). Context Vector and RNN were used in this work. The dataset was developed by them and it contains 4000 parallel sentences and 2839 unique English words and 3527 unique Bangla words.

2.3 Research Summary

In this research, we used machine learning methods for standard Bangla to Chittagongian translation. Data collection was a vital portion of our research work. Instead of focusing on a particular field for data collection, we tried to use various sources to gather as much data as we can. Our work totally depends on the amount of data we can gather. We have collected data from various books, newspapers, social media posts, and comments. Apart from these, we collected data from dramas and videos based on the Chittagongian language. Also, we took help from some native people to translate their language. Then we pre-

processed it and applied the data to our model and trained it. Our training provided us with decent accuracy.

2.4 Scope of the problem

Translation work based on Machine Learning is not a very new concept in the area of research. But there is no significant work based on Bangla dialect translation. So we choose to introduce Bangla dialect translation using Machine Learning. Because of a poor dataset, we could not get good enough results on our first attempt. But with time we increased the amount of data in our dataset and noticed a difference in our output. We used a Supervised algorithm in our translation approach for gaining a better result.

2.5 Challenges

The main challenge in our research was data collection. There was no available dataset for our work and as a result, we had to build our own dataset. For the construction of the dataset, we took help from some native speakers. But the problem there was as we were collecting data from a dialect-based work and each speaker pronounced words differently. For that, the written form of our dataset was very hard to build. The words in the dataset were spelled differently as they were pronounced differently. Constructing a dataset where the same word has the same spelling was a hard nut to crack. Again we used Deep Learning and it only provides appropriate results when the dataset is rich enough. Working with less data produces a result that is not good for the system.

Chapter 3

RESEARCH METHODOLOGY

3.1 Introduction

Here, we will examine the entire strategy of the exploration action. We will discuss the methods and techniques we used to get a better result as our output. We will display a short description of each vital part of our research.

In our examination, we have used the Machine learning model for translation. Machine learning algorithms are used to include computer learning from provided data so that they do particular tasks.

The Machine Learning approach is classified into three classes:

i) Supervised learning ii) Unsupervised learning iii) Reinforcement learning

For this research, we have developed a layered model consisting of LSTM, Embedding, Repeat Vector, Time Distributed Dense layer. The machine learning model requires a satisfactory dataset to produce a correct outcome. To apply the calculation, the dataset should be gathered and preprocessed. Then, each segment of the methodology is examined separately. Given all areas are followed when the examination work is finished. The process of our work is shortly described with a diagram below

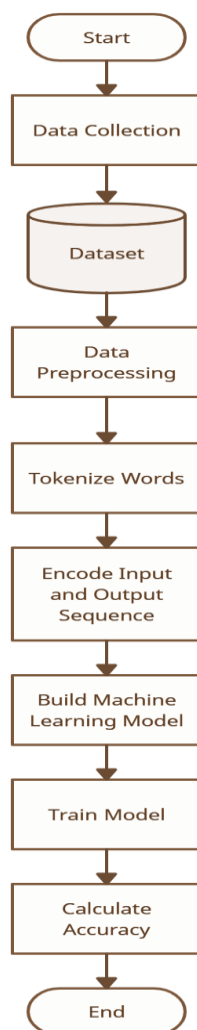


Figure 3.1.1: Workflow for Chittagonian to Standard Bangla Conversion

Data gathering is the primary step containing Bangla and Chittagonian dialects. Then we have created pairs for similar sentences. Then the noise from the dataset was cleaned and developed a clean pair. Then we used Pickle API to store the pair of the clean text to file. The dataset is then suffered randomly. Because we wanted to remove similar context data. The dataset is distributed into 80:20 ratio for train and test purposes. Per words are separated and tokenized to create a vocabulary. When we were generating a sentence we took these tokens and checked in the vocabulary. We will encode the input for Chittagonian and output for the Bangla language. Then one hot encoding for output

sequencing. Because the model will predict the probability of each word in the vocabulary for output. The encoder-decoder LSTM model is used here. The architecture of the model input sequence is encoded by a front-end model encoder then decoded word by word by a backend model the decoder. To develop the model sequential API was used because it allows creating model layer by layer. In this model LSTM, Embedding, Repeat Vector, Time Distributed Dense layer is used. The model is trained using Adam approach to stochastic gradient descent and minimizes the categorical loss function. Because we have framed the prediction problems as multiclass classification. The model is trained for 100 epochs and a batch size of 2 examples. Here we used checkpointing to ensure that each time the model skill on the test set improves the model is saved to file. After that the score of the model and the BLEU for the translation.

3.2 Research Subject and Instrumentation

Our research topic is “An extensive approach of converting Chittagongian to Standard Bangla”. This is a unique work based on the Bangla dialect. Research work based on the Bangla dialect is a new field and has not been explored widely yet. This is significant work in Bangla NLP. A layered model is used to do the translation work and a good set of computers with GPU and other instruments is necessary for that. The list of instruments used in this work is given below:

Hardware and Software:

- Intel Core i5 including minimum 16GB RAM
- 1 TB HDD
- 250 GB SSD
- Google Colab including 12GB GPU and 350 GB RAM

Advancement Tools:

- Windows 10
- Python 3.7
- Pickle
- TensorFlow Backend Engine
- Matplotlib
- NumPy

3.3 Data Collection and Preprocessing

The toughest part of our work was the collection of data. It is a unique work and there is no significant work using Bangla dialect. Our model demands a huge dataset and there is no available dataset. So we had to build our own dataset. The technique and source for mapping Chittagongian and standard Bangla sentences were not available. So initially we took data from [16]. Different dramas were also very helpful for data collection. We have taken some standard Bangla sentences from the newspaper, stories, and articles and translated them with the help of the native Chittagong people. Translating word by word from standard Bangla to Chittagongian is not an easy task. For word-by-word translation, we have created a dictionary. The dictionary helped us to avoid misspelling and the wrong translation of any word. We then used different social media platforms for data collection. Comments and posts written in the Chittagongian dialect were useful for data collection. We took data from posts and comments and then translated it word by word with the help of our own built dictionary. We watched dramas and different videos based on Chittagongian dialect and took data, and converted those into text format from video and audio format. We did the same for Chittagongian folk songs for data collection.

There are not many Chittagongian folk songs available online. Most of these songs do not have any subtitles. So it was very challenging for us to collect data from these sources. Again working with dialect comes with another challenge which is pronunciation of words. Same words are sometimes pronounced differently in different areas of a large region. Depending on the pronunciation of the speakers it was quite difficult to convert data into text form. We also used the application “Chittagong Language চট্টগ্রামের ভাষা” to collect sentences for our dataset. Songs based on Chittagongian dialect were also a good source for our data collection. We avoided any hypothetical meaning of the word and used only the actual meaning. Using this technique, we avoided the possibility of having multiple meanings of a single word. Words with multiple meanings are a big problem in the sector of translation. It reduces the accuracy of the translation and low accuracy is not appreciated. We used the same spelling for only one word. No different words have the same spelling in our dataset. Again we have not used the same sentences multiple times. It provided us with the redundancy of the same data. These were the methods we used for data collection. To apply models in this dataset we had to pre-process it.

3.3.a Data Cleaning

Data Cleaning is a vital part of the research domain. You cannot get better results without a clean dataset. Also, we need to clean data to apply the working models and codes. In this part, we manually check extra spaces, punctuation marks, incomplete sentences, inappropriately organized sentences and remove them. Again using the code we remove the punctuation marks. The pre-processing steps are given below:

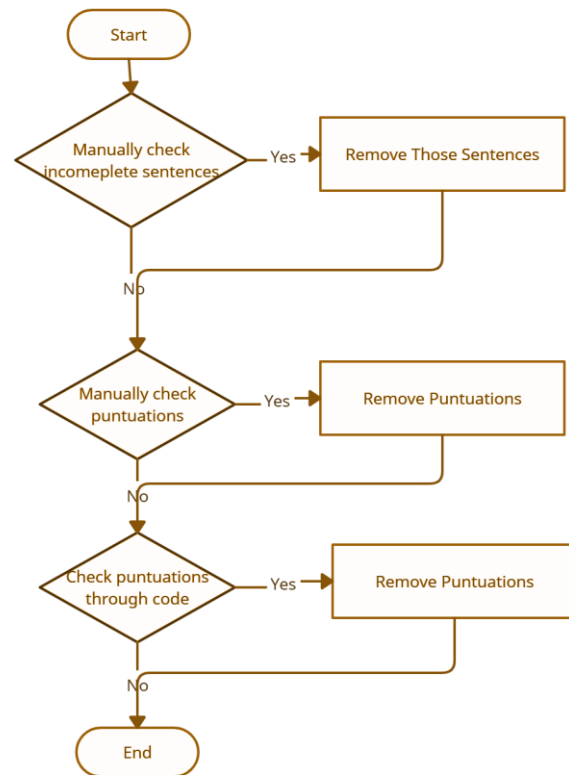


Figure 3.3.a.1: Preprocessing Steps

3.4 Statistical Analysis

1. Our dataset represents Bangla and Chittagonian sentences for Chittagonian to Bangla translation. The dataset consists of 4005 unique sentences. We built this dataset on our own.
2. Dataset has two parts, one is for Bangla sentences and another is for Chittagonian Sentences.
3. Sentences that remain the same even after conversion is placed in the same line in our dataset.
4. Both sentences are separated by tab characters.
5. Lines are separated by new line characters.
6. Our dataset is stored in a text file where the augmentation is .txt.

7. We used 100% of our dataset.

A short preview of our dataset is shown below

একদিন বুড়ি চিন্তা করল তার মুরগি	একদিন বুড়ি চিন্তা করিল কুরো
আমি যে খেতে দেই যদি	অ্যাঁই যে হাইতে দ্যা যদি
সে একটা করে ডিম দেয়	ইতে উজ্জো গরে ডিম দে
তার খাবার দিলে বেশি করে ডিম দিবে	ইতার হানা দিলে বেশি গরে ডিম দিতি
আর আমার বেশি টাকা ঘরে আসবে	আর অ্যাঁর বেশি টিয়া ঘরত আইবু
বুড়ি মুরগির খাওয়া বেশি করে দেয়	বুড়ি কুরোর হাওন বেশি গরে দে
মুরগি বেশি খাবার খেয়ে বেশি করে ডিম দেয়	কুরো বেশি হানা হাই বেশি গরে ডিম দে
কিন্তু পরের দিন আর ডিম দেয় না	কিন্তু ফরর দিন আর ডিম ন দে
আবার ডিম দিয়েছে মুরগি	আবার ডিম দিইয়্যা কুরো
মুরগি একটি করে ডিম দিতে লাগে	কুরো উজ্জো গরে ডিম দন লার
সে ডিম দেওয়া বন্ধ করে দেয়	ইতে ডিম দিবে বন্ধ গরে দে
একটা বড় ব্যবসায়ী ছিল	উজ্জো উঁর ব্যবসায়ী আইসসিল
তার ছিল অনেক ধন সম্পত্তি	ইতার আইসসিল বহুত ধন সম্পত্তি
তার একটা বড় কুকুর ছিল	ইতার উজ্জো উঁর কুত্তা
দুইটি ভাই অনেক বছর বড়	দুয়ো বন্দা বহুত বছর উঁর
বয়স যখন বেশি	বয়স যেতে বেশি
আর তখন থেকে দেখা করে আর লেখাপড়া খরচ দেয়	আর তখন তুন দেখা গরে আর লেখাপড়া খরচ দে
এখন অনেক বড় একটা বাড়ি	এখন বহুত উঁর উজ্জো বাড়ি
আর কথা বলার সুযোগ দেয় না	আর হথা হইবের সুজুগ ন দে
তাকে নিয়ে চলে আসে	ইতারে লইয়ে চলি অ্যাঁয়ে
অনেক ভয় পেয়েছে বুঝা যাচ্ছে	বহুত ডর পাইয়ে বুজা যার
ভয়ে কিছু করতে যেতে চায় না	ডরে কিছু গরিতে যাইতাম ন চায়
এসব দেখে তাকে সবাই বলল	এগিন চ্যাঁই ইতারে বেজ্জনে হইলো
না আপনি যাবেন না	না অনে ন যাইবেন
আমি আর করবো না আজকে কিছু	অ্যাঁই আর ন গইজ্জুম আজিয়ে কিসু
তারা আবার আসবে	ইতারা আবার আইবু
আর আমাকে সাথে করে নিয়ে যাবে	আর অ্যাঁরে লগে গরে লইয়ে যাইবু
তুই আমাকে কি দিতে পারবি	তুঁই অ্যাঁরে কি দন ফারিবি
সে জন্য ওকে সুযোগ করে দিতে চাই	ইতে লাই সুজুগ গরে দন চাই
তোরা তো বুঝিস সব	তুঁরা তো বুঝস হক্কল
কি ভুল বলছি বলবেন	কি ভুল হইন্দি হইয়্যন
তোকে কেন যেতে দিবে বল তো	তোঁরে কিন্নাই যাইতাম দিতি হ তো
তোর কষ্ট হবে শাড়ি পড়ে	তুঁর কষ্ট আইবু শাড়ি ফড়ি
তাই বলছি সব করে নিতে	তাই হইন্দি হক্কল গরে লইত
আচ্ছা তুই কাজ করে শেষ কর	আইচ্ছা তুঁই কাম গরে শেষ গর

Figure 3.4.1: Dataset Overview

3.5 Implementation Requirements

In this section, we'll talk about our problem and how we're going to solve it.

3.5.a Problem Discussion

With this research, we want to minimize the communication barrier between Chittagonian and other dialect people. Also trying to save the beautiful dialect from us. To solve this problem, we developed a system having a layered model. By using a split method, we have split our dataset for training and testing and applied it to the model.

3.5.b Split Method

Data splitting means partitioning the data into two separate parts. Mainly we use it for cross-validation. Here one portion is for developing the model and the other is for evaluating the model performance. So we divide our dataset into two parts. One is for training and the other is for testing which is 80:20. Before splitting the dataset we have shuffled the whole dataset to avoid sequences of similar context. And after that, we train the model which consists of LSTM, Embedding, Repeat Vector, Time Distributed Dense layer.

3.5.c LSTM

Long Short Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order in sequence prediction problems. LSTM is used for complex problem domains like machine translation, speech recognition, text classification, and more. It's a complex area of deep learning. It is effective in predicting the long sequence

of data like sentences. It has a feedback loop in its architecture and a memory unit to withhold the past information for a longer time for making an effective prediction.

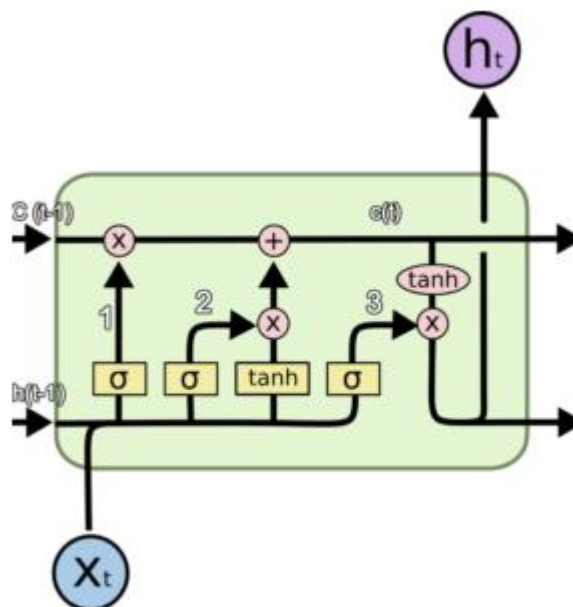


Figure 3.5.c: Architecture of a LSTM unit

3.5.d Embedding

The Embedding layer is characterized as the primary hidden layer of a network. It must indicate three arguments:

- **input_dim**: This can be the measure of the vocabulary within the content information. Like, in case your information is numbers encoded to values between 0 to 10, at that point the estimate of the vocabulary would be 11 words.
- **output_dim**: This can be the estimate of the vector space in which words will be embedded. It characterizes the measure of the output vectors from this layer for each word. Like, it can be 32 or 100 or indeed bigger.
- **input_length**: This is often the length of input arrangements, as you'd characterize for any input layer of a Keras model. For illustration, on the off chance that all of

your input records are 1000 words, this would be 1000.

3.5.e Repeat Vector

The Repeat Vector layer includes an additional dimension to the dataset. Like, on the off chance that you've got an input of shape (group measure, input measure) and you need to feed that to a GRU layer, you'll be able to utilize a Repeat Vector layer to change over the input to a tensor with shape. It acts as a bridge between the encoder and decoder modules. It plans the 2D array input for the primary LSTM layer in the decoder. The decoder layer is outlined to unfurl the encoding. Subsequently, the decoder layers are stacked within the turn around arrangement of the encoder.

3.5.f Time Distributed Layer

Time Distributed Dense applies the same thick layer to each time step amid GRU/LSTM cell unrolling. That's why the error function will be between the anticipated name grouping and the genuine label sequence.

The Dense layer will only be applied once in the final cell if `return sequences=False` is used. When RNNs are employed to solve classification issues, this is usually the case.

If `return sequences=True`, the Dense layer, like Time Distributed Dense, is utilized to apply at each timestep.

Chapter 4

Experimental Results and Discussion

4.1 Introduction

Language conversion is a very difficult task. And if you try to do that with the machine learning approach then the difficulty level goes up. For every work, accuracy prediction is very important. Translation of language is not an easy task. It is more difficult when one of the languages does not have a written form. After the collection of data, we need to preprocess the dataset to remove errors and incomplete sentences from the dataset. At that point, the dataset needs to be prepared for the model to learn the machine. Machine Learning offers a better result than any other approach to translation. It produces more appropriate results depending on the model. We trained our model using google collab. Google collab provides free GPU administration to the client. We got a maximum accuracy of 99.3%, 72.5% for train and test respectively.

4.2 Experimental Results

There is no machine on planet earth that offers 100% accurate results. Every machine has some value or energy loss. So we focused on how much accurate result we could produce with our model. Our model gives a reasonable result but sometimes it reacts to invalid data connected to the dataset. But the greatest number of the reactions suggest an ideal outcome. Our dataset contains 4005 sentences for both Bangla and Chittagongian dialects. The total dataset is used in our work. The dataset was separated into two parts: training and testing. The ratio for training and testing was 80:20. We used a customized multilayered model and got accuracy for the train 99.3% and for the test 72.5%. We have calculated the BLEU score to identify the translation quality. Below table shows these score along with model configuration:

TABLE 1: BLEU SCORES AND MODEL CONFIGURATION

Dataset Size	Splitted Dataset	Model Configuration	BLEU Scores
4005	Train: 3204, Test: 801	Epochs: 100 Batch Size: 2 Verbose: 1 Number of Memory Units: 64	Train: BLEU-1: 0.666609 BLEU-2: 0.529437 BLEU-3: 0.461442 BLEU-4: 0.335299 Test: BLEU-1: 0.302418 BLEU-2: 0.155284 BLEU-3: 0.119055 BLEU-4: 0.059103

4.3 Descriptive Analysis

We have used this model for English to German language translation and got an impressive result. Then we used our dataset in the model. When we used the dataset, we started with a tiny quantity of data and gradually grew it. We applied this strategy to find out the difference in output depending on the size of the data. Our finding is shown below with a related table:

TABLE 2: TRAINING AND TESTING RESULTS

Date	Data Volume	Value Loss	Model Accuracy	BLEU Scores
10.05.2021	150	4.63	-	BLEU-1: 0.00 BLEU-2: 0.00 BLEU-3: 0.00 BLEU-4: 0.00
30.05.2021	1291	2.39	-	BLEU-1: 0.073043 BLEU-2: 0.008674 BLEU-3: 0.053373 BLEU-4: 0.084062
19.06.2021	2378 (Less Cleaned)	2.63	-	BLEU Train: BLEU-1: 0.084194 BLEU-2: 0.010794 BLEU-3: 0.058274 BLEU-4: 0.088827 BLEU Test: BLEU-1: 0.080876

				BLEU-2: 0.014719 BLEU-3: 0.071372 BLEU-4: 0.105910
29.06.2021	1995	2.63	-	BLEU Train: BLEU-1: 0.084194 BLEU-2: 0.010794 BLEU-3: 0.058274 BLEU-4: 0.088827 BLEU Test: BLEU-1: 0.080876 BLEU-2: 0.014719 BLEU-3: 0.071372 BLEU-4: 0.105910
17.07.2021	3266	Train: 2.215, Test: 2.710	Train: 0.760, Test: 0.698	Train: BLEU-1: 0.142232 BLEU-2: 0.029212 BLEU-3: 0.118037 BLEU-4: 0.167353 Test: BLEU-1: 0.135553 BLEU-2: 0.025865 BLEU-3: 0.109653 BLEU-4: 0.157342
4.12.2021	4005	Train: 0.041, Test: 2.676	Train: 0.993, Test: 0.725	Train: BLEU-1: 0.666609 BLEU-2: 0.529437 BLEU-3: 0.461442 BLEU-4: 0.335299 Test: BLEU-1: 0.302418 BLEU-2: 0.155284 BLEU-3: 0.119055 BLEU-4: 0.059103

4.4 Summary

In short, we can say that we have applied the idea that we had dreamt of. We wanted to translate the Chittagongian dialect into the Standard Bangla language and at last, we did it. And we are pretty much successful in our attempt.

Chapter 5

Conclusion and Future Work

5.1 Summary of the Study

The main goal of this research is to create a system that can minimize the barrier between Chittagonian and other dialects. We have done this research with Machine Learning and created a customized model to solve the problem. As no remarkable work has been done yet and no publicly available dataset is there, we faced a lot of difficulties to complete the work. It takes almost a half year to complete. Our full work flow is given below:

5.2 Conclusion

In our work, we have created a system for Chittagonian to Bangla conversion. For this we created a layer model which consists of LSTM, Repeat Vector, Embedding and Time Distributed Dense layer. This system can convert Chittagonian sentences to Standard Bangla sentences. For this work, we needed a large amount of data. But there is no such work and dataset. So we made our own dataset. But the dataset is not rich enough for this work. That's why we gained very little BLEU score. But with our dataset the model performed well and got 99.3%, 72.5% accuracy for train and test respectively.

5.3 Recommendations

As our work is not giving that many BLEU scores due to the small number of datasets that's why first we will expand the dataset. Now our accuracy of the work is fully dependent on the size of our dataset. Further, we will try to create a better-layered model which will reduce the dataset dependency. Some recommendations are given below:

- More accurately clean the dataset
- Expand the dataset
- Create a better-layered model
- Try to get more accurate results

5.4 Implication for Further Study

Our research has some restrictions and limitations. Which inhibits us to produce the optimal result. Here we are using a model which is not very much optimized. In the future, we will work to make the model more optimized. Again our model is totally dataset dependent. So we will try to reduce the dependency and reduce the current value loss. Also we need to enrich our dataset to gain better output. Now our work is totally in text format further we will try to create a system which will be based on speech. Again there are many beautiful dialects available in the Bangla language. We will endeavour to work with such dialects in the future.

Reference

- [1] K. Marimuthu, S. L. Devi, “Automatic Conversion of Dialectal Tamil Text to Standard Written Tamil Text Using FSTs”, Proceedings of The 2014 Joint Meeting of SIGMORPHON and SIGFSM, pp. 37-45, June 2014.
- [2] M. N. Y. Ali, M. Z. H. Sarker, G. F. Ahmed, J. K. Das, “Conversion of Bangla Sentence into Universal Networking Language Expression”, International Journal of Computer Science Issues (IJCSI), vol. 8(2), pp. 64, 2011.
- [3] A. Singh, P. Singh, “Punjabi Dialects Conversion System for Malwai and Doabi Dialects”, Indian Journal of Science and Technology, vol. 8(27), pp. 1-6, 2015.
- [4] H. R. Milon, S. N. U. Sabbir, A. Inan, N. Hossain, “A Comprehensive Dialect Conversion Approach from Chittagonian to Standard Bangla”, 2020 IEEE Region 10 Symposium (TENSymp), pp. 214-217, June 2020.
- [5] M. A. Hoque, “Chittagonian Variety: Dialect, Language, or Semi-Language?”, IIUC STUDIES, 2015.
- [6] H. Choudhary, A. K. Pathak, R. R. Saha, P. Kumaraguru, “Neural Machine Translation for English-Tamil”, Proceedings of The Third Conference on Machine Translation: Shared Task Papers, vol. 2, pp. 770-775, October 2018.
- [7] P. Nakov, “Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing”, Proceedings of The Third Workshop on Statistical Machine Translation, pp. 147-150, June 2008.
- [8] S. Singh, M. A. Kumar, K. P. Soman, “Attention Based English to Punjabi Neural Machine Translation”, Journal of Intelligent & Fuzzy Systems, vol. 34(3), pp. 1551-1559, 2018.
- [9] O. Dhariya, S. Malviya, U. S. Tiwary, “A Hybrid Approach for Hindi-English Machine Translation”, 2017 International Conference on Information Networking (ICOIN), pp. 389-394, January 2017.
- [10] A. Lopes, R. Nogueira, R. Lotufo, H. Pedrini, “Lite Training Strategies for Portuguese-English and English-Portuguese Translation”, Proceedings of The 5th Conference on Machine Translation (WMT), arXiv preprint arXiv:2008.08769, 2020.
- [11] F. M. Tyers, “Rule-based Breton to French Machine Translation”, *EAMT*, 2010.

- [12] M. S. Rahman, M. F. Mridha, S. R. Poddar, M. N. Huda, “Open Morphological Machine Translation: Bangla to English”, 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 460-465, October 2010.
- [13] S. B. Uddin, D. M. F. Hossain, K. Biswas, “Bangla to English Text Conversion using OpenNLP Tools”, Journal of Science and Technology, vol. 8, January 2013.
- [14] M. A. Hasan, F. Alam, S. A. Chowdhury, N. Khan, “Neural Machine Translation for The Bangla-English Language Pair”, 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1-6, December 2019.
- [15] S. Siddique, T. Ahmed, M. R. A. Talukder, M. M. Uddin, “English to Bangla Machine Translation Using Recurrent Neural Network”, International Journal of Future Computer and Communication, vol. 9, arXiv preprint arXiv:2106.07225, 2021.
- [16] N. M. Rafiq, Chottogramer Ancholik Bhashar Ovidhan, 2nd Edition, Balaka Prokashon, 2017, pp. 209-266.

ORIGINALITY REPORT			
18%	16%	6%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%	
2	Submitted to Daffodil International University Student Paper	3%	
3	www.scribd.com Internet Source	2%	
4	machinelearningmastery.com Internet Source	2%	
5	Submitted to Federal University of Technology Student Paper	1%	
6	Submitted to Indian Institute of Technology Student Paper	<1%	
7	docplayer.net Internet Source	<1%	
8	Hafizur Rahman Milon, Sheikh Nasir Uddin Sabbir, Azfar Inan, Nahid Hossain. "A Comprehensive Dialect Conversion Approach from Chittagonian to Standard Bangla", 2020 IEEE Region 10 Symposium (TENSYP), 2020 Publication	<1%	