

Bangla Grammar and Spelling Check Using Machine Learning

by

Foysal Ahmed

19101535

Md Shahriar Khan

22241119

MD Emon Arafin

22241120

Abdullah Al Abir

22241118

Mumtahina Begum

19101306

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 2023

© 2023. Brac University
All rights reserved.

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.

Student's Full Name & Signature:



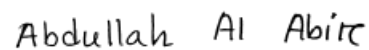
Foyzal Ahmed
19101535



Md Shahriar Khan
22241119



MD Emon Arafin
22241120



Abdullah Al Abir
22241118



Mumtahina Begum
19101306

Approval

The thesis titled “Bangla Grammar and Spelling Check Using Machine Learning” submitted by

1. Foysal Ahmed(19101535)
2. Md Shahriar Khan(22241119)
3. MD Emon Arafin(22241120)
4. Abdullah Al Abir(22241118)
5. Mumtahina Begum(19101306)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 19, 2023.

Examining Committee:

Supervisor:

(Member)

**Annajiat
Alim
Rasel** Digitally signed by
Annajiat Alim Rasel
DN: cn=Annajiat Alim
Rasel, o=Brac University,
ou=CSE Department,
email=annajiat@bracu.ac.
bd, c=BD
Date: 2023.01.14 23:04:00
+06'00'

Annajiat Alim Rasel

Senior Lecturer

Department of Computer Science and Engineering
Brac University

Co-Supervisor:

(Member)



Dewan Ziaul Karim

Lecturer

Department of Computer Science and Engineering
Brac University

Head of Department:

(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor

Department of Computer Science and Engineering
Brac University

Abstract

Bangla, or Bengali, is one of the world's most spoken languages, with hundreds of millions of native speakers worldwide. Thousands of books are written in the Bangla language every year, and millions of people register in Bangla daily. But there are only a few researches conducted on Bangla Grammar and Spelling correction because of the lack of Bangla resources and the complexity of the Bangla language. This paper is concerned with implementing a Machine Learning based model to detect grammar and spelling errors in Bangla writing. There are many machine learning algorithms to see mistakes in writing. This research uses Levenshtein distance and Double Metaphone algorithms to detect spelling errors. For grammar, Recurrent Neural Network based sequential model is used with an accuracy of 89%. We have created a Bangla monolingual corpus containing three hundred thousand sentences for this paper. Therefore, we expect this research to make Bangla writing easier and more fascinating for everyone.

Keywords: Bangla language; Machine Learning; Bangla Grammar and Spelling; Checker; Double Metaphone; Bangla Corpus; Neural Network

Acknowledgement

Firstly, all praise to the Almighty Allah for whom our thesis has been completed without any significant interruption.

Secondly, to our supervisor Mr. Annajiat Alim Rasel, and co-supervisor, Mr. Dewan Ziaul Karim, sir for their kind support and advice throughout our research.

And finally, to our parents, it may not be possible without their support and prayers.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Research Problem	1
1.2 Research Objectives	2
2 Literature Review	4
2.1 Related Works	4
3 Development of Corpus and Lexicon	7
3.1 Development of Corpus and Lexicon	7
3.1.1 Corpus	7
3.1.2 Lexicon	8
4 Development of Spell and Grammar Checker	10
4.1 Spell Checker and Correct Word Prediction	10
4.2 Grammar Checker	11
4.2.1 Context Pre-Processing	11
4.2.2 Word Embeddings	12
4.2.3 Sequence to Sequence Model	12
5 Result and Conclusion	14
5.1 Result Analysis	14
5.1.1 Data Training:	14
5.1.2 Result:	15
5.1.3 Future works:	15
5.2 Conclusion	16

List of Figures

3.1	Development of Corpus	8
4.1	Diagram of model	11
5.1	Epoch vs Loss graph	14
5.2	Epoch vs Accuracy graph	15

List of Tables

3.1	Unigram frequency distribution of corpus	9
4.1	Sample input for Spelling Checker	10

Chapter 1

Introduction

Bangla language is a member of the Indo-European family. Roughly 250 million people speak Bangla as their first or second language worldwide, mainly in Bangladesh and some regions of India like West Bengal, Assam, and Tripura[9]. Not only speaking, but Bangla is also rich in literature. Thousands of books are published every year in both Bangladesh and India. Some of the greatest minds of Bangla literature are Rabindranath Tagore, Kazi Nazrul Islam, Satyajit Ray, Sharat Chandra Chattopadhyay, Humayun Ahmed, etc. But Bangla is comparatively a more complex language than other languages. Bangla language has two primary written forms, “Sadhu Bhasha”, used mainly in the early 19th century and late 20th century. It has similarities with the Sanskrit language. Another form of Bangla written language is “Cholito Bhasha”, which is much more simplified than previous. These two forms of writing have huge differences. Grammatical rules, vocabulary, and pronunciation are completely different in most cases. So, while creating a grammar and spelling checker, both forms must be considered. Besides written forms, Bangla is prosperous with eleven vowels, forty consonants, and compound alphabets, whereas the English language only has twenty-six alphabets. So, checking for spelling mistakes is difficult in Bangla Language. There are also writing differences between Bangladesh and India. In Bangladesh, the Bangla writing rules and spelling of words are decided by Bangla Academy. Currently, a proofreader checks all the Bangla books for spelling and grammatical errors, which is time-consuming and tedious.

1.1 Research Problem

There have been many Machine Learning approaches to detect errors in English writing, but only a few researches have been done on Bangla language error detection. As Bangla is a complex language with various grammatical rules and a vast vocabulary, only a few papers are based on Bangla writing error detection.

Every day, the Bangla language is expanding. Not only by the number of native speakers but also Bangla books are being published every day for educational and entertainment purposes. Besides written literature, Bangla is also used for day-to-day communication, social media, and other virtual platforms. Bangla is also the official language of Bangladesh; all government files and documents are written in Bangla. So, a unified grammar and error-checking system for Bangla Language is necessary. For literature, the manuscript is sent for proofreading before publishing.

Proofreading is a time-consuming and costly process as it includes manual labor. While writing, there can be different types of errors. These can be categorized into three types. The first one is a Lexical error. A lexical error means the misspelling of a word. For example, “Office” is spelled as “Offyce”. The second one is Syntactic error, where words are spelled right, but there is an error in the organization of the terms. For instance, “people elderly” is used instead of “Elderly people”. Lastly, Semantic error means words in the sentence do not make sense. For example, “He is mopping the sky.” Here, the sentence is correct but does not have any meaning. The proposed system has to consider all the possible errors in the writing and suggest a correction.

Semantic errors involve typographical errors, writing errors, grammatical errors, homophone and homonym errors, etc. In these mistakes, the writer skips the alphabet or writes the wrong word in the text, but the term is still correct. So, semantic errors can not be ignored while checking for the error.

In Bangladesh, we follow Bangla Academy’s direction for spelling. But there are differences between Bangla Academy and West Bengal dialects and spelling. In this paper, we only train words for spelling from Bangla Academy. But even after only following Bangla Academy, there is a problem. As mentioned before, Bangla is expanding day by day. So is Bangla Academy. Bangla Academy is accepting new spelling for words and changing previously used spelling for the adaptability of the Bangla language. The proposed model must regularly be updated to the latest spelling corpus to prevent this kind of false error. Previously, most of the Bangla corpora collection was carried out manually or semi-automatically. Manual methods may not be appropriate for compiling incredible amounts of the dataset. Using corpora as training resource datasets in ML and Language models is unavoidable with the rising development of language technology. It has been demonstrated that an extensive dataset, even one that is noisy, consistently outperforms a smaller number of data in probabilistic machine learning. To create a monolingual Bangla corpus, we tried an automated data-gathering method. Automatic text collecting makes corpus data administration, electronic storage, and dynamic expansion easier. From the linguistics of canon, the ideal corpus size is not clearly defined. The wordstock of the principle, which estimates lexical variety, in general, ensures the optimality of the canon. From the linguistics of canon, the ideal corpus size is not clearly defined. The wordstock of the corpus, which estimates lexical variety, in general, ensures the optimality of the corpus.

1.2 Research Objectives

This research aims to develop an error detection system in Bangla texts, including grammatical and spelling errors using the Double Metaphone algorithm. The input data will be sent for pre-processing and encoding and then it will be checked for spelling and grammar errors. Correction suggestions will be suggested based on the detected errors. The objectives of this research are:

1. To deeply understand Machine Learning and how it works.
2. To deeply understand different errors in the Bangla language and different error detection algorithms.
3. To develop a model for writing error detection in the Bangla language using the Double Metaphone algorithm.

4. To evaluate the model.
5. To offer recommendations on improving the model.
6. Ultimately, to make Bangla writing correctly easier to mass people.

Chapter 2

Literature Review

With time technology is also flourishing day by day. With this rapidly changing technology more and more people are showing their immense interest in the use of technology and with great interest, there comes great demand for new technology which will make human life easier and more comfortable. Along with English grammar and spelling checking applications, there are increasing demands for Bangla spelling and grammar checks also. A common question always rises with this demand “Is it possible to cover the whole diversity of Bangla grammar and involve those in a single application?”. We can assume that the answer would be 100 percent positive or on the verge of that. The most important thing in this project is it needs to be upgradable as the whole literature of the Bangla language is so diverse and rapidly changing.

2.1 Related Works

This feature aspires to critically review previous relevant work in the field of Bangla spelling and grammar corrector systems in the context of machine learning and specifically in the context of the double Metaphone algorithm. We research the various methods utilized for the primary outcomes attained and we demonstrate how the Bangla spelling and grammar corrector system has its distinct challenges due to the complexity of the Bangla spelling and grammar rule, complicated dataset, and shortage of research in this field that makes the Bangla spelling and grammar corrector system more challenging.

A spell and grammar checker is undoubtedly necessary for Bangla native speakers because it is used by millions of people all over the world. According to the author, they have developed a merged spell and grammar corrector system that continuously catches different spelling and grammatical errors which is based on a huge amount of corpus and lexicon [8]. The problem with building this system was a shortage of research.

The complex orthographic Bangla grammar and spelling rules cause unexpected performance for most of the algorithms. Therefore, the double Metaphone algorithm can be used for spelling correctors and language model probability is a suitable option for grammar correctors [8].

In the subject of NLP, checking for faults in English composition is a crucial task. The main goal of this job is to identify and repair grammatical faults in English sentences. Grammatical error detection and correction are important applications

in fully automated spelling and punctuation of English texts and the ground of English learning aids. With English's expanding importance on a worldwide scale, great progress has been achieved in the method of identifying grammatical errors in English. This study suggests a brand-new, machine-learning-based technique for identifying grammatical errors in English text. This study first builds a Seq2Seq-based model for grammatical mistake identification. Second, this work leverages the Transformers model to provide a method for detecting and fixing grammatical errors [11].

A typographical error occurred when inputting the text document. Search engines, information retrieval, emails, and other applications necessitate user typing. To correct misspellings in such apps, a strong spell-checker is required. Spell-checkers for western languages, such as English, are quite powerful and can handle any form of spelling problem, however, spell-checkers for Indian languages, such as Hindi, Urdu, Bengali, Kannada, Assamese, and others, are very rudimentary. Traditional methodologies such as statistical and rule-based procedures are used to create these spell-checkers. This article introduces HINDIA, an innovative methodology for handling Hindi spelling problems, one of India's most widely spoken languages. It detects and corrects spelling errors using a deep-learning algorithm [10].

According to Andrew R. Golding, two approaches may be used to address context-sensitive spelling (Mitsubishi Electric Research Laboratories). The first method involves the presence of specific words at a particular distance from the uncertain target word, while the second method utilizes part-of-speech identifiers and word patterns. The procedure is to combine the evidence provided by the component technique before resolving the problem using the most persuasive evidence. For this, a novel hybrid approach utilizing Bayesian classifiers is given, and performance gains are shown [1].

Lawrence Philips discovered an advanced version of the phonetic algorithm as the Double Metaphone Algorithm. The work of this was to index the words with their pronunciation. Lawrence Philips brought the third version of this Metaphone algorithm in 2009 which is open to spelling the characteristics of all other languages with the English language [7].

Banglalekha dataset is a dataset that includes 84 characters, 50 Bangla basic characters, 10 Bangla numbers, and 24 compound characters. For each of the 84 characters, 2000 handwriting samples were gathered, scanned, and pre-processed. After removing errors and scribbles, the final dataset had 1,66,105 handwritten character pictures. The collection also includes labels that indicate the age and gender of the participants who provided the samples. This dataset could be used to investigate the impact of gender and age on handwriting in addition to optical handwriting recognition research [4].

It's not easy to auto-correct a missing word in a phrase. It is also considered more difficult for Bengali speakers. Our thorough analysis reveals that no substantial research on this topic has been done for the Bengali language. In this research, we offer a method for detecting missing words and providing a suggestion list that matches the missing word with an accuracy of 82.82 percent. The n-gram model was used to determine whether a word was missing between two words in a phrase. Then, after determining the likely words for the missing word, we utilized probability scoring to rank the proposal list [6].

Another novel method for the problem of Bangla sentence correction and auto-

completion is utilizing a sequence-to-sequence model with encoder-decoder architecture[5]. An innovative solution involves the employment of a bidirectional dynamic recurrent neural network (BDRNN) with long short-term memory (LSTM) cells as the encoder and a bespoke recurrent neural network with LSTM cells and an attention mechanism as the decoder. The attention mechanism is a crucial component of the proposed method, which enables the decoder to construct words based on distant context hidden in the input sequence of words and is therefore applicable to sentences of varying lengths. The authors have done an excellent job of explicating the various components and configurations used in the proposed method and the rationale behind their selection. The authors have also done an excellent job at creating a standard benchmark dataset for this work and achieving 79% accuracy on the test dataset, which is a good result. However, this research is limited to the Bangla language, and the dataset employed may not be particularly vast.

Chapter 3

Development of Corpus and Lexicon

3.1 Development of Corpus and Lexicon

3.1.1 Corpus

Corpus is a large collection of texts. In this paper, we have developed a corpus for spelling and grammar checking. We have collected text data from various sources. Most of the existing corpora contain news articles. We have collected news articles from an existing corpus SUMono[3]. The newspapers were most helpful as they have maintained article archives and are updated daily with fresh content. Every online newspaper story is typically written, edited, reviewed, categorized, and published by a small group of intelligent people. Along with this, we have manually collected the literature of famous Bengali writers like Kazi Nazrul Islam from websites. After the compilation of the text, our corpus has 1,712,745 (1.71 million) words in total. The development process of the test corpus is shown in figure 3.1:

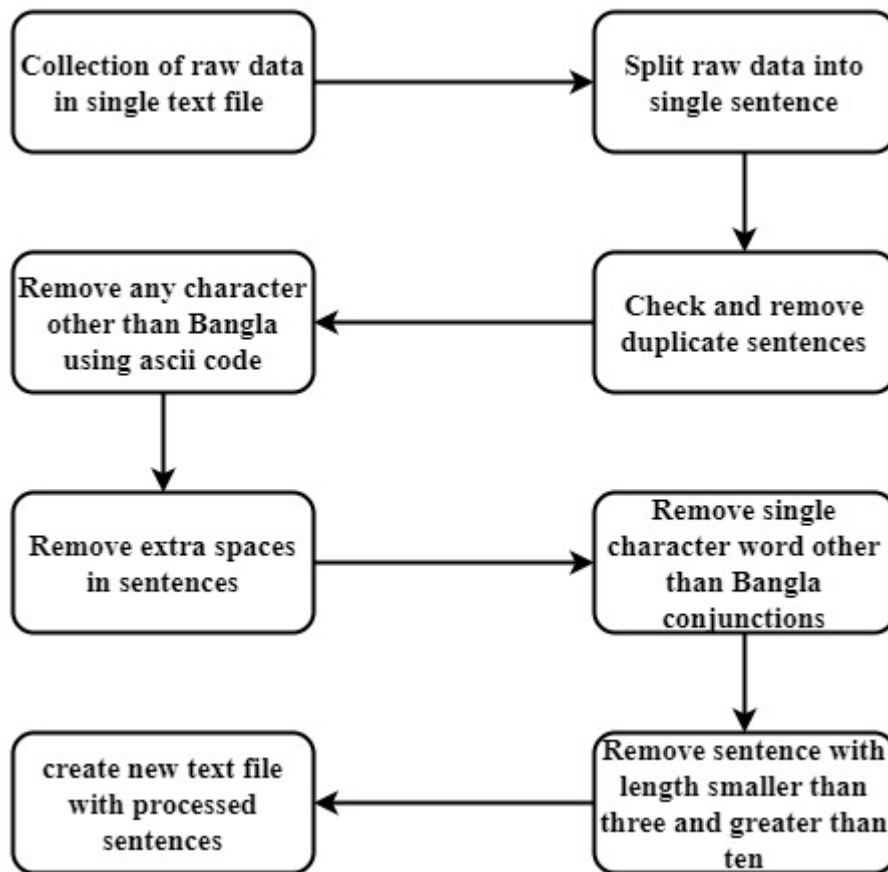


Figure 3.1: Development of Corpus

3.1.2 Lexicon

An index of lexemes is a lexicon. Bound morphemes, figurative phrases, and compound words are also included in lexicons. On the contrary hand, bound morphemes are typically absent from a regular dictionary, which only comprises root words. A lexicon is thus an essential component of a spell checker. Our vocabulary was developed by selecting words from the corpus. Although we manually collected the data and the existing corpus was processed, we had to remove all the punctuation marks and quotation marks from the corpus. The redundant lexemes are then removed from the lexicon file, leaving just the unique lexemes. To create the lexicon, we extracted all the unique lexemes from the corpus. In the lexicon, we got 136,802 (one hundred thirty-six thousand eight hundred and two) words. The unigram frequency distribution of the top twenty words is shown in table 3.1:

Word	Frequency	Percentage in corpus	Word	Frequency	Percentage in corpus
আমি	22602	1.32%	আমরা	8874	0.52%
না	21163	1.24%	আর	8690	0.51%
আমার	13081	0.77%	আমাদের	8507	0.50%
কি	12910	0.76%	করতে	7935	0.46%
এই	12740	0.75%	এবং	7782	0.46%
করে	11980	0.70%	হয়	7696	0.45%
করা	11099	0.65%	হবে	7557	0.44%
তুমি	10919	0.64%	আছে	7311	0.42%
জন্য	9635	0.56%	থেকে	7198	0.42%
এটা	9583	0.56%	তার	7092	0.41%

Table 3.1: Unigram frequency distribution of corpus

The articles are divided into parts during the tokenization phase, and the parts are then divided into words. Tokenization in Bangla is relatively simple compared to some other languages where white space does not resemble word boundaries. In Bangla, words are typically separated by white spaces or by punctuation marks. Punctuation marks, brackets, numerals, and hyphens are all regarded as word boundaries during the tokenization step. Stop words are also referred to as a language's function terms. Stop words and function words don't add to the text's meaning. Therefore, they are of little use in knowledge engineering or information retrieval. To find the content terms, words of this type were deleted from some of the corpora.

Chapter 4

Development of Spell and Grammar Checker

4.1 Spell Checker and Correct Word Prediction

For comprehensive spell-checking, phonetic algorithms are necessary. A phonetic algorithm called Soundex was created in the early 1900s [2]. An assortment of consonant letters with comparable sounds are phonetically encoded using the Soundex method. Apart from the vowel at the start of the word, vowels are often excluded from the procedure. The outdated phonetic algorithm Soundex has several flaws. Another phonetic algorithm known as Metaphone was created in 1990 by Lawrence Philips and it differs from the previous Soundex algorithm in several ways. The Metaphone method is more complex than the earlier algorithms since it has additional rules to deal with various spelling irregularities. The same inventor of Metaphone put forward Double Metaphone in 2000 [] which was an updated version of the original Metaphone. This incredibly complex phonetic algorithm has all the varied guidelines needed to control different elocution styles. Additionally, since Double Metaphone takes into account the pronunciation of various names, it makes name searches easier. For a single word, a double metaphone offers a major and a secondary code. We are now ahead of other phonetic algorithms thanks to this. Bangla contains several terms that can be pronounced differently depending on the circumstances, hence Double Metaphone provides the ideal answer. The edit distance, on the other hand, measures the distinction between two strings or words by computing the smallest number of operations needed to change one string into the other. Using the edit distance algorithm, we created the spelling checker which can detect wrong Bangla words and suggest the correct word based on Levenshtein's distance. Sample outputs are given below:

Input Word	Result	Prediction
শান্তিরক্ষক	Accurate	None
জনগ	Inaccurate	জনে, জগ, জনগন, জন, জন্ম, জনি, জাগ, জনেয

Table 4.1: Sample input for Spelling Checker

4.2 Grammar Checker

In this part, we will discuss the approach used in our work. Figure 4.1 depicts a system diagram of our methods for reference.

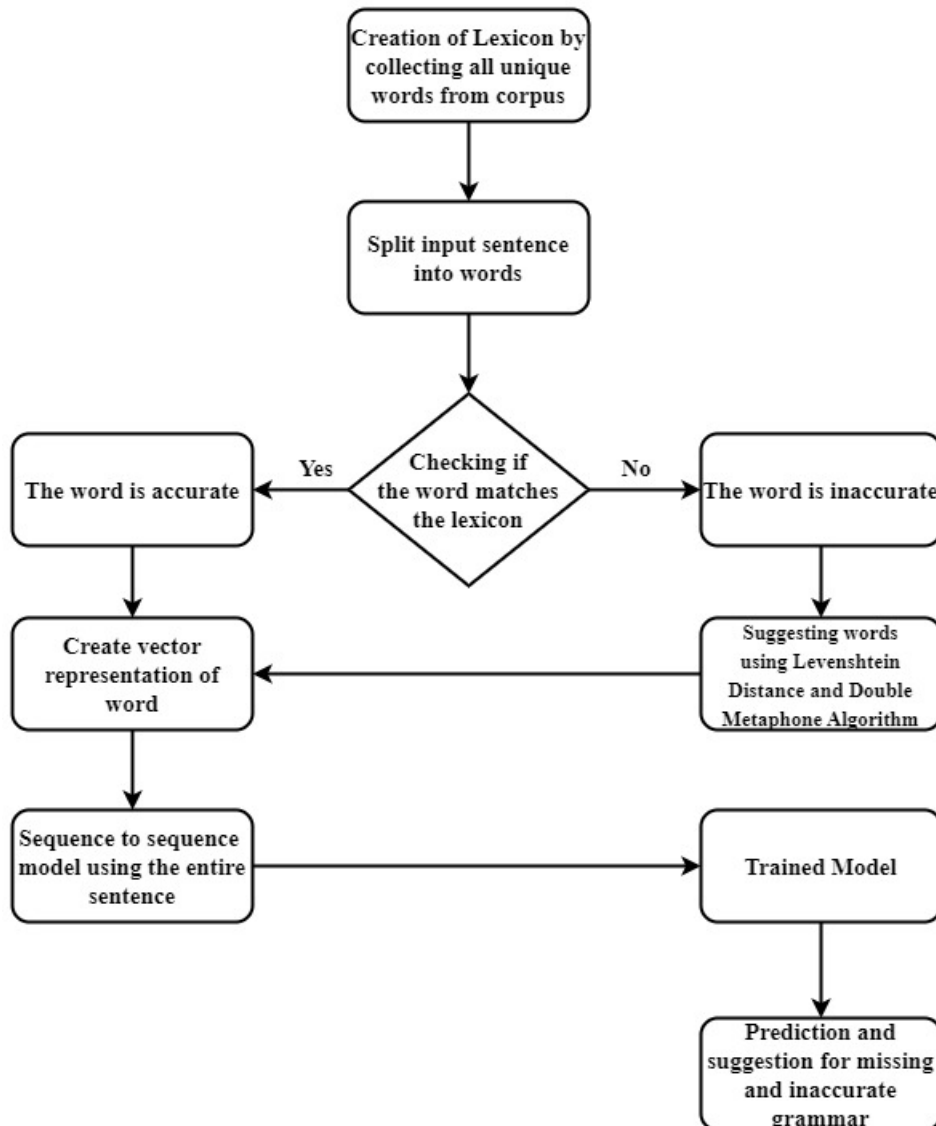


Figure 4.1: Diagram of model

4.2.1 Context Pre-Processing

The pre-processing for sentences was the following stage just after gathering Bangla sentences, which came after that. During this stage of the process, each sentence included in the dataset entered was judged to be a correct sentence. After that, we used random perturbation to construct the wrong sentences by inserting noise into every one of the sentences. Using a method that was both regulated and randomized, the following three categories of flaws were inserted through every sentence:

Auto-complete: A location at random was chosen inside the proper text provided, and the sentence itself was cut in half to create two portions. The first portion was

taken into consideration as the input sentence, as well as the second portion, which consisted of the auto-completed portion, was taken into consideration as the output.

Wrong Arrangement: A random choice was made of two words from the text, and their places inside the sentence have been switched around in order to produce a mis-arrangement mistake.

Missing Words: A word was randomly chosen inside the text and then removed to create the appearance of an absent word in the sentence. As a result, every gathered sentence was turned into a set of sentences. Following these steps, we produced a huge dataset consisting of 6 million input-output pairings.

4.2.2 Word Embeddings

During this stage, we built the lexicon and used tokenization to break down phrases. After that, the frequencies of appearance were determined for each word in the lexicon, taking into account all of the words' appearances throughout the entire dataset. Following the development of the lexicon, we have substituted the uncommon phrases and words with the sign **UNK**, which denotes the term **UNKNOWN**. We defined a term as unusual if it had a frequency three times less than the average. The word **NUM** was substituted for every instance of a number wherever it appeared in the text. The word embeddings that were constructed from the lexicon were then used to substitute every sentence pair that was a part of the previously processed dataset. After that, the frequency calculations were completed, and the lexicon was generated. A real numbering system was used to assign a number to each word in the corpus, along with the UNK token. The numbers were assigned to every word in the order of their frequency of occurrence in the lexicon. The encodings of the vector for input phrases may be seen in Fig. 4(b), which relates to the item illustration.

4.2.3 Sequence to Sequence Model

After effectively constructing the corpus, correctly translating the data with the help of words, and embedding it appropriately for this model, the next step is to train our model with the use of the sequence-to-sequence model. Encoder-decoder architecture is what we've utilized in the technique that we've suggested. Bidirectional Dynamic Recurrent Neural Network (BDRNN) describes the neural network we employ as an encoder. A specialized RNN that gives the user greater control is known as a decoder. The attention mechanism has accomplished communication between the encoder and decoder. Long-term, short-term memory cells, or LSTM cells, were employed for all RNNs. This part explains the proposed method's constituent components and configurations.

Recurrent Neural Network (RNN): The output for the current RNN is returned for the following input steps, creating a loop. Consequently, networks can predict a sequence at any time depending on the inputs and predictions made in prior phases, making them well-suited for sequences and lists. An unfolded RNN essentially equals a line of information-forwarding feed-forward networks. Figure 6 depicts the RNN as well as its comparable input network.

LSTM Cell: Assume that we are attempting to anticipate the final word given each word in natural language processing tasks such as sentence completion. The most current information in the network is required for the forecasting job. However, if the predictions rely on the environment, recurrent neural networks must need long-term memory cells.

Encoder-Decoder Architecture: An attention mechanism was used in the Encoder-Decoder structure. Figure 8 depicts our architecture's block diagram. Note that the input sentences for natural language processing and especially sentence corrections are of different lengths.

In addition, the output length may depend on the input phrase (in the case of missing words, etc.). Bidirectional RNN with LSTM cells is given all the vectorized words during the encoding step. Each cell receives data from left to right, and it learns words depending on the present state's and previous states' taken input. After that, the encoder creates a hidden layer as well as an output. It gathers the hidden layer of the encoder and develops content or words for the corrected or modified work. In this entire encoder-decoder system, the attention mechanism employs a representation of a vector between the decoder and encoder. It takes output for each unit as well as models for language by analyzing the word allocation from a broader perspective. Thus, the decoder can synthesize words depending on the distant information concealed in the input word sequence and can function with input sentences of different lengths.

Chapter 5

Result and Conclusion

5.1 Result Analysis

5.1.1 Data Training:

This part of the report represents an analysis of our training, testing data, accuracy, and potential loss of any epoch. Also, the required time to evaluate a particular train model concerning the test model is mentioned in this report section. Here the number of passes for training data (epoch) was divided into different segments and shuffled randomly. The total number of the epoch was 50. We used soft-max cross entropy to determine the loss margin as we faced a particular amount of value loss over training each period. The first epoch segment showed an epoch loss of 4.73, the highest of all the other epochs, and the lowest epoch loss was recorded at 0.15 in the forty-ninth epoch. The epoch loss gradually decreased with the following segments. The epoch loss in the tenth segment was 0.69, and the epoch loss in the fifteenth segment was 0.48. The graph of Epoch vs Loss is shown in figure 5.1:

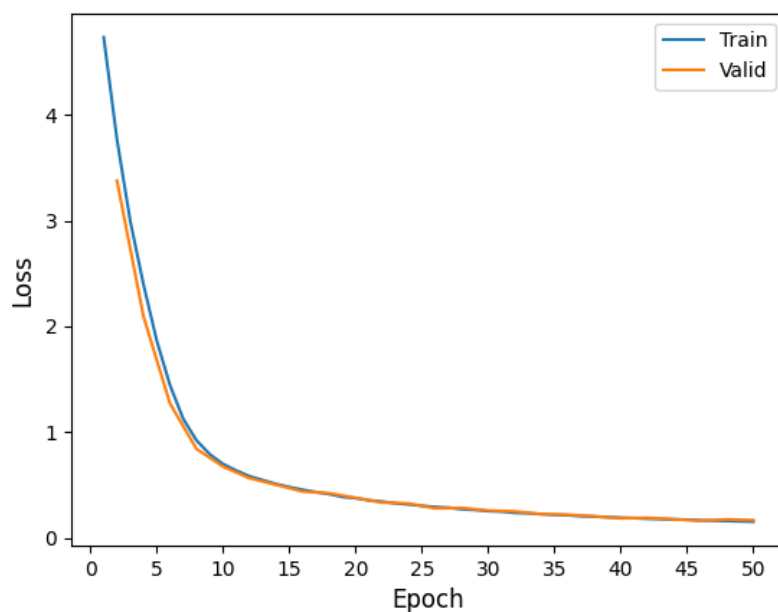


Figure 5.1: Epoch vs Loss graph

5.1.2 Result:

Our method did not over-fit the data set as the training loss function and validation loss function are almost similar. To make the output understandable, every sentence was taken in the network through encoding and decoded in the form of human understanding. Since our method did not over-fit, and we got almost similar loss functions in both the training loss function and validation loss function, we moved forward with the accuracy testing. In terms of testing accuracy, the lowest accuracy recorded was 33%, and the highest testing accuracy we could achieve was 89%.

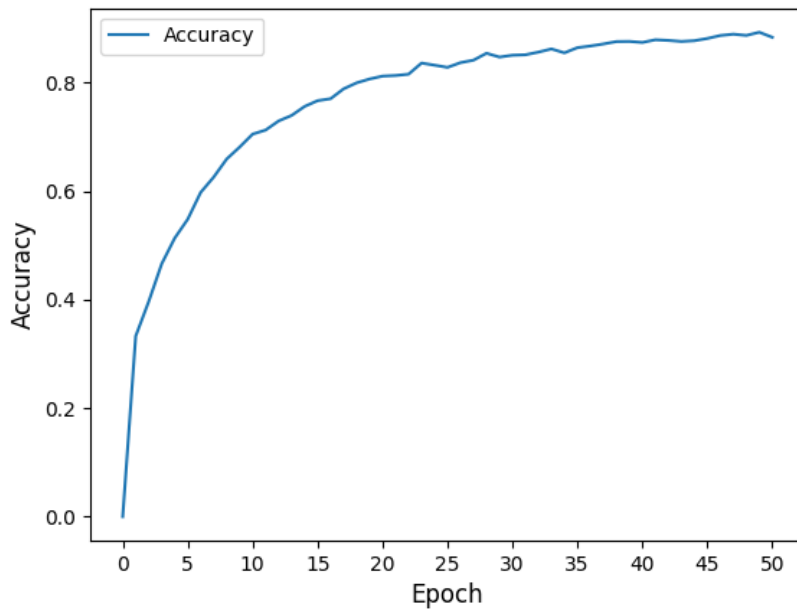


Figure 5.2: Epoch vs Accuracy graph

5.1.3 Future works:

So far we have accomplished a certain accuracy but we are always open and trying to emphasize more on further research with this model. For our further research target -

1. Concatenate this model with other models to have a better accuracy .
2. Making the dataset more diverse.
3. Updating the decoder more human-friendly.
4. Build an web-based application for Bangla Grammar and Spelling checking.

5.2 Conclusion

The suggested system includes a grammar and spelling checker for Bangla. The paper shows each stage of creating the corpus, lexicon, and spelling and grammar checking in detail. Any other limited resource language with a structure akin to Bengali can utilize a similar strategy. This can be used soon to create an open-source mobile app or to provide consumers with a browser extension. Users' consent will be required to collect helpful information for the system, which may be utilized to enhance the spelling and grammar checker program in its upcoming iteration.

Bibliography

- [1] A. R. Golding and Y. Schabes, “Combining Trigram-based and feature-based methods for context-sensitive spelling correction,” en, in *Proceedings of the 34th annual meeting on Association for Computational Linguistics -*, Santa Cruz, California: Association for Computational Linguistics, 1996, pp. 71–78. DOI: 10.3115/981863.981873. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=981863.981873> (visited on 01/15/2023).
- [2] M. T. Hoque and M. Kaykobad, “Coding system for bangla spell checker,” Dec. 2002, pp. 186–190.
- [3] M. A. A. Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, “Sumono: A representative modern bengali corpus,” *SUST Journal of Science and Technology*, vol. 21, no. 1, pp. 78–86, 2014.
- [4] M. Biswas, R. Islam, G. K. Shom, *et al.*, “BanglaLekha-Isolated: A multi-purpose comprehensive dataset of Handwritten Bangla Isolated characters,” en, *Data in Brief*, vol. 12, pp. 103–107, Jun. 2017, ISSN: 23523409. DOI: 10.1016/j.dib.2017.03.035. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2352340917301117> (visited on 01/15/2023).
- [5] S. Islam, M. F. Sarkar, T. Hussain, M. M. Hasan, D. M. Farid, and S. Shatabda, “Bangla sentence correction using deep neural network based sequence to sequence learning,” in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh: IEEE, Dec. 2018, pp. 1–6, ISBN: 9781538692424. DOI: 10.1109/ICCITECHN.2018.8631974. [Online]. Available: <https://ieeexplore.ieee.org/document/8631974/> (visited on 01/24/2023).
- [6] M. F. Mridha, M. M. Rana, M. A. Hamid, M. E. A. Khan, M. M. Ahmed, and M. T. Sultan, “An approach for detection and correction of missing word in bengali sentence,” in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox’sBazar, Bangladesh: IEEE, Feb. 2019, pp. 1–4, ISBN: 9781538691113. DOI: 10.1109/ECACE.2019.8679416. [Online]. Available: <https://ieeexplore.ieee.org/document/8679416/> (visited on 01/15/2023).
- [7] A. L. Fradkov, “Early history of machine learning,” en, *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1385–1390, 2020, ISSN: 24058963. DOI: 10.1016/j.ifacol.2020.12.1888. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2405896320325027> (visited on 01/15/2023).

- [8] N. Hossain, S. Islam, and M. N. Huda, “Development of bangla spell and grammar checkers: Resource creation and evaluation,” *IEEE Access*, vol. 9, pp. 141 079–141 097, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3119627. [Online]. Available: <https://ieeexplore.ieee.org/document/9568876/> (visited on 01/15/2023).
- [9] M. A. Jishan, K. R. Mahmud, A. K. A. Azad, M. R. A. Rashid, B. Paul, and M. S. Alam, “Bangla language textual image description by hybrid neural network model,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, p. 757, Feb. 2021, ISSN: 2502-4760, 2502-4752. DOI: 10.11591/ijeecs.v21.i2.pp757-767. [Online]. Available: <http://ijeecs.iaescore.com/index.php/IJECS/article/view/22296> (visited on 01/15/2023).
- [10] S. Singh and S. Singh, “HINDIA: A deep-learning-based model for spell-checking of Hindi language,” en, *Neural Computing and Applications*, vol. 33, no. 8, pp. 3825–3840, Apr. 2021, ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-020-05207-9. [Online]. Available: <https://link.springer.com/10.1007/s00521-020-05207-9> (visited on 01/15/2023).
- [11] J. Zhu, X. Shi, and S. Zhang, “Machine learning-based grammar error detection method in english composition,” en, *Scientific Programming*, vol. 2021, R. Ali, Ed., pp. 1–10, Dec. 2021, ISSN: 1875-919X, 1058-9244. DOI: 10.1155/2021/4213791. [Online]. Available: <https://www.hindawi.com/journals/sp/2021/4213791/> (visited on 01/15/2023).