

**ANN Based POS Tagging in Bangla
BY**

Md Mahmudul Hasan Pritom ID: 161-15-7026

AND Syed Nabil Azam ID: 161-15-6781

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

Supervised By

**Md. Tarek Habib Assistant Professor Department of CSE Daffodil
International University Co-Supervised By**

**Md. Jueal Mia Lecturer Department of CSE Daffodil International
University**



**DAFFODIL INTERNATIONAL UNIVERSITY DHAKA,
BANGLADESH DECEMBER 2019**

APPROVAL

This Research Project titled “ANN Based Parts of Speech Tagging in Bangla”, submitted by MD Mahmudul Hasan Pritom, ID No: 161-15-7026 and Syed Nabil Azam, ID No: 161-15-6781 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 07/12/2019.

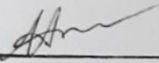
BOARD OF EXAMINERS



Dr. Syed Akhter Hossain
Professor and Head

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

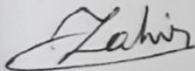
Chairman



Nazmun Nessa Moon
Assistant Professor

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

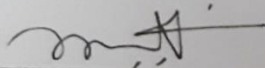
Internal Examiner



Gazi Zahirul Islam
Assistant Professor

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

Internal Examiner



Dr. Mohammad Shorif Uddin
Professor

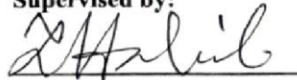
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

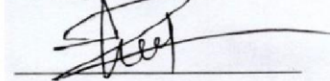
We hereby declare that, this project has been done by us under the supervision of **Md. Tarek Habib, Assistant Professor**, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

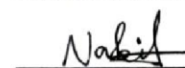


Md. Jucal Mia
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md Mahmudul Hasan Pritom
ID: 161-15-7026
Department of CSE
Daffodil International University



Syed Nabil Azam
ID: 161-15-6781
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to Almighty ALLAH 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 **Md. Tarek Habib, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge and keen interest of our supervisor in the field of “*Neural Networking*” 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 **Dr. Syed Akhter Hossain, Professor 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

POS Tagging in Bangla is a procedure of identify the parts of speech of given data sets. If any person does not know about the right Parts of Speech of a Bengali word then this project will be helpful for them. Trigrams'*n*'Tags (TnT) Tagging is a part of Hidden Markov Model Viterbi Algorithm. By the use of Trigrams'*n*'Tags (TnT) Tagging method Parts of Speech tagging become more easier. Trigrams'*n*'Tags (TnT) Tagging method is used here for tagging the unknown word by the use of external editable corpus which was been tagged before implement the algorithm. For research I have taken 30787 words and 1860 sentences for tagging from a newspaper web portal, and also took some data for which are untagged for tagged by this algorithm. Firstly I tagged the 30787 word for creating the editable corpus. Secondly I collect the untagged data, and then implement the algorithm on this untagged data for further tagging.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
Table of Contents	vi
List of Figures	ix
List of Tables	x

CHAPTER

CHAPTER 1: INTRODUCTION

1.1 Introduction	11
1.2 Motivation	12
1.3 Rationale of the Study	12
1.4 Research Question	13
1.5 Expected Output	13
1.6 Report Layout	13

CHAPTER 2: BACKGROUND STUDY

2.1 Introduction	14
	18
2.2 Related Works	
2.3 Research Summery	19
2.4 Scope of the Problem	19
2.5 Challanges	19

CHAPTER 3: Research Methodology

	20
3.1 Introduction	
3.2 Research Subject and Instrumentation	20
3.3 Data Collection Procedure	20
3.4 Statistical Analysis	20
	21
3.5 Implementaion Requirements	

CHAPTER 4: Experimental Results and Discussion

4.1 Introduction	23
4.2 Experimental Result	23
4.3 Descriptive Analysis	23

4.4 Summery	28
CHAPTER 5:	
SUMMERY,CONCLUSION,RECOMMENDATION AND	
IMPLICATION FOR FUTURE RESEARCH	
5.1 Summery of the Study	29
5.2 Conclusion	30
5.3 Implication for Further Study	30
APPENDIX	31
REFERENCES	32

LIST OF FIGURES

FIGURES	PAGE NO
Figure 4.2: Experimental Result	22
Figure 4.3: Corpus	23
Figure 4.3: Tagset	24
Figure 4.3: Untagged Dataset	24
Figure 4.3: New Tagged Dataset	25
Figure 4.3: Indian Corpus	25
Figure 4.3: Reading Bangla Dataset	25
Figure 4.3: Count Dataset	26
Figure 4.3: Untagged Data	26
Figure 4.3: New Tagged Outcome	26

LIST OF TABLES

TABLES	PAGE NO
Table 2.2: Comparison Table	16
Table 2.2: Comparison Table	17

CHAPTER 1

INTRODUCTION

1.1 Introduction

There are so many languages in the world. People use language to communicate with one another in their day to day life. Each and every country of the world has its own core language. People use language to express their opinion, desire and thought with others. Normally in a language, there are many things exist like sentence, word, letter etc. A sentence normally consists of numbers of letters and words. Basically Part of Speech refers every single word in a sentence. Every word is a part of speech that's depends on the usage of the words that used in a sentence.

Part-of-Speech Tagging is a process of identify a word in a text as corresponding to a specific part of speech, based on text's context. It is known as grammatical tagging. In the past, people tag words with their own intellectual but nowadays people are using some methods or algorithms to specify the part of speech of a sentence. Part of Speech (PoS) Tagging is much more complex than it seems because a word can refer other meanings of part of speech at many times. In natural languages it is very difficult to identify the tagging of part of speech because part of speech is not easy, much more complex for its unspoken situation or condition in a sentence [1]. In the research work, we are trying to identify the part of speech tag of words using types of Hidden Markov Model.

Bangla is one of the top most spoken languages in the world around 200 million native speakers. The grammatical rules of Bangla is derived from Sanskrit language. Although a large number of people use Bangla as their main or first language, still there has not been sufficient and appropriate research for Bangla in area of natural language processing. For our research, we used different toolkit and dataset. We used the toolkit to implement the algorithm for our experiment [2]. In this research, we used Trigrams'n'Tags (TnT) Tagging for our experiment. Trigrams'n'Tags (TnT) Tagging is used to identify the unspecified limit or boundary of a Hidden Markov Model (HMM). The main paradigm

used for smoothing is linear interpolation, the various weights area unit determined by deleted interpolation. Unknown words area unit handled by a suffix trie and serial abstraction [3].

1.2 Motivation

Every country has many different languages. People use languages to express their thoughts and ideas with one another. Language plays a part of any country and its people. People love their own language and use it various purposes for doing something better. People of different country, worked part of speech tagging on their main language. Part of speech is done in the past in languages like Hindi, Arabic, Portuguese, Marathi, Italian etc. using algorithms and methods. In Bangla language also people worked on it earlier using different types of algorithms and methods to solve tagging. As a people of Bengali nation, we wanted to work on this topic to tagging Bangla part of speech using various algorithms. In the past, people got success using many algorithms and methods to solve the Bangla Part of speech tagging, method like Brill, Trigram, Hidden Markov Method etc. that's why we wanted to worked on Trigrams'n'Tags (TnT) Tagging to solve the Bangla part of speech tagging because no one tried before with this tagset for tagging Bangla POS. This is the reason we are enthusiastic and passionate to working on it.

1.3 Rationale of the Study

In everyday life, we used many words and sentences to express our intentions and wants with others. We used so many words and its synonyms to complete our speech, let others know our thoughts about things.

Part of Speech Tagging is a process of identify and specify every word in a given sentence. A sentence normally consists of more than one words. After completing the tagging process, if we run the methods to show which types of part of speech are there in the sentence. It will tell us what kind of part of speech is out there in that sentence with efficiently and accurately. We are applying some methods and algorithms to predict the part of speech of the given sentences. We want to do this POS tagging with Bangla language. We will give some data that is sentences, which is consists of words as input.

Then we will apply some methods and algorithms of natural language processing so that as a result it will show us the tagging of words in those given sentences.

1.4 Research Questions

Part of Speech tagging helps to understand each and every word's part of speech in a sentence. Many words which we even don't know the meaning if we tag the word which used in a sentence with the context, we can easily know the unfamiliar word's part of speech. If the given input is accurate, then after applying different types of algorithms and methods on it still it will give us correct part of speech tagging of words with precise accuracy. The accuracies of various methods and algorithms can be different but it will definitely show us our desired result.

1.5 Expected Output

- A different approach to find out the part of speech of words using various kinds of methods and algorithms.
- The more dataset is used as input then the more accuracy it will give as output of this research.
- Safe and efficient ways to find out about unfamiliar words.
- Make easier to know about part of speech of words in Bangla.

1.6 Report Layout

- The research maintains general accurate content, lots of ideas about the topics and hypothesizes as a standard level.
- This research idea is supervised and authorized by people of in this fields.

CHAPTER 2

BACKGROUND

2.1 Introduction

There are many languages in the world. Every country has its main language and use many languages as their second language around the world. Specifically, Language is a communication system of humans that use of much more complex system. Today the number of languages of human in the world varies somewhere between 5000 and 7000 [4]. Humans learn language in his early childhood through interaction with other people.

Language evolves over time. The most widely spoken languages in the world as diverse as English, Hindi, Chinese, Bangla, French, German, Portuguese, Arabic etc.

As we are students computer science has much more effects than we thought in every aspects in our today's life. Now everything we do, we are doing with the help of these. So, we are tried to come up with an idea of our own Bangla language. An idea of part of speech tagging with Bangla words using some methods and algorithms of natural language processing. After talking and sharing of views with our honorable supervisor we decided finally to work on this research topic.

This research is about tagging Bangla words with using some NLP algorithms which will do better and make the Bangla word tagging much easier for others.

2.2 Related Works

There are many works done with Part of Speech Tagging before. But somehow our idea is different and we wanted to work on Bangla word part of speech tagging with Bangla Part of Speech tagset. No one did similar like this before as far as we know.

Part of Speech Bangla word tagging was done earlier with using some other algorithms and methods. But we are working with Trigrams'n'Tags(TnT) Tagging of Viterbi algorithm for Markov models.

ANN based POS Tagging for Nepali text, in this article there are three techniques used for solving POS for Nepali text tagging. For solving, they used Hidden Markov Model. They created two different tag set. They used Radial Basis Function, General Regression, Neural Networks and Feed Forward Neural Networks for solving POS tagging for Nepali text [5].

HMM based POS tagger in Hindi, the initial step for developing of NLP application is POS tagging. For developing Hindi POS tagging, the tagger machine translation and tagger also used Name Entity Recognition for which word tagged in noun. This project is a successful project [6].

Using Wiktionary to build an Italian POS tagger, the tagger of this project collected data or word from Wikipedia by Wiktionary and the tagger developed this project because in Italian language the documented resources are in limited number. The tagger used Brills method to developing this project [7].

Tagging Urdu sentences from English POS taggers, the tagger collected 10 sentences from twitter by twitter API and the data was raw and translated the data through google translator. After translating the data, they started to work with those (Urdu to English) data [8].

Implementation of Kadazan Tagger based on Brills method, Kadazan is a native language of Brunei and some native people of Malaysia. For Kadazan language, the POS tagging never been developed so the tagger eagerly wanted to developed POS tagging for their native language. The tagger used Brills method to developing this project [9].

A corpus based study of (kare) in Bangla , Bangla is a well-known language around the world. The taggers wanted to tagged (kare) by POS. The taggers collected the data from EMILLE corpus and Ananda Bazar corpus and tried to develop the project. But somehow the project become unsuccessful for some paucity of space [10].

Arabic POS tagging using Quran corpus, Arabic language is one of the most widely used languages in the world. POS tagging for Arabic language are relatively unexplored. For this reason the tagger compare the performance of some POS tagging techniques for Arabic using Quran corpus. These techniques include N-Gram, Brill, HMM, TNT taggers etc. The tagger wants to maximize the performances using those technique [11].

A feasible corpus for Persian POS tagging, in this paper, a description is given of a test collection for Persian POS tagged. Persian corpus with over two million tagged words. The original collection had a tag set of 550 tags that are more than what any machine learning algorithm can handle. It was created using Maximum Likelihood Estimation (MLE) for guessing the correct tags in Persian. The best accuracy that was achieved by MLE tagging was 95.43% [12].

POS tagging of Marathi text using Trigram method, in this paper the tagger presented a Marathi part of speech tagger. It is morphologically rich language. The general approach used for development of tagger is statistical using Trigram method. The main concept of Trigram is to explore the most likely POS for a token based on given information of previous two tags by calculating probabilities to determine which the best sequence of tags [13].

An unsupervised POS tagger for Bangla language, in this paper the tagger giving an overview of different approach to POS tagging and describe what has done so far for Bangla. The tagger described the POS tag set and the corpus used in that. They used 54 tag sets developed for Bangla language [14].

POS tagging in Portuguese language, a unified spelling system for Portuguese has been recently approved and its implementation process has already started in some countries. The POS taggers for Portuguese are specifically built for a particularly variety. This paper presents different dictionaries of the new orthography (Spelling Agreement) as well as a new freely available testing corpus, containing different varieties and textual typologies [15].

Related work for POS tagging

Ref.	Size of Data	Technique	Accuracy
1. ANN based POS tagging for Nepali text	42100 words testing set consists 6000 words	Hidden Markov Model, Corpus	98.32%
2. HMM based POS tagger for Hindi	358288 words, Test corpus 11720 words	Hidden Markov Model, Corpus	92%
3. Using Wiktionary to build an Italian POS tagger	100000 words	Brills method	92.9%
4. Tagging Urdu sentences for English POS taggers	10 sentences from twitter	Kappa Statistics	96.4%
5. Implementation of Kadazan tagger based on Brills method	5663 words	Brills method, Corpus	93%
6. POS tagging of Marathi text using Trigram method	2300 sentences, 48635 words	Trigram method, corpus	91.63%

7. An unsupervised POS tagger for Bangla language	18110 tokens and 4760 words	Baum-Welch algorithm, Corpus collected from newspaper	This project has not been successful
8. A corpus for Persian tagging	400000 words	Corpus	95%

Table 1: Comparison Table

Although there are so many efforts in tagging POS in Bangla language but no research became successful to develop an unsupervised data for POS tagging. We tried to tagged the datas and predict the Bangla POS using Trigrams'n'Tags (TnT) .

2.3 Research Summary

We gave so much effort to gain as much knowledge as we can for this project. We almost follow every paper written on POS tagging specially Bangla POS tagging. In the work has done on Bangla POS tagging but did not succeeded on the project.

We studied every paper written on POS tagging in different language around the world. That gave us idea about POS Tagging. Then we selected the papers according to our topic which will help us more. We read those selected papers start to finish and read those paper again and again to understand the earlier paper authors purposes and intentions. We tried to make every paper's summary to understand the core things and to make it easy for us on this project. We can remember some papers to make it clear.

Bengali Part of Speech tagging using Indian corpus: This is a paper where we get to know about development process of Bangla POS tagging and used algorithms and methods is also described in this paper.

Like these paper we also took the assist and help of some papers which is around 15 to 20 in number. We gathered information and knowledge, analyze them properly and applied in our research. These studies were very important for us. We would not be able to come this far without the help of these papers.

2.4 Scope of the Problem

Everything has an opposite side. The research we are doing was not successful before. So the chances of problems are not less. Problems can be happened anywhere.

We tried to make tagging of Bangla POS. We applied some algorithms in these but we want to check other algorithms implementing on Bangla POS tag sets.

2.5 Challenges

We have faced a lot of challenges while doing this research. We tried our best to justify with the research.

In coding we have faced so many challenges. After some time, we stuck in a situation where we could not understand how to solve this problem. We appeared with many similar challenges and we are tried to overcome those situations.

We had a little knowledge about NLP before doing this research we have to learn many things which is not easy for us to digest within short time.

We are gone through some ups and downs. Some parts of our code were not working properly at a certain time where everything was fine then the result was not accurate. But at last we are able to find the desired result and now we are hoping for the best.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

“ANN Based POS Tagging in Bangla” is the method where we can train the data easily and when we run the program it will check the other untagged dataset and tagged it and send us the output that which Bangla word is in which Parts of Speech.

3.2 Research Subject and Instrumentation

Our research topic's is Parts of Speech Tagging in Bengali Language. It is a famous topic's till now. This Project was implemented by various algorithm for different languages. For Bengali Language various algorithm implemented at the past.

We did our project by using Trigrams'n'Tags (TnT) Tagging. And by the Grace of Almighty we made it. We tagged the Bengali words from a news paper web portal by manually and the editor is Text Document.

3.3 Data Collection Procedure

In a Parts of Speech Tagging project need a large number of dataset. We collected the data from a Newspaper's web portal which name is The Prothom Alo. Approximately the dataset/Corpus contain above then 30787 word and 1860 sentences.

All of them are tagged by manually. And we need also some data for running the result which are untagged. We also collected the data from a Newspaper's Portal.

3.4 Statistical Analysis

The most important work for this research project is collecting data and tagging them manually. This is the hardest work in this project. After getting then we tagged them and made them as like a corpus. Tag set contains 8 different tags. After training the corpus we run the project by the use of software “Jupyter” and untagged data becomes tagged by the use of corpus. It shows us the actual outcome of the project (tagged data).

3.5 Implementation Requirements

In this section we are talking about the instruments what we need to do for this project.

This is a researched based project and this type of project doesn't need any types of hardware instruments.

All instruments we need for solving this project are Software based.

The software related things what we need are:

- Windows operating system(windows 10)
- NLP(Natural Language Processing)
- Python
- Jupyter
- NLTK(Natural Language Toolkit)
- Text Document

NLP (Natural Language Processing)

Natural language process (NLP) is a Artificial Intelligence based language which work with the Human Language. It may be a subfield of linguistics, engineering, info engineering, and computer science involved with the interactions between computers and human (natural) languages, specifically the way to program computers to method and analyze massive amounts of linguistic communication knowledge.

Python

Python is a High Level language for solving or programming on Artificial Intelligence based work.

NLTK (Natural Language Toolkit)

It's a library function for python in the sector of NLP (Natural Language Processing). It works for statistical NLP in English written in Python programming language.

Text Document

Text Document is a software where we can work with all types of text. By this software we can edit cut copy paste delete the text. Mainly all kind of works could do with text by this software.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

Experiment of a program is the most important part. Because it shows what has done in the coding side. From experiments we can find if there is anything wrong or any error happened in the coding side. After experiment we can sort out the error and resolve it.

4.2 Experimental Result

This is an experimental result. Our Project is optimistically done and the accuracy is 65%.

```
pos_tagger = tnt.TnT()  
pos_tagger.train(train_data)  
pos_tagger.evaluate(test_data)
```

```
Out[6]: 0.6538591169531514
```

Figure 4.2: Containing the Output

4.3 Descriptive Analysis

In This section we will tell everything about the code

4.3.1 Corpus

We have a large number of corpus but in the recent experiment we took only 1860 tagged sentences for experiment and our Corpus is

```

<Sentence id=1>
বাংলাদেশের B অগ্রগতি B উদাহরণ B দেওয়ার K মতোই O I_Sym
</Sentence>
<Sentence id=2>
অর্থনীতি B ও O আর্থসামাজিক B বেশির Bn ভাগ B সূচকে B বাংলাদেশ B ছাড়িয়ে K গেছে K দক্ষিণ এশিয়াকে B I_Sym
</Sentence>
<Sentence id=3>
নিম্ন Bn আয়ের B দেশগুলোকে B ছাড়িয়েছে K তো O অনেক O আগেই O
</Sentence>
<Sentence id=4>
আন্তর্জাতিক B মুদ্রা B তহবিল B (আইএমএফ) B গত O সপ্তাহেই O একটি Bn প্রতিবেদন B প্রকাশ B করে K বলেছে Bn ,,_Sym একটি K জনবহুল B ও O নিম্ন Bn আয়ের B দেশ B হিসেবে O
বাংলাদেশ B যেভাবে O প্রবৃদ্ধির Bn সঙ্গে O মারিযা Bn দূর K এবং O বৈষম্য Bn কমানোকে K সংযুক্ত B করেছে K ,,_Sym তা S অত্যন্ত Bn উদ্বেগযোগ্য B I_Sym
</Sentence>
<Sentence id=5>
সবাইকে S অন্তর্ভুক্ত B করে K প্রবৃদ্ধি Bn অর্জনের B ক্ষেত্রে O বাংলাদেশ B এখন Bn উদাহরণ B দেওয়ার K মতো O একটি Bn দেশ B I_Sym
</Sentence>
<Sentence id=6>
আবার O আরেকটি Bn আন্তর্জাতিক B দাতা B সংস্থা B বিশ্বব্যাপ্ত B একটি Bn টেবিল B উপস্থাপন B করে K দেখিয়েছে B ,,_Sym প্রধান Bn ১২ N টি Bn সূচকের B মধ্যে O ১০ N টিতেই Bn
বাংলাদেশ B দক্ষিণ এশিয়া B এবং O অন্য S নিম্ন Bn আয়ের B দেশের B তুলনায় K এগিয়ে K গেছে K বা O যাচ্ছে K I_Sym
</Sentence>
<Sentence id=7>
তবে O অন্তর্ভুক্তিমূলক Bn প্রবৃদ্ধি Bn অর্জনে B বাংলাদেশের B বড় Bn ধরনের B সাফল্য Bn থাকলে K ও O রাজনীতি B চলাছে K ঠিক Bn উল্টো Bn পথে Bn I_Sym
</Sentence>
<Sentence id=8>
অন্তর্ভুক্তিমূলক Bn রাজনীতির B অভাবে Bn অর্থনীতির B সাফল্য Bn পিছিয়ে K পড়ছে K বলে K মনে B করছেন K অর্থনীতিবিদ B ও O বিশেষজ্ঞরা B I_Sym
</Sentence>
<Sentence id=9>
অল Bn য়ে K পড়ছে K দেশের B অর্থনীতি B ,,_Sym কমছে K প্রবৃদ্ধি Bn I_Sym
</Sentence>
<Sentence id=10>
ফলে B সামাজিক B সূচকগুলো B ও O হ্রাসের Bn মধ্যে O পড়ে K গেছে K I_Sym
</Sentence>
<Sentence id=11>
এ O রকম B এক Bn অনিশ্চিত Bn অবস্থার B মধ্যেই O নতুন Bn বছরে B প্রবেশ K করছে K বাংলাদেশ B I_Sym
</Sentence>

```

Figure 4.3.1: Sample Tagged Data

4.3.2 Tagset

We took 8 specific tags to our Tagset those are:

- B (Noun/বিশেষ্য)
- S (Pronoun/সিনব াম)
- Bn (Adjective/বিশেষণ)
- K (Verb/ক্রিয়া)
- O (Adverb/অিয়ই)
- N (Number/সংখ্যা)
- E (English Letter)
- Sym (Symbol)


```

_B
_Bn
_S
_O
_K
_N
_E
_Sym

```

Figure 4.3.1: Sample Tagset

4.3.3 Untagged Dataset

If we want to run the program obviously we need some untagged data which will be tagged with the help of tagged corpus/ online direct corpus. We took some data for the experimental purpose.

Those data collected from a newspapers web portal.

These Data are In Bengali word.

```

word_to_be_tagged = u"কুমিল্লার লাকসামের কবুতর বাজার এলাকায় ট্রাক্টরের চাপায় হৃদয় মিয়া (৪০) নামের এক ব্যক্তি মারা গেছেন। তিনি ট্রাকচালকের স
tokenized = nltk.word_tokenize(word_to_be_tagged)

print(pos_tagger.tag(tokenized))

```

Figure 4.3.3: Sample Untagged Input

4.3.4 New Tagged Data

We send the untagged data and this data compared with the corpus tagged dataset, then this new tagged data shows to us.

```
[('কুমিল্লার', 'B'), ('লাকসামের', 'Unk'), ('কবুতর', 'Unk'), ('বজোর', 'B'), ('এলাকায়', 'Unk'), ('ট্রাক্টরের', 'Unk'), ('চাপায়', 'Unk'), ('হৃদয়', 'Unk'), ('মিয়া', 'Unk'), ('', None), ('৪০', 'N'), ('', None), ('নামের', 'B'), ('এক', 'BN'), ('ব্যক্তি', 'S'), ('মার', 'BN'), ('গেছেন', 'Unk'), ('তিনি', 'S'), ('ট্রাকচালকের', None), ('সহকারী', 'BN'), ('ছিলেন', 'Unk'), ('গতকাল', 'S'), ('মঙ্গলবার', 'B'), ('সকাল', 'B'), ('ছয়টার', 'Unk'), ('দিকে', 'S'), ('এ', 'S'), ('দুঘটিনা', 'Unk'), ('ঘটো', 'Unk'), ('একই', 'BN'), ('দিন', 'B'), ('দাউদকান্দি', 'Unk'), ('উপজেলায়', 'Unk'), ('সড়ক', 'Unk'), ('দুঘটিনায়', 'Unk'), ('নুরুল', None), ('ইসলাম', None), ('', None), ('২৫', 'N'), ('', None), ('নামের', 'B'), ('এক', 'BN'), ('ট্রাকচালক', 'B'), ('নিহত', 'BN'), ('হনা', 'Unk'), ('পুলিশ', 'B'), ('জানায়', 'Unk'), ('', 'SYM'), ('লাকসামে', 'Unk'), ('নিহত', 'BN'), ('হৃদয়ের', 'Unk'), ('বাড়ি', 'Unk'), ('নারায়ণ গঞ্জের', 'Unk'), ('নিতাইগঞ্জ', 'Unk'), ('এলাকায়', 'Unk'), ('লাকসাম', 'Unk'), ('পৌর', 'B'), ('এলাকার', 'B'), ('কবুতর', 'Unk'), ('বাজারে', 'Unk'), ('একটি', 'BN'), ('ট্রাক', 'B'), ('থেকে', 'O'), ('পপ্য', 'Unk'), ('নামাচ্ছিলেন', 'Unk'), ('হৃদয়', 'Unk'), ('এ', 'S'), ('সময়', 'B'), ('ওই', 'S'), ('পথ', 'B'), ('দিয়ে', 'Unk'), ('একটি', 'BN'), ('ট্রাক্টর', 'Unk'), ('যাচ্ছিল', 'Unk'), ('ওই', 'S'), ('ট্রাক্টরের', 'Unk'), ('চাপায়', 'Unk'), ('তিনি', 'S'), ('নিহত', 'BN'), ('হনা', 'Unk'), ('দাউদকান্দি', 'Unk'), ('হাইওয়ে', 'Unk'), ('পুলিশ', 'B'), ('ও', 'O'), ('প্রত্যক্ষদর্শীরা', 'B'), ('জানান', 'K'), ('', 'SYM'), ('গতকাল', 'S'), ('সকাল', 'B'), ('ছয়টার', 'Unk'), ('দিকে', 'S'), ('ঢাকা-চট্টগ্রাম', 'Unk'), ('মহাসড়কে', 'Unk'), ('দাউদকান্দি', 'Unk'), ('উপজেলার', 'B'), ('রায়পুর', 'Unk'), ('সেতুর', 'B'), ('কাছে', 'S'), ('চট্টগ্রামগামী', 'Unk'), ('একটি', 'BN'), ('ট্রাক', 'B'), ('নিয়ন্ত্রণ', 'Unk'), ('হারিয়ে', 'Unk'), ('রাস্তার', 'B'), ('পাশে', 'BN'), ('গাছের', 'Unk'), ('সঙ্গে', 'O'), ('ধাক্কা', 'B'), ('খায়', 'Unk'), ('এতে', 'S'), ('ট্রাকের', 'Unk'), ('চালক', 'Unk'), ('নুরুল', None), ('ইসলাম', 'B'), ('ঘটনাস্থলেই', 'B'), ('নিহত', 'BN'), ('হনা', 'Unk'), ('নিহত', 'BN'), ('ট্রাকচালকের', None), ('বাড়ি', 'Unk'), ('চট্টগ্রামের', 'B'), ('গুপ্তপুরের', 'Unk'), ('হাসনাবাদ', 'Unk'), ('এলাকায়', 'Unk')]
```

Figure 4.3.4: New Tagged Output

4.3.5 Explanation of program

At First we took Indian Corpus as like model

```
import nltk
from nltk.corpus import indian
from nltk.tag import tnt
import string
```

Figure 4.3.5: Indian Corpus

Secondly we read the Corpus what we need and we trained in our Program for further process

```
tagged_set = 'C:\\Users\\HunTer\\Desktop\\TNT tagger pos tagging\\bangla.pos'
word_set = indian.sents(tagged_set)
count = 0
```

Figure 4.3.5: Reading Bangla Dataset

```

for sen in word_set:
    count = count + 1
    sen = "".join([" " + i if not i.startswith("'") and i not in string.punctuation else i for i in sen]).strip()
    print (sen)
print (count)

train_perc = .9

train_rows = int(train_perc*count)
test_rows = train_rows + 1

print (train_rows, test_rows)

```

Figure 4.3.5: Count Dataset

At the Third step we read the untagged Dataset what we have to tagged by the use of corpus.

```

word_to_be_tagged = u"কুমিল্লার লাকসামের কবুতর বাজার এলাকায় ট্রাক্টরের চাপায় হৃদয় মিয়া (৪০) নামের এক ব্যক্তি মারা গেছেন। তিনি ট্রাকচালকের স
tokenized = nltk.word_tokenize(word_to_be_tagged)

print(pos_tagger.tag(tokenized))

```

Figure 4.3.5: Untagged Data

In the Fourth Section we get some new tagged data by tokenize which were untagged

```

[(('কুমিল্লার', 'B'), ('লাকসামের', 'Unk'), ('কবুতর', 'Unk'), ('বাজার', 'B'), ('এলাকায়', 'Unk'), ('ট্রাক্টরের', 'Unk'), ('চাপায়', 'Unk'), ('হৃদয়', 'Unk'), ('মিয়া', 'Unk'), ('(', None), ('৪০', 'N'), (',', None), ('নামের', 'B'), ('এক', 'BN'), ('ব্যক্তি', 'S'), ('মারা', 'BN'), ('গেছেন', 'Unk'), ('তিনি', 'S'), ('ট্রাকচালকের', None), ('সহকারী', 'BN'), ('ছিলেন', 'Unk'), ('গতকাল', 'S'), ('মঙ্গলবার', 'B'), ('সকাল', 'B'), ('১০টার', 'Unk'), ('দিকে', 'S'), ('এ', 'S'), ('দুঘটনা', 'Unk'), ('ঘটো', 'Unk'), ('একই', 'BN'), ('দিন', 'B'), ('দাউদকান্দি', 'Unk'), ('উপজেলায়', 'Unk'), ('সড়ক', 'Unk'), ('দুঘটনায়', 'Unk'), ('নুরুল', None), ('ইসলাম', None), ('(', None), ('২৫', 'N'), (',', None), ('নামের', 'B'), ('এক', 'BN'), ('ট্রাকচালক', 'B'), ('নিহত', 'BN'), ('হন', 'Unk'), ('পুলিশ', 'B'), ('জানায়', 'Unk'), (',', 'SYM'), ('লাকসামে', 'Unk'), ('নিহত', 'BN'), ('হৃদয়ের', 'Unk'), ('বাড়ি', 'Unk'), ('নারায়ণ গঞ্জের', 'Unk'), ('নিতাইগঞ্জ', 'Unk'), ('এলাকায়', 'Unk'), ('লাকসাম', 'Unk'), ('পৌর', 'B'), ('এলাকার', 'B'), ('কবুতর', 'Unk'), ('বাজারে', 'Unk'), ('একটি', 'BN'), ('ট্রাক', 'B'), ('থেকে', 'O'), ('পশ্য', 'Unk'), ('নামাছিলেন', 'Unk'), ('হৃদয়', 'Unk'), ('এ', 'S'), ('সময়', 'B'), ('গুই', 'S'), ('পথ', 'B'), ('দিয়ে', 'Unk'), ('একটি', 'BN'), ('ট্রাক্টর', 'Unk'), ('যাচ্ছিল', 'Unk'), ('গুই', 'S'), ('ট্রাক্টরের', 'Unk'), ('চাপায়', 'Unk'), ('তিনি', 'S'), ('নিহত', 'BN'), ('হন', 'Unk'), ('দাউদকান্দি', 'Unk'), ('হাইওয়ে', 'Unk'), ('পুলিশ', 'B'), ('গু', 'O'), ('প্রত্যক্ষদর্শীরা', 'B'), ('জানান', 'K'), (',', 'SYM'), ('গতকাল', 'S'), ('সকাল', 'B'), ('ছয়টার', 'Unk'), ('দিকে', 'S'), ('ঢাকা-চট্টগ্রাম', 'Unk'), ('মহাসড়কে', 'Unk'), ('দাউদকান্দি', 'Unk'), ('উপজেলার', 'B'), ('রায়পুর', 'Unk'), ('সেতুর', 'B'), ('কাছে', 'S'), ('চট্টগ্রামগামী', 'Unk'), ('একটি', 'BN'), ('ট্রাক', 'B'), ('নিয়ন্ত্রণ', 'Unk'), ('হারিয়ে', 'Unk'), ('রাস্তার', 'B'), ('পাশে', 'BN'), ('গাছের', 'Unk'), ('সঙ্গে', 'O'), ('ধাক্কা', 'B'), ('খায়', 'Unk'), ('এতে', 'S'), ('ট্রাকের', 'Unk'), ('চালক', 'Unk'), ('নুরুল', None), ('ইসলাম', 'B'), ('ঘটনাস্থলেই', 'B'), ('নিহত', 'BN'), ('হন', 'Unk'), ('নিহত', 'BN'), ('ট্রাকচালকের', None), ('বাড়ি', 'Unk'), ('চট্টগ্রামের', 'B'), ('গুতপুরের', 'Unk'), ('হাসনাবাদ', 'Unk'), ('এলাকায়', 'Unk')])

```

Figure 4.3.5: New Tagged Outcome

4.4 Summary

We implement the algorithm with a little untagged data with a little containing data Corpus. At next moment we will train this algorithm with the full size corpus.

When the corpus containing data will be more then the algorithm will also gives us more accuracy or output and more tags will be used for Bengali language, as the result this algorithm gives us more to learn about every words Part of Speech.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the study

In this whole process of our work we learned such things we don't know about it much, we had less knowledge and idea about it before. We had a little idea about POS tagging system and how does it works. We only aware that it like a process of identify a word's part of speech. But the field of tagging POS word is much more than our thought.

In this research we worked with quite a lot of software that we were not aware much about all of these before.

We actually started working with this POS tagging topic knowing close to nothing. We just determined that we have to do this research properly at any cost. So we started working and discover and explore so many new things we need. When we found out we need something then we started to look for it and tried to manage it. So, here we in our research report we tried to explain about what we are doing in our research about the topic.

We described our working procedure, objectives, methodology, problems that we faced and implementation of the code about the research project. We have analyzed this to get better accuracy of the POS tag sets.

5.2 Conclusion

In this research work, we have tested with necessary required rules for our research project. There are many worked on POS tagging around the world but not that much work was done on Bangla POS tagging. People are less aware about these types of things. We have to make sure that what we are trying to showing that people are understanding these things in a simple way.

There are many algorithms of NLP to identify the word's parts of speech. People uses so many algorithms to solve it. Some algorithms give much better accuracies than other algorithms. It's not all about which one gives better accuracy but also tests algorithms to check the accuracy of given data sets. We are trying our best to implement algorithms and find out the word's parts of speech. This research project is a small effort from us.

5.3 Implication for Further Study References

There can be much to do in the future on Bangla POS tagging. We want to apply some more algorithms and methods in the future to give this research project a new dimension. Neural Networking is a vast field to work on. There is so much more chances and opportunities to show some creativities. This research project is for all the Bengali people who are seeking inspiration and motivation to work with. This study will help them to work and study about this topic in the near future. It will make their research much easier.

As this is our first research about any topic, so we realize that there are some lacking in it. We have some planning and thinking with this topic in the future.

This research kind of an experiment with words to identify the word's POS for us. We will try to make this research much more advanced and dynamic.

We have got some ideas and thoughts about our research project. Because of shortage of time, we could not do it now but hopefully we can implement those ideas and techniques in the near future. Also, shortage of funding is an issue. So, everything is not done yet. We save that for the coming days and we hope that we will continue to work on this research project with adding something new that not done before.

Appendix

Appendix A: Related issues

NLP (Natural Language Processing): Artificial intelligence and the field of computer science are mainly concerned with interactions between computers and human languages. NLP works like a sub platform. In these interactions between computers and human languages plays an important role. NLP is generally used in to process and analyze the data then program the computers data.

HMM (Hidden Markov Model): The model is in hidden state with Markov process. Hidden Markov Model is a type of Markov model which is statistical. It can be represented as dynamic Bayesian network. L.E. Baum worked on this model's mathematics with his coworkers. In the hidden Markov method, the state is invisible normally in kind of hidden forms. If the parameters of this model are known exactly but still it showed to as a hidden Markov model.

References

[1] “Part-of-Speech Tagging”, available online: https://en.wikipedia.org/wiki/Part-of-speech_tagging [Last accessed 4th Nov 2019]

[2] Garg, Navneet, Vishal Goyal, and Suman Preet. "Rule based Hindi part of speech tagger." *Proceedings of COLING 2012: Demonstration Papers*. 2012.

[3] “Trigrams'n'Tags”, available online: <http://www.coli.uni-saarland.de/~thorsten/tnt/>

[Last accessed 4th Nov 2019]

[4] “Language”, available online: <https://en.wikipedia.org/wiki/Language>

[Last accessed 4th Nov 2019]

[5] Yajnik, Archit. "ANN Based POS Tagging For Nepali Text."

[6] Joshi, N., Darbari, H., & Mathur, I. (2013). HMM based POS tagger for Hindi. In *Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013)*.

[7] De Smedt, Tom, Fabio Marfia, Matteo Matteucci, and Walter Daelemans. "Using wiktionary to build an italian part-of-speech tagger." In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pp. 1-8. Springer, Cham, 2014.

[8] Naseem, A., Anwar, M., Ahmed, S., Satti, Q.A., Hashmi, F.R. and Malik, T., 2017.

Tagging Urdu Sentences from English POS Taggers. *Corpus*, 96(1), p.2.

[9] Alex M, Zakaria LQ. Brill's rule-based part of speech tagger for kadazan. *International Journal on Recent Trends in Engineering & Technology*. 2014 Jan 1;10(1):75.

[10] Biswas, Priyanka, et al. "A Corpus-based Study of কেরা (kare) in Bangla:

Theoretical and Computational Perspectives."

©Daffodil International University

- [11] Alashqar, A. M. (2012, May). A comparative study on Arabic POS tagging using Quran corpus. In *2012 8th International Conference on Informatics and Systems (INFOS)* (pp. NLP-29). IEEE.
- [12] Oroumchian, F., Tasharofi, S., Amiri, H., Hojjat, H., & Raja, F. (2006). Creating a feasible corpus for Persian POS tagging. *Department of Electrical and Computer Engineering, University of Tehran*.
- [13] Singh, Jyoti, Nisheeth Joshi, and Iti Mathur. "Part of speech tagging of Marathi text using trigram method." *arXiv preprint arXiv:1307.4299* (2013).
- [14] Ali, H., 2010. An unsupervised parts-of-speech tagger for the bangla language. *Department of Computer Science, University of British Columbia, 20*, pp.1-8.
- [15] Garcia M, Gamallo P, Gayo I, Cruz MA. PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*. 2014(53):95-101.
- [16] Khan, N., Habib, M. T., Alam, M. J., Rahman, R., UzZaman, N., & Khan, M. (2006). History (forward n-gram) or Future (backward n-gram)? Which model to consider for n-gram analysis in Bangla?.