Bachelor of Science in Computer Science & Engineering

Developing a Supervised System using Lexicon Database for Bangla Word Sense Disambiguation

by

Tareq Ahamed Shuvo

ID: 1504089

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

Developing a Supervised System using Lexicon Database for Bangla Word Sense Disambiguation

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Tareq Ahamed Shuvo
ID: 1504089

Supervised by

Lamia Alam
Assistant Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled '**Developing a Supervised System using Lexicon Database for Bangla Word Sense Disambiguation'** submitted by ID: 1504089, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

____ Chairman

Lamia Alam
Assistant Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

____ Member (Ex-Officio)

Dr. Asaduzzaman
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

____ Member (External)

Dr. Mahfuzulhoq Chowdhury
Assistant Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date: 18.04.2021

Acknowledgements

The long journey of this project comes to an end, not because of just me. From my teachers to my batch-mates everyone has a contribution to this project. First of all, I would like to express my heartfelt gratitude towards my supervisor Lamia Alam (Assistant Professor, Department of Computer Science & Engineering, CUET). I am grateful to her for her continuous guidance and simultaneous response to help. Besides, I would like to thank my batch-mates for helping me mentally and technically. And last but not least, my parents whom unconditional love and support help me to make this possible.

Abstract

A Supervised System using Lexicon Database for Bangla Word Sense Disambiguation is the process of figuring out the actual meaning of an ambiguous Bangla word based on the context. Ambiguous word implies a word having multiple meanings. A Bangla POS tagging system with Bangla Lexicon database of 1,17,042 words has been integrated with this system. Besides, a contribution of 64,760,021 (about 65 million) sentences and 854,589,384 (about 1 billion) words have been made to Bangla Corpus Database. The evaluation of result reveals that Bangla Disambiguation, Ambiguous Word Detection and Bangla POS Tagging system have 84.6%, 88.2% and 92.7% accuracy respectively.

Keywords: Bangla Word Sense Disambiguation, Bangla POS Tagging, Ambiguous Detection, Bangla Lexicon, Bangla Corpus.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Bangla Word Sense Disambiguation (BWSD) means recognizing the actual meaning or sense of an ambiguous Bangla word in a particular context. And words that have different meanings in different contexts are called ambiguous words. Lexical ambiguity is the presence of two or more possible meanings for a single word. It's also called semantic ambiguity or homonymy. It differs from syntactic ambiguity, which is the presence of two or more possible meanings within a sentence or sequence of words.

WSD-system is very important in the Language Technology fields. For instance, sometimes search results on the web are not appropriate because of ambiguity in the words. If it is possible to figure out the actual sense or meaning of those search words then the most appropriate results can be found. This is one example. There are tons of other uses and fields, likely, machine translating, extracting summary, categorizing contents, processing natural languages and so on.

Because of its importance, many researchers already have worked on Word Sense Disambiguation System. But unfortunately, most of them are done in English language and for Bangla language, the number is very little whereas Bangla is the 4th largest language in the world.

As human beings are blessed with inherent linguistic competence, they can easily find out the contextual sense of an ambiguous word, but it is a difficult task for a computer to do this. Because a computer thinks differently, works differently. However, at this current time, many researchers are trying to do this in Bangla and in different approaches. Both supervised and unsupervised methods are being applied to solve ambiguities.

1.2 Challenges

Some challenges have been raised while working with Bangla language processing. Among these -

- Community support in Bangla is comparatively lesser than any other languages like English, Hindi, Spanish, etc.
- Insufficient words in Bangla Dictionary (Electronic format)
- Insufficient Bangla Corpus or Lexicon Database.
- Unavailability of Bangla Natural Language Processing Toolkits.
- Unavailability of Bangla POS tagging system.

1.3 Applications

Bangla Word Sense Disambiguation has various usages and applications. Some of them are to get better -

- search results with ambiguous phrases or clauses
- summarization of Bangla essay, paragraphs, etc. writings.
- translation from Bangla language to other languages and vice-versa.

1.4 Motivation

When we search for something on search engines, we usually get better results in English than in Bangla, especially with ambiguous phrases. Not only that, translation of foreign language to Bangla is horrible whereas other languages like Spanish, French, Hindi, etc. give us much better results.

All of these problems can be solved with the help of the Bangla Disambiguous System. Although there are few Bangla Disambiguous System which has been developed in recent year, however, this is not mature enough. With the vision of better performance in terms of Bangla corpus or lexicon database size, we have decided to **develop a Supervised System using Lexicon Database for Bangla Word Sense Disambiguation**.

1.5 Contributions

The primary focus of our project is to develop a Supervised System using Lexicon Database for Bangla Word Sense Disambiguation. All the contributions are as follows -

- Addition of **1,17,042 words** to a **Bangla Dictionary or Lexicon Database**
- Enhancement of a Bangla Corpus Database which contains 64,760,021 (**about 65 million**) sentences and 854,589,384 (**about 1 billion**) words.
- Development of a **Bangla Natural Language Toolkits**.
- Improvement of a **POS tagging system** for Bangla Language.
- Development of a **system for Bangla Word Sense Disambiguation**.
- Evaluation of the system performance.

1.6 Thesis Organization

The rest of this report is organized as follows:

- **Chapter 2** provides a brief summary of different methods for disambiguation and previous research work on Bangla Word Sense Disambiguation.
- **Chapter 3** depicts the procedures of the proposed methodology.
- **Chapter 4** provides the result generated by our developed system in detail with an explanation.
- **Chapter 5** gives the overall summary of this project and provides some future recommendations as well.

1.7 Conclusion

In this introduction chapter, we introduced our system what are we going do to in the next chapters. We tried to explain, what is Bangla Word Sense Disambiguation System, what is it used for, what makes us motivated to develop this, the difficulties to develop it and so on.

Chapter 2

Literature Review

2.1 Introduction

There are different approaches to implement Word Sense Disambiguation. In this chapter, we will shortly describe some of the disambiguation approaches. At the end of this section, it contains some of the related work of the researchers.

2.1.1 Word N-grams

To understand this concept, we had better start with few examples where it is used. Spell checking, grammar checking, auto completion, sentence completion, etc. are the use cases of word N-grams. A contiguous sequence of N words from a given sample of the text is called an N-gram. For instances,

Text	**N-gram**
Hello	1-gram
Hey there	2-gram
How are you	3-gram
Nice to met you	4-gram

Table 2.1: Word N-grams

If a system is N-gram system, then that system works with N contiguous sequence of words. For example, a 4-gram sentence completion system will suggest or complete four words for a given input.

2.1.2 Different Methods of Word Sense Disambiguation

In recent years, many works have been done on different approaches to solving the disambiguation problem. It has been done using supervised and unsupervised learning approaches and some of the researchers tried to solve this by combining both. Some of the methods has been briefly stated below -

2.1.2.1 Dictionary-based approach

Dictionary-based approach or also known as **Lexicon-based approach** which is one kind of knowledge-based approach that relies on the dictionary or lexicon. The results obtained using this method are quite impressive. This method gives better result if the used dictionry or lexicon is larger in size. We have used this method in our project because of its better performance and being able to manage a larger Bangla lexicon and Corpus datasets. A detailed discussion will be in the next chapters. An overview of this method is depicted in the figure: 3.1

2.1.2.2 Naïve Bayes classifier

The Bayes Theorem is used to build a set of classification algorithms known as Naive Bayes classifiers. It is a family of algorithms that share a common concept, namely that each pair of features being classified is independent of the others.

$$P(class/features) = \frac{P(class) * P(features/class)}{P(features)} \tag{2.1}$$

- P(class/features) : Posterior Probability
- P(class) : Class Prior Probability
- P(features/class) : Likelihood
- P(features) : Predictor Prior Probability

2.1.2.3 K-Nearest Neighbors Algorithm (K-NN)

In 1951, it was first developed by Evelyn Fix and Joseph Hodges and later expanded by Thomas Cover which is still used for classification in Machine Learning. Word Sense Disambiguation can be done using this method as well.

Regression can be done using this method but in the case of classification (figure: 2.1), at first, K-NN calculates the distances between a query and all of the samples in the data, selecting the K closest samples to the query and then vote for the most frequent label.

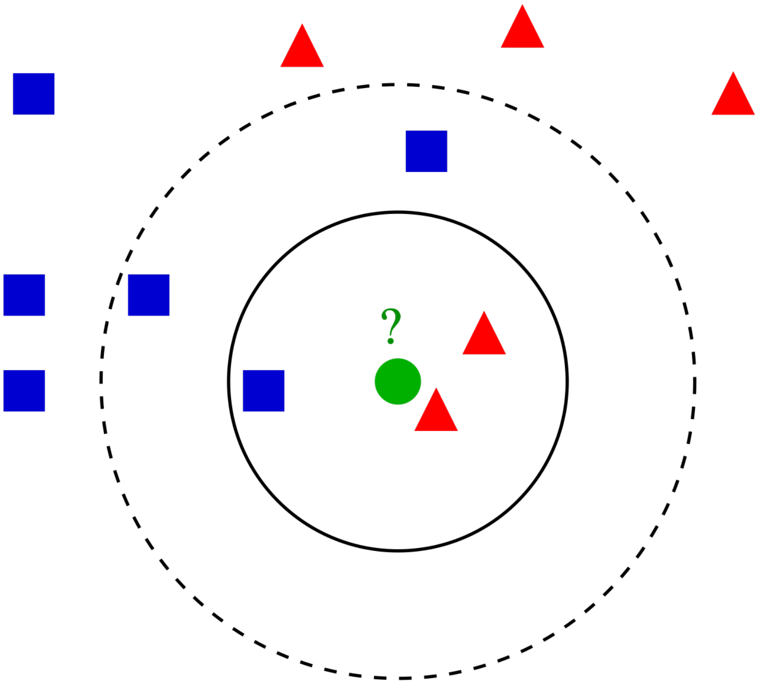


Figure 2.1: Classification using K-NN

2.1.2.4 Conceptual Density (CD) method

In 1996, it was first introduced by Agirre and Rigau (figure: 2.2) as a measure of the correlation between the sense of a given word and its context which can be used to disambiguate a word based on context. It is based on the use of WordNet's wide-coverage noun taxonomy and the notion of conceptual distance among concepts.

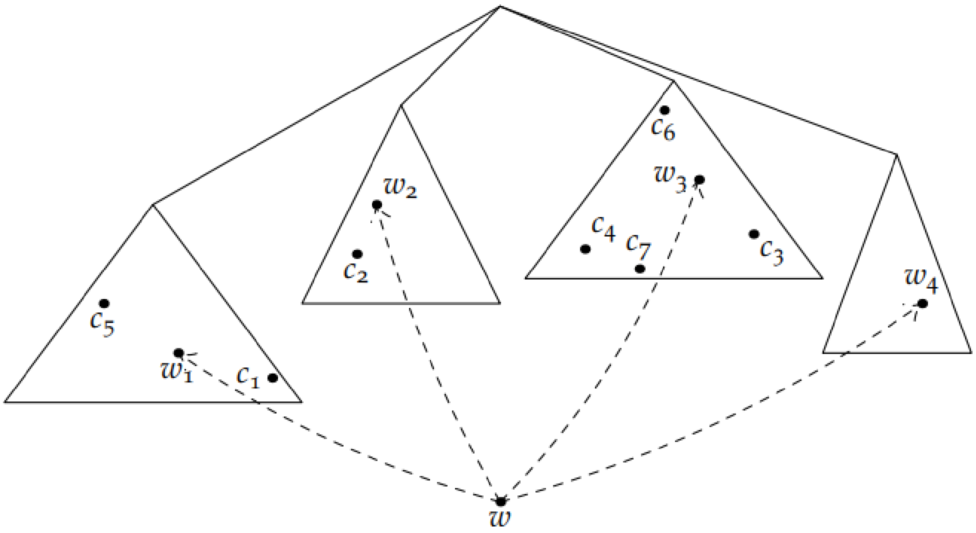


Figure 2.2: Illustration of the conceptual density measure by Agirre and Rigau (1996)

2.2 Related Literature Review

Works on Natural Language Processing in Bangla are really limited to their demands. In recent years, it has been found that some researchers work on it but which is not really satisfactory. But among those, some works are really praiseworthy in terms of accuracy and innovation. Recently, a review on Word Sense Disambiguation has been published by Chingakham PonyKumar Singh, et. al. [1] where they emphasized the data resources on wordnet and corpus. In the same year, B.H. Manjunatha Kumar and M Siddappa [2] have proposed to use the semantic relations of words to solve Word Sense Disambiguation (WSD) for the Kannada language. They tried to make the overlaps between two concepts of semantic relationships like hypernymy and hyponymy. A knowledge-based dictionary approach [3] is used by A. Haque and M. M. Hoque to detect ambiguous words and corresponding meanings. It provides 82.40% accuracy. Another knowledge-based approach [4] using Bengali WordNet was done by Pal et al. They actually tried to figure out the maximum number of overlap among ambiguous words' definitions along with collocating words in the sentence as well as synonymous words of these collocating words. They achieved 75% accuracy testing on 9 common Bengali ambiguous words. They classified Bangla Sentences [5] using WordNet. N. N. Islam, N. Nawaz, M. Tanzil and M. M. Ali have done a theoretic Bangla Word Sense Disambiguation technique [6] using Lexeme List. A k-Nearest Neighbor (k- NN) algorithm was proposed [7] and achieved an accuracy of over 71%. Naïve Bayesian Algorithm [8] was implemented for Word Sense Disambiguation. Yarowsky algorithm [9] was implemented as a solution by D. McCarthy named Word Sense Disambiguation from the University Of Melbourne in 2011 and analyzed its behavior with respect to program parameters. Using the Statistical Approach [10] 82% accuracy was found. A hybrid approach [11] combining supervised and unsupervised learning was done by Pal et al. on the English language. But the outcome was not so satisfactory. To enrich Bangla and make it more compatible with a computer, a lot of researches and projects should be done. Especially the accuracy and the size of the working data are not up to the mark. The size of the data is really a big factor because bigger data gives more accurate results. Unavailability of sufficient words in the Bangla dictionary (Lexicon) and contents in computer format is the main obstacle to progress. So the focus of our project will be to increase data size both training testing data as well as the number of words in the Bangla dictionary to achieve higher accuracy.

2.3 Conclusion

In this chapter, we have discussed the previous works on Word Sense Disambiguation especially in Bangla. Definitely, different approaches showed different results and have different limitations. All of those limitations cannot be solved at a time but in the next chapter, we will discuss a methodology and its procedures to overcome some of the limitations.

Chapter 3

Methodology

3.1 Introduction

In this chapter, we will discuss how we have developed a **Supervised System using Lexicon Database for Bangla Word Sense Disambiguation** and how it works. We will demonstrate its procedures step by step and depict its overview.

3.2 Diagram/Overview of Framework

The key objective of our work is to develop a system that can identify ambiguous words from Bangla sentences and figure out the actual meaning of that ambiguous word in that context. An overview of such a system has been depicted in the figure: 3.1.

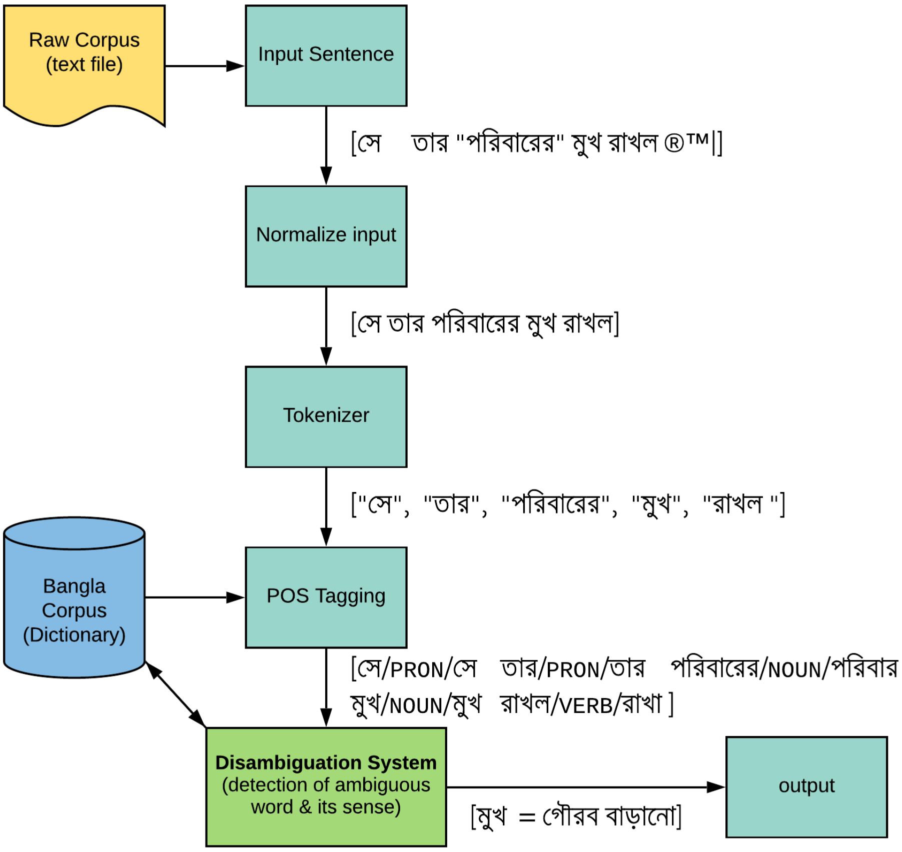

Raw Corpus (text file)
Input Sentence
[সে   তার "পরিবারের" মুখ রাখল ®™|]
Normalize input
[সে তার পরিবারের মুখ রাখল]
Tokenizer
["সে", "তার", "পরিবারের", "মুখ", "রাখল "]
Bangla Corpus (Dictionary)
POS Tagging
[সে/PRON/সে   তার/PRON/তার   পরিবারের/NOUN/পরিবার
মুখ/NOUN/মুখ   রাখল/VERB/রাখা]
Disambiguation System
(detection of ambiguous word & its sense)
[মুখ = গৌরব বাড়ানো]
output

Figure 3.1: Overview of the whole system (BWSD)

Details of each step of the Bangla Word Sense Disambiguation System (BWSD) have been discussed in the next sections.

3.3 Detailed Explanation

The overview of the whole system has been depicted in the figure: 3.1 and now the description of those steps has been explained in two sections. These are -

3.3.1 Data collection

Bigger Bangla Corpus and Bangla Lexicon databases must give better results in disambiguation. So, to enrich those databases, we had to do web scraping from the web. The resources we used to do web scraping -

- **Softwares**
  - Python Scrapy as a web scraper

    - NodeJS for HTTP/HTTPS request
    - Python for HTTP/HTTPS request
    - Javascript to handle DOM
    - jQuery to extract and sanitize data
    - Sqlite3 as a database
- **Hardware**
    - Laptop: Acer, Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.20GHz, 8 GB RAM, 64-bit Windows 10

But just using those tools and writing scripts were not enough for web scraping because -

- web is a collection of unordered data.
- every website has its own design and structure which makes us hard to use a single pattern.
- it is full of unusual & dangerous characters and fonts.
- it is full of malware and malicious Javascript codes.
- Google reCAPTCHA makes it hard to scrape and so on.

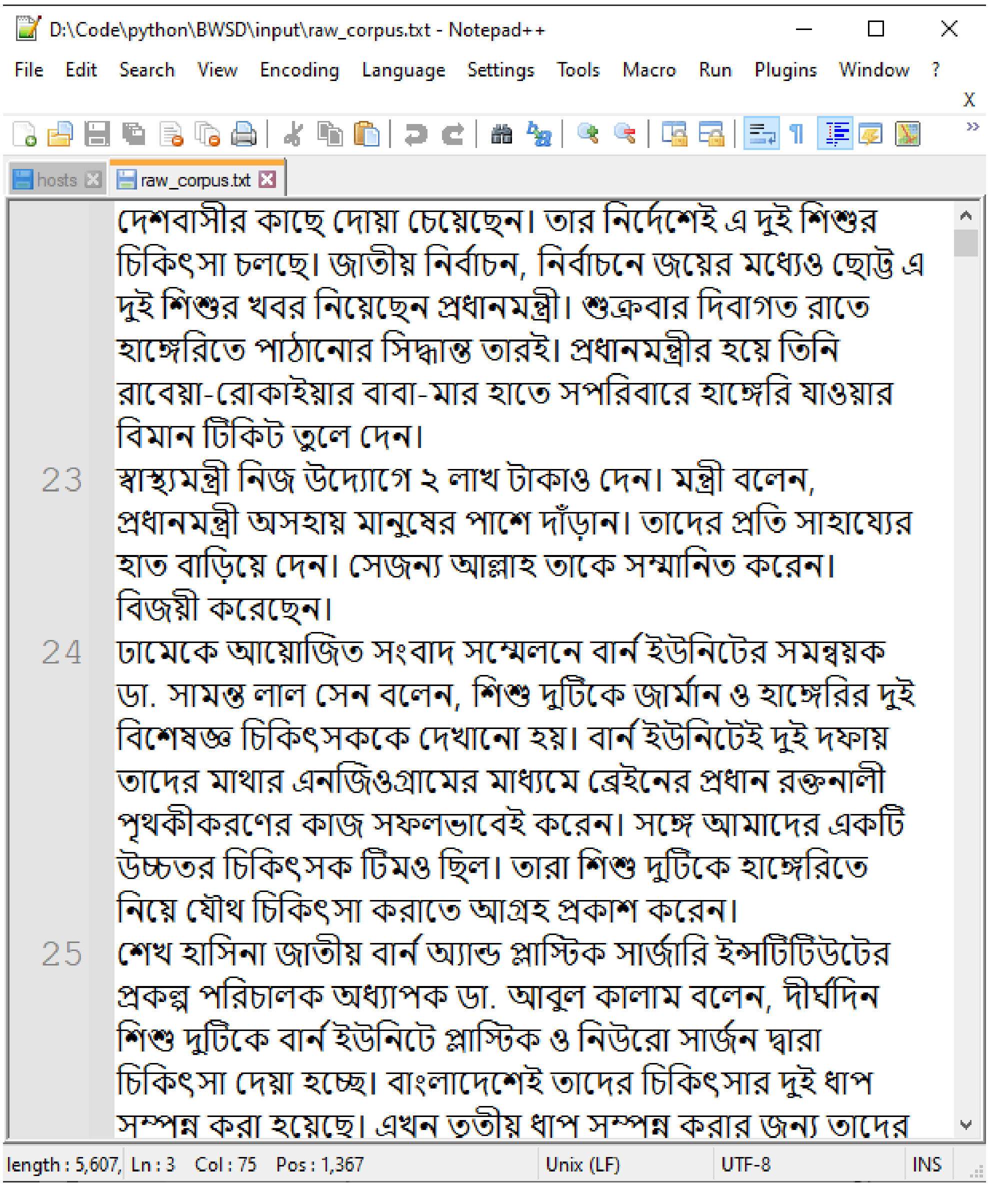
All of these problems caused our system to run into problems and sometimes even crash. So to overcome these, we had to follow a semi-automated process that is, scripting and manual human labour.

As a result of all those struggles, we have been able to collect data worth about 20GB which includes Bangla Lexicon of 1,17,042 words and Bangla Corpus of 64,760,021 (about 65 million) sentences and 854,589,384 (about 1 billion) words.

Our online sources from which we have collected our Bangla data -

- Anandabazar [12]
- Banglatribune [13]
- Bdnews24 [14]

- Banglanews24 [15]
- Banglapedia [16]
- Ittefaq [17]
- Jugantor [18]
- Kalerkantho [19]
- Prothomalo [20]
- Samakal [21]
- Somewhereinblog [22]



দেশবাসীর কাছে দোয়া চেয়েছেন। তার নির্দেশেই এ দুই শিশুর চিকিৎসা চলছে। জাতীয় নির্বাচন, নির্বাচনে জয়ের মধ্যেও ছোট্ট এ দুই শিশুর খবর নিয়েছেন প্রধানমন্ত্রী। শুক্রবার দিবাগত রাতে হাঙ্গেরিতে পাঠানোর সিদ্ধান্ত তারই। প্রধানমন্ত্রীর হয়ে তিনি রাবেয়া-রোকাইয়ার বাবা-মার হাতে সপরিবারে হাঙ্গেরি যাওয়ার বিমান টিকিট তুলে দেন।

স্বাস্থ্যমন্ত্রী নিজ উদ্যোগে ২ লাখ টাকাও দেন। মন্ত্রী বলেন, প্রধানমন্ত্রী অসহায় মানুষের পাশে দাঁড়ান। তাদের প্রতি সাহায্যের হাত বাড়িয়ে দেন। সেজন্য আল্লাহ তাকে সম্মানিত করেন। বিজয়ী করেছেন।

ঢামেকে আয়োজিত সংবাদ সম্মেলনে বার্ন ইউনিটের সমন্বয়ক ডা. সামন্ত লাল সেন বলেন, শিশু দুটিকে জার্মান ও হাঙ্গেরির দুই বিশেষজ্ঞ চিকিৎসককে দেখানো হয়। বার্ন ইউনিটেই দুই দফায় তাদের মাথার এনজিওগ্রামের মাধ্যমে ব্রেইনের প্রধান রক্তনালী পৃথকীকরণের কাজ সফলভাবেই করেন। সঙ্গে আমাদের একটি উচ্চতর চিকিৎসক টিমও ছিল। তারা শিশু দুটিকে হাঙ্গেরিতে নিয়ে যৌথ চিকিৎসা করাতে আগ্রহ প্রকাশ করেন।

শেখ হাসিনা জাতীয় বার্ন অ্যান্ড প্লাস্টিক সার্জারি ইন্সটিটিউটের প্রকল্প পরিচালক অধ্যাপক ডা. আবুল কালাম বলেন, দীর্ঘদিন শিশু দুটিকে বার্ন ইউনিটে প্লাস্টিক ও নিউরো সার্জন দ্বারা চিকিৎসা দেয়া হচ্ছে। বাংলাদেশেই তাদের চিকিৎসার দুই ধাপ সম্পন্ন করা হয়েছে। এখন তৃতীয় ধাপ সম্পন্ন করার জন্য তাদের

Figure 3.2: A screen-shot of sample raw Bangla Corpus

3.3.1.1 Training and Test Data

The most vital point as well as an issue in Machine Learning, especially Natural Language Processing, is to collect data. The bigger data (training data) we have, the precious outcomes will be. If we have huge collections of ambiguous sentences, then we can train our system more preciously. Our system will be aware of the more peculiar structure of sentence patterns.

We know, the internet is the biggest place to search and collect data. We will do Web scraping also known as web harvesting, or web data extraction - which means extracting data from the web. There are lots of Bangla news portals, blogs, forums, social media, Facebook, Twitter, Wikipedia and so on sources to collect Bangla data.

3.3.1.2 Dictionary (Bangla Corpus)

This is a dictionary-based approach so we need an enriched Bangla dictionary. Although we have an enriched collection of Bangla dictionaries from various publications all these are hardcover i.e. printed version. To work with our project we need a database-based dictionary which is really rare and limited. So here we have to enrich and increase our database-based dictionary and of course, it is in a manual way. More words will give us better POS tagging and results.

3.3.2 Steps of the BWSD System

3.3.2.1 Input Corpus/Sentence

The first step of our system to take input from the user which will be disambiguated later to the actual meaning of that ambiguous word in that context. Users may input anything from garbage to friendly input which we need to sanitize before further use. Otherwise, it may cause our system into trouble or even crash. So to do this, the next step of the system is Normalize step or also known as sanitize step.

3.3.2.2 Normalizing Input

Users can input anything, from garbage to proper string, to a system but to work with a system needs a systematic sequence of strings. To function our system properly, we have to ensure that input pattern. So our system will sanitize and normalize user input and make it perfect for our system to work with. As a normalizing process, we will remove unwanted, harmful, special characters likely, ({[]})=%$#@! *:;<>?/| etc. and non-Bangla words because it works with Bangla only.

For instance,
Input: **সে তার "পরিবারের" (মুখ) রাখল|**
Ouput: **সে তার পরিবারের মুখ রাখল**

3.3.2.3 Tokenize

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords called tokens. This is required for POS tagging and figuring out feature words, head words and the rest of the process.

For instance,
From: **সে তার পরিবারের মুখ রাখল**
To: **["সে", "তার", "পরিবারের", "মুখ", "রাখল"]**

3.3.2.4 POS Tagging

In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST) is the process of identification of words part-of-speech based on both its definition and its context. This is one of the big challenges because no such tools are available in Bangla, although there are huge collections of POS tagging tools available in English both online and offline. So we have to develop our own version of the Bangla POS tagging tool for our system. Our system has 7 different parts of speech, likely, Noun, Pronoun, Verb, Adjective, Adverb, Preposition and Conjunction. Notation of the tags has been given in figure: 3.1.

Tag name	**Tag Description**
NOUN	NOUN
PRON	PRONOUN
VERB	VERB
ADJ	ADJECTIVE
ADV	ADVERB
PREP	PREPOSITION
CONJ	CONJUNCTION

Table 3.1: POS Tagging tags

For instance,

From: **["সে", "তার", "পরিবারের", "মুখ", "রাখল"]**

To: সে/**PRON**/সে তার/**PRON**/তার পরিবারের/**NOUN**/পরিবার মুখ/**NOUN**/মুখ রাখল/**VERB**/রাখা

3.3.3 Implementation

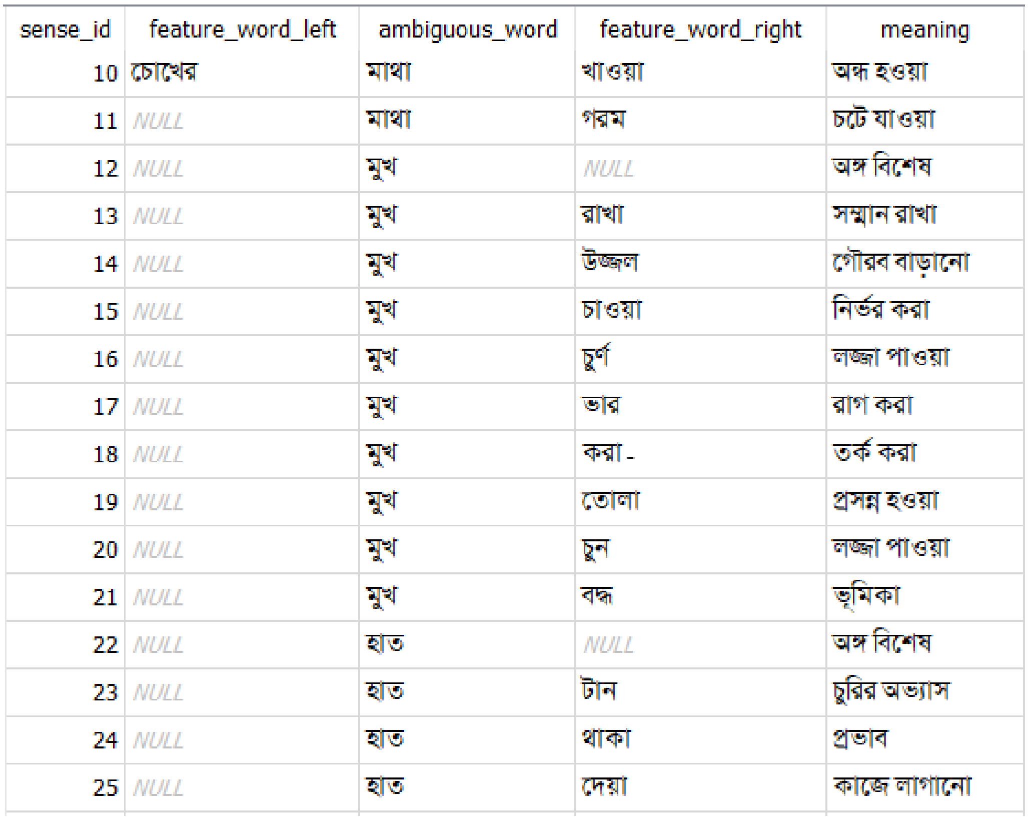
It is time to deep dive into the systems algorithm. How will this system detect ambiguous words from sentences? How will it figure out the actual meaning or sense of that ambiguous word based on both its definition and its context?

Our system is a 2-gram system and it has been developed using **"Supervised approach using Bangla Lexicon Database for Disambiguation"**

Since it is a supervised method, hence we will teach our system first. Firstly, our system will detect **ambiguous word** which is also known as **head word**. The words before and after the head word are regarded as **feature word**. Here are the steps which will be followed by Disambiguation System

- **Step 1:** POS tagging will be done on an input sentence before handing it over to the Disambiguation System. Now it has both POS tagged-sentence and access to Bangla Dictionary (Bangla text corpus).
- **Step 2:** To find the head word i.e. ambiguous word, the system will check every word in POS tagged sentence and their corresponding meaning from Bangla Dictionary. Words having more than one meaning will be treated as ambiguous word.
- **Step 3:** To find feature words, our system will use an ambiguous word. Word before and after head word (ambiguous word) will be counted as right feature word and left feature word respectively. If there is no word before

head word like head word is at the starting of the sentence, then the right feature word is null. The same goes for the left feature word when head word is at the last of the sentence.



sense_id	feature_word_left	ambiguous_word	feature_word_right	meaning
10	চোখের	মাথা	খাওয়া	অন্ধ হওয়া
11	NULL	মাথা	গরম	চটে যাওয়া
12	NULL	মুখ	NULL	অঙ্গ বিশেষ
13	NULL	মুখ	রাখা	সম্মান রাখা
14	NULL	মুখ	উজ্জল	গৌরব বাড়ানো
15	NULL	মুখ	চাওয়া	নির্ভর করা
16	NULL	মুখ	চুর্ণ	লজ্জা পাওয়া
17	NULL	মুখ	ভার	রাগ করা
18	NULL	মুখ	করা -	তর্ক করা
19	NULL	মুখ	তোলা	প্রসন্ন হওয়া
20	NULL	মুখ	চুন	লজ্জা পাওয়া
21	NULL	মুখ	বন্ধ	ভূমিকা
22	NULL	হাত	NULL	অঙ্গ বিশেষ
23	NULL	হাত	টান	চুরির অভ্যাস
24	NULL	হাত	থাকা	প্রভাব
25	NULL	হাত	দেয়া	কাজে লাগানো

Figure 3.3: Head words with left right feature words along with their senses in Bangla Corpus.

- **Step 4:** After detecting the head word (ambiguous word), right feature word and left feature word, now the Disambiguation System will find the actual meaning of that ambiguous word. Now our system will try to find matching with either **head word** + **right feature word** or **head word** + **left feature word** against our Bangla Text Corpus (Database) like figure: 3.3. This matching will give us the actual ambiguous meaning of that word. For instance, consider this sentence **"সে তার পরিবারের মুখ রাখল"**. Here "মুখ" is head word (ambiguous word) because it has more than one meaning enlisted in Bangla Dictionary. The word before "মুখ" is "পরিবারের" which is the right feature word and after word is which is the left feature word. Now no matching will be found for "পরিবার + মুখ" in our database but one matching will be found for "মুখ + রাখা" which mean . So here the actual meaning of the ambiguous word "মুখ" is "গৌরব বাড়ানো".

3.3.3.1 Resources

The resources both hardware and software have been used to implement this project are -

- **Hardware:**
  - Laptop: Acer, Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.20GHz, 8 GB RAM, 64-bit Windows 10
- **Software:**
  - Editor (IDE): PyCharm by JetBrains
  - Text Editor: Visual Studio Code
  - Anaconda Python Distribution
- **Programming Languages:**
  - Python as the main language
  - HTML, CSS, JS as web design
  - Flask as webserver

3.4 Conclusion

This chapter gives an overview of the whole system where we mentioned what methodology we used and how we have implemented this. The next chapter will give us the results generated by this system and the evaluation of performance.

Chapter 4

Results and Discussions

4.1 Introduction

In the previous chapter, our developed system's methodology and procedures have been explained elaborately. Here we will evaluate our system output results and discuss them in detail.

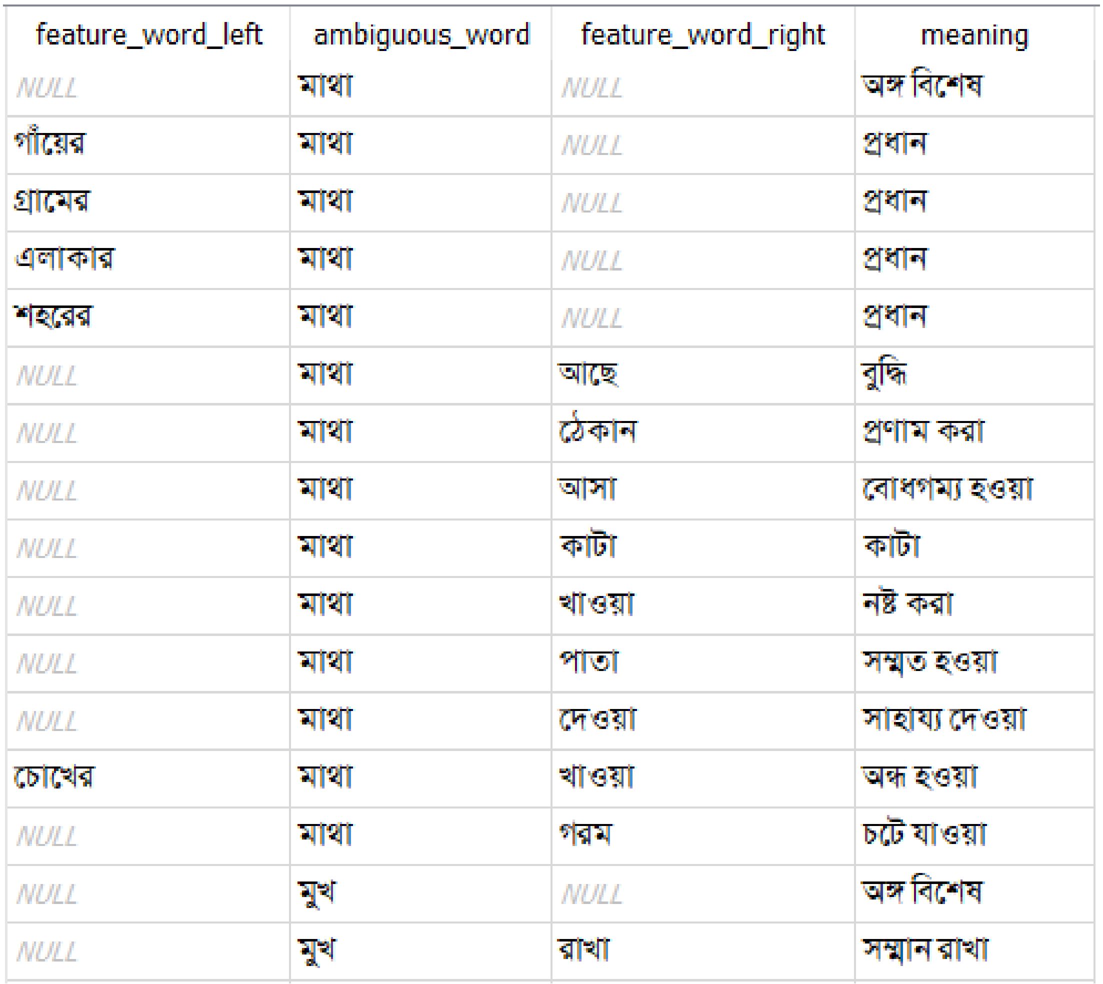
Runtime environment: Acer, Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.20 GHz, 8 GB RAM, 64-bit Windows 10 and Python language.
This is the runtime environment for our system where we run it. All results used here have been generated by the system using above mentioned environment.

4.2 Dataset Description

The dataset we used in this project can be divided into three categories. These are the Labeled dataset, Training dataset, and Bangla corpus.

4.2.1 Labeled Dataset

Labeled data is an essential part of supervised learning. In supervised learning, a machine is taught to identify or detect something based on features. The more labeled data we have, the more the machine will learn. A snapshot of the labeled dataset is given in figure: 4.1



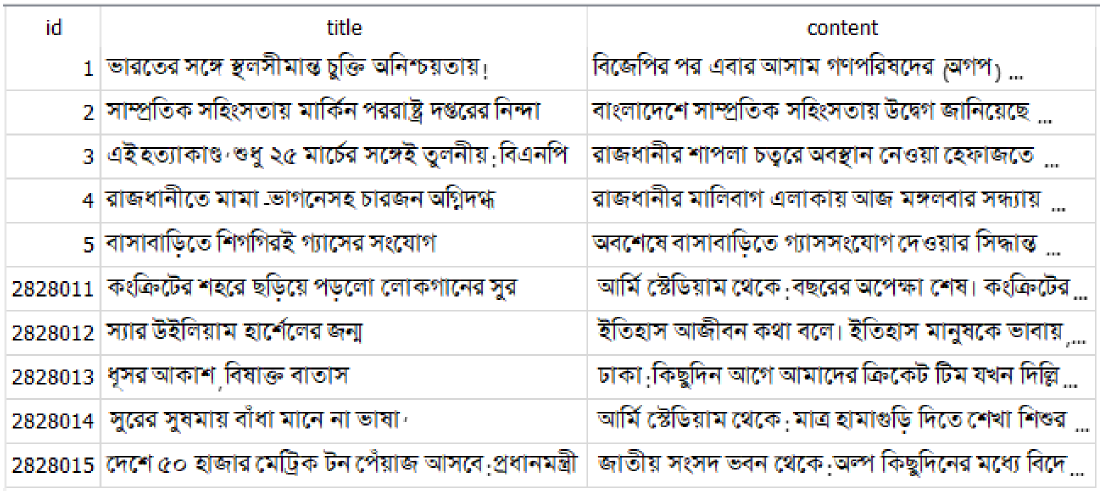
feature_word_left	ambiguous_word	feature_word_right	meaning
NULL	মাথা	NULL	অঙ্গ বিশেষ
গাঁয়ের	মাথা	NULL	প্রধান
গ্রামের	মাথা	NULL	প্রধান
এলাকার	মাথা	NULL	প্রধান
শহরের	মাথা	NULL	প্রধান
NULL	মাথা	আছে	বুদ্ধি
NULL	মাথা	ঠেকান	প্রণাম করা
NULL	মাথা	আসা	বোধগম্য হওয়া
NULL	মাথা	কাটা	কাটা
NULL	মাথা	খাওয়া	নষ্ট করা
NULL	মাথা	পাতা	সম্মত হওয়া
NULL	মাথা	দেওয়া	সাহায্য দেওয়া
চোখের	মাথা	খাওয়া	অন্ধ হওয়া
NULL	মাথা	গরম	চটে যাওয়া
NULL	মুখ	NULL	অঙ্গ বিশেষ
NULL	মুখ	রাখা	সম্মান রাখা

Figure 4.1: A snapshot of labeled dataset

4.2.2 Training dataset

To train a model a training dataset is needed. A bigger training dataset helps to get various patterns which consequently gives better accuracy.

Our training dataset has 64,760,021 (**about 65 million**) sentences and 854,589,384 (**about 1 billion**) words. And this dataset has 2828015 (**about 3 million**) records or rows with id, vendor, link, title, content, category, tags and status. Figure: 4.2 shows the first five and last five records of training dataset.



id	title	content
1	ভারতের সঙ্গে স্থলসীমান্ত চুক্তি অনিশ্চয়তায়!	বিজেপির পর এবার আসাম গণপরিষদের (অগপ) ...
2	সাম্প্রতিক সহিংসতায় মার্কিন পররাষ্ট্র দপ্তরের নিন্দা	বাংলাদেশে সাম্প্রতিক সহিংসতায় উদ্বেগ জানিয়েছে ...
3	এই হত্যাকাণ্ড শুধু ২৫ মার্চের সঙ্গেই তুলনীয়: বিএনপি	রাজধানীর শাপলা চত্বরে অবস্থান নেওয়া হেফাজতে ...
4	রাজধানীতে মামা-ভাগনেসহ চারজন অগ্নিদগ্ধ	রাজধানীর মালিবাগ এলাকায় আজ মঙ্গলবার সন্ধ্যায় ...
5	বাসাবাড়িতে শিগগিরই গ্যাসের সংযোগ	অবশেষে বাসাবাড়িতে গ্যাসসংযোগ দেওয়ার সিদ্ধান্ত ...
2828011	কংক্রিটের শহরে ছড়িয়ে পড়লো লোকগানের সুর	আর্মি স্টেডিয়াম থেকে: বছরের অপেক্ষা শেষ। কংক্রিটের ...
2828012	স্যার উইলিয়াম হার্শেলের জন্য	ইতিহাস আজীবন কথা বলে। ইতিহাস মানুষকে ভাবায়,...
2828013	ধূসর আকাশ, বিষাক্ত বাতাস	ঢাকা: কিছুদিন আগে আমাদের ক্রিকেট টিম যখন দিল্লি ...
2828014	সুরের সুষমায় বাঁধা মানে না ভাষা	আর্মি স্টেডিয়াম থেকে: মাত্র হামাগুড়ি দিতে শেখা শিশুর ...
2828015	দেশে ৫০ হাজার মেট্রিক টন পেঁয়াজ আসবে: প্রধানমন্ত্রী	জাতীয় সংসদ ভবন থেকে: অল্প কিছুদিনের মধ্যে বিদে ...

Figure 4.2: A snapshot of training dataset which has 2,828,015 records

4.2.3 Bangla corpus

This dataset has been used for POS tagging. Dataset with a larger amount of words helps to do better POS tagging, as a result, better word sense disambiguation. Our Bangla corpus has more than **1,17,042 words** along with POS or Parts Of Speech labeling.

4.3 Evaluation of Framework

The development of Bangla Word Sense Disambiguation is done with pure Python language. No framework has been used to develop it. Along with Python, HTML, CSS and pure JavaScript have been used for the web interface and Flash for the webserver. This developed system has three modes for three different purposes. The first two modes are essential, likely training mode and testing mode. These two modes are part of Supervised Machine Learning and the third mode is Graphical User Interface or GUI mode. We have used a webserver in this mode. Those three modes of our developed system are Training mode, Testing mode, and Web mode.

4.3.1 Training Mode

When this project runs in training mode, it trains the model. The training mode in supervised learning is used to train the model how to detect or identify something based on features along with labeled data. Here in this project, we train our model to detect the ambiguous words and disambiguate their meaning based on **head word feature word**. Labeled data is also known as the training set. In this system, labeled data is provided from the sqlite3 database to train that model. A snapshot of labeled data is shown in the figure: 4.1.

4.3.2 Testing Mode

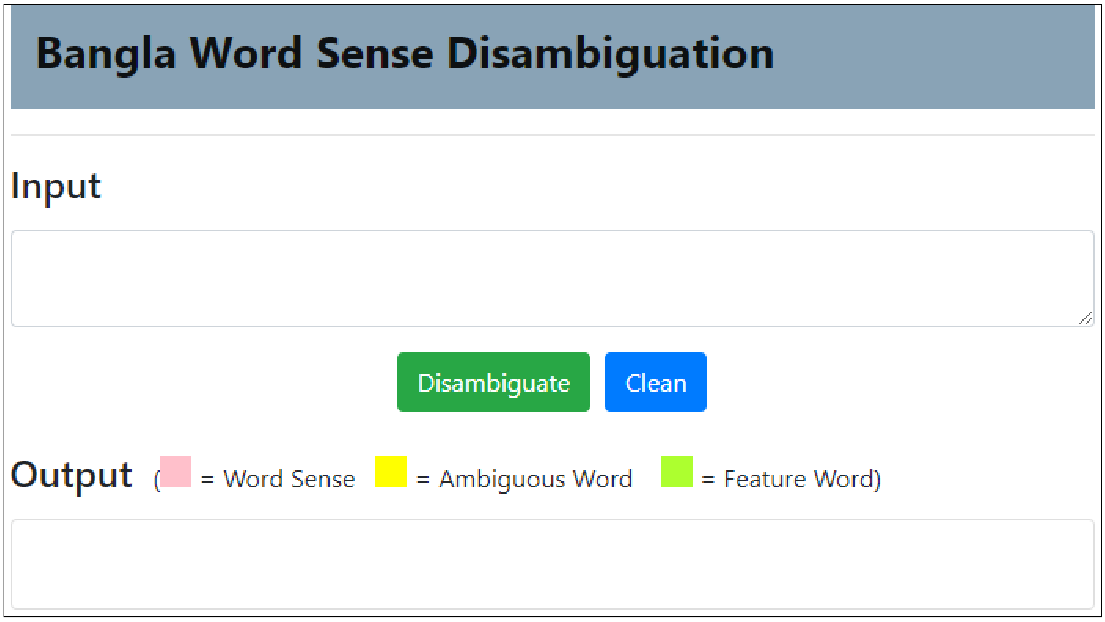
The system run in this mode tries to predict outcomes for unforeseen data. After completion of training, the test is done on that system to evaluate its performance that is, how accurately it can detect the ambiguous word and disambiguate its meaning based on the previous training. It takes a bulk amount of input from a text file containing testing data and output results as a text file as well.

For instance,
Input: সে তার পরিবারের মুখ রাখল
Output: সে/**pron**/সে   তার/**pron**/তার   পরিবারের/**verb**/পরিবার [aw_id=2, sense13=সম্মান রাখা]
মুখ/**adv**/মুখ   রাখল/**verb**/রাখা

4.3.3 Web Mode

It is also known as **GUI mode** (figure: 4.3). It has input section (figure: 4.4) and output section (figure: 4.5). User can input single or multiple lines into the input section and their corresponding results will be displayed on the output section with colored labels, where -

- Pink = Word Sense (the actual meaning of that ambiguous word based on given context)
- Yellow = Ambiguous word in a given sentence.
- Green = Feature word either left or right of an ambiguous word.



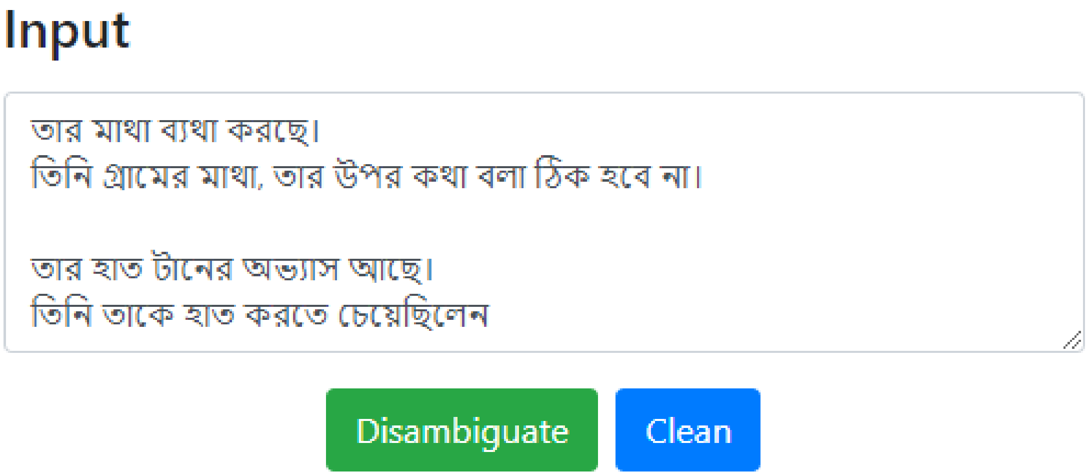
Bangla Word Sense Disambiguation

Input

Output (🟥 = Word Sense 🟨 = Ambiguous Word 🟩 = Feature Word)

Figure 4.3: Bangla Word Sense Disambiguation System GUI

Input box is editable and We can input single or multiple sentences at a time. **Clear** button cleans the input box into blank and **Disambiguation** button displays the disambiguated results into the output box.



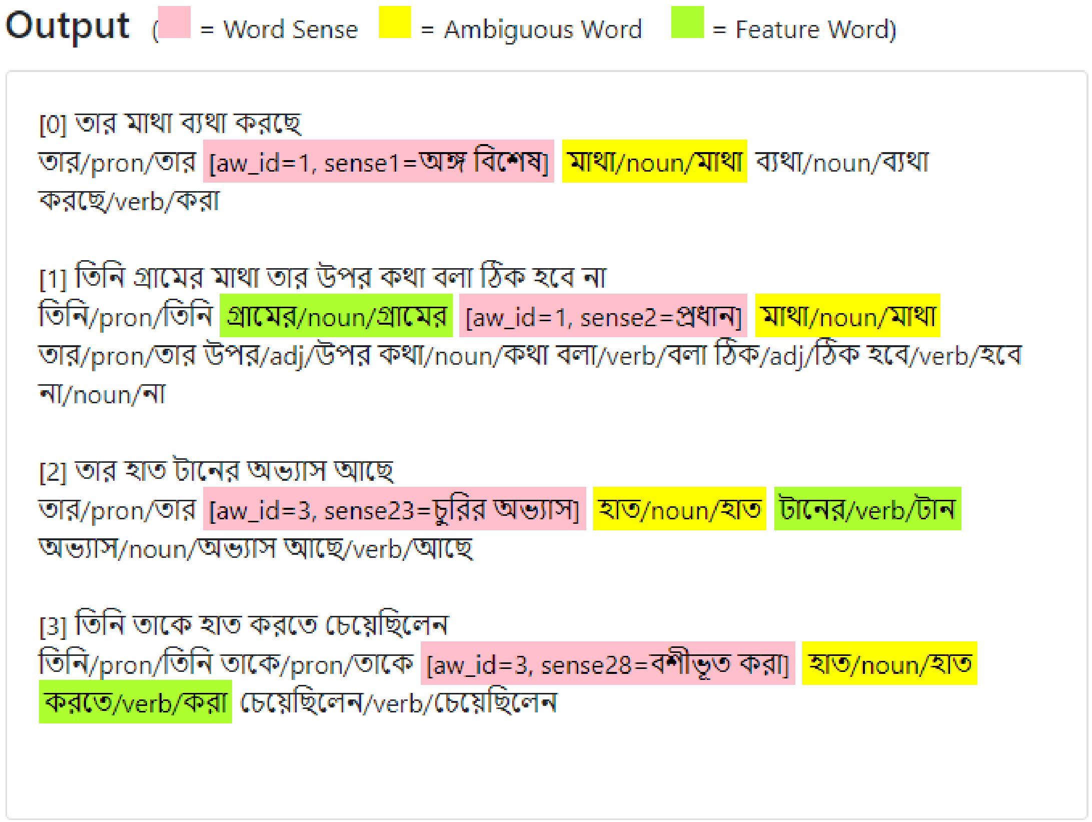
Input

তার মাথা ব্যথা করছে।
তিনি গ্রামের মাথা, তার উপর কথা বলা ঠিক হবে না।

তার হাত টানের অভ্যাস আছে।
তিনি তাকে হাত করতে চেয়েছিলেন

Figure 4.4: Input into Bangla Word Sense Disambiguation System

Output box is non-editable and read-only. Disambiguated results with colored background display here.



Output (🟥 = Word Sense 🟨 = Ambiguous Word 🟩 = Feature Word)

[0] তার মাথা ব্যথা করছে
তার/pron/তার [aw_id=1, sense1=অঙ্গ বিশেষ] মাথা/noun/মাথা ব্যথা/noun/ব্যথা
করছে/verb/করা

[1] তিনি গ্রামের মাথা তার উপর কথা বলা ঠিক হবে না
তিনি/pron/তিনি গ্রামের/noun/গ্রামের [aw_id=1, sense2=প্রধান] মাথা/noun/মাথা
তার/pron/তার উপর/adj/উপর কথা/noun/কথা বলা/verb/বলা ঠিক/adj/ঠিক হবে/verb/হবে
না/noun/না

[2] তার হাত টানের অভ্যাস আছে
তার/pron/তার [aw_id=3, sense23=চুরির অভ্যাস] হাত/noun/হাত টানের/verb/টান
অভ্যাস/noun/অভ্যাস আছে/verb/আছে

[3] তিনি তাকে হাত করতে চেয়েছিলেন
তিনি/pron/তিনি তাকে/pron/তাকে [aw_id=3, sense28=বশীভূত করা] হাত/noun/হাত
করতে/verb/করা চেয়েছিলেন/verb/চেয়েছিলেন

Figure 4.5: Output from Bangla Word Sense Disambiguation System

4.4 Evaluation of Performance

Earlier we mentioned (section 4.1) the runtime environment used for this system to execute. All training sets and testing sets were executed in that mentioned environment.

We trained our model with a training set, then we evaluated the system performance by providing testing sets. To evaluated the system performance, the following formulae have been used -

$$\text{accuracy rate} = \left(\frac{\text{no. of total items} - \text{no. of total wrong items}}{\text{no. of total items}}\right) \times 100\% \quad (4.1)$$

or

$$\text{accuracy rate} = \left(\frac{\text{no. of total correct items}}{\text{no. of total items}}\right) \times 100\% \quad (4.2)$$

and

$$\text{error rate} = \left(\frac{\text{no. of total wrong items}}{\text{no. of total items}}\right) \times 100\% \quad (4.3)$$

To calculate accuracy rate equation: 4.1 and error rate equation: 4.3 has been used in this report.

4.4.1 Result of POS Tagging

At first, we evaluated the POS Tagging system by providing the testing sets (figure: 4.1). The first testing set for the POS Tagging system contains around 100077 sentences and out of 1321016 words, it accurately tagged 1228545 words which gives us an accuracy rate of 93%. The second testing test was conducted with 90605 sentences and out of 1195986 words, it accurately did 1136186 words which is a 95% accuracy rate. And our last was comparatively large than the previous, it has 117580 sentences, successfully done 1396850 out of 1552056 words which is 90%. The overall accuracy rate is 92.7% which is an acceptable figure.

Our system gave an overall 7.8% error or inaccuracy to do POS Tagging because of -

- limitation of enough Bangla corpus in electronic format.

- limitation of enough contexts.
- some of the contexts are too hard to detect.

No. of Input Sentences	No. of Words	No. of Accurate POS Tagging	Accuracy Rate	Overall Accuracy Rate
100077	1321016	1228545	93%	
90605	1195986	1136186	95%	92.7%
117580	1552056	1396850	90%	

Table 4.1: Accuracy Rate based on POS Tagging

4.4.2 Result of Ambiguous Word Detection

This section is done with five testing sets which contain 500357, 507310, 450067, 478940, 400632 sentences respectively. Among them, the first set gave 91% accuracy by successfully detecting 7793 out of 8564 ambiguous words. The second testing set was 90% accurate (7760 out of 8623 ambiguous words) and similarly third, fourth, and fifth testing sets were 87%, 87%, and 86% accurate respectively.

As a result, the overall accuracy rate to detect ambiguous words is 88.2% which is less than POS tagging accuracy (92.7%). Accurate POS Tagging does not always give correct ambiguous word detection because some of the sentence structures are too much complex to normalize. Consequently, our overall result reduced to 88.2% from 92.7%.

No. of Input Sentences	No. of Available Ambiguous Words	No. of Detected Ambiguous Words	Accuracy Rate	Overall Accuracy Rate
500357	8564	7793	91%	
507310	8623	7760	90%	
450067	8492	7389	87%	88.2%
478940	8649	7519	87%	
400632	8167	7052	86%	

Table 4.2: Accuracy Rate based on Ambiguous Word Detection

4.4.3 Result of Disambiguation

The last section or the main section of this project is disambiguation (more specifically Bangla Word Sense Disambiguation) accuracy. It has also conducted with five testing sets like the previous section. Actually ambiguous word detection and disambiguation are done simultaneously. But the outcomes were different from the ambiguous word detection accuracy. The testing sets of first, second, third, fourth, and fifth gave accuracy rate 85%, 84%, 86%, 85%, and 83% respectively for disambiguation whereas, the accuracy rate for ambiguous word detection were 91%, 90%, 87%, 87%, and 86% respectively.

The overall accuracy rate is 84.6% which is further reduced from 88.2%. This further reduction happens for complexity arises during disambiguation. Detection of ambiguous work is one of the sub-work of disambiguation. A disambiguation system first detects ambiguous word then use its algorithm to disambiguate that word in that context. This additional algorithm creates additional complexity, consequently additional errors in the result.

No. of Input Sentences	No. of Ambiguous Sentences	No. of Accurate Disambiguation	Accuracy Rate	Overall Accuracy Rate
500357	8564	7279	85%	
507310	8623	7243	84%	
450067	8492	7303	86%	84.6%
478940	8649	7351	85%	
400632	8167	6778	83%	

Table 4.3: Accuracy Rate based on Disambiguation

4.4.4 Overview of Accuracy

An overview of POS Tagging, Ambiguous Word Detection, and Disambiguation Accuracy rate corresponding to their testing sets is shown in the graph: 4.6. Three testing sets have been used for POS Tagging System and five for both Ambiguous Word Detection and Disambiguation System.

The overall results produced by POS Tagging, Ambiguous Word Detection, and

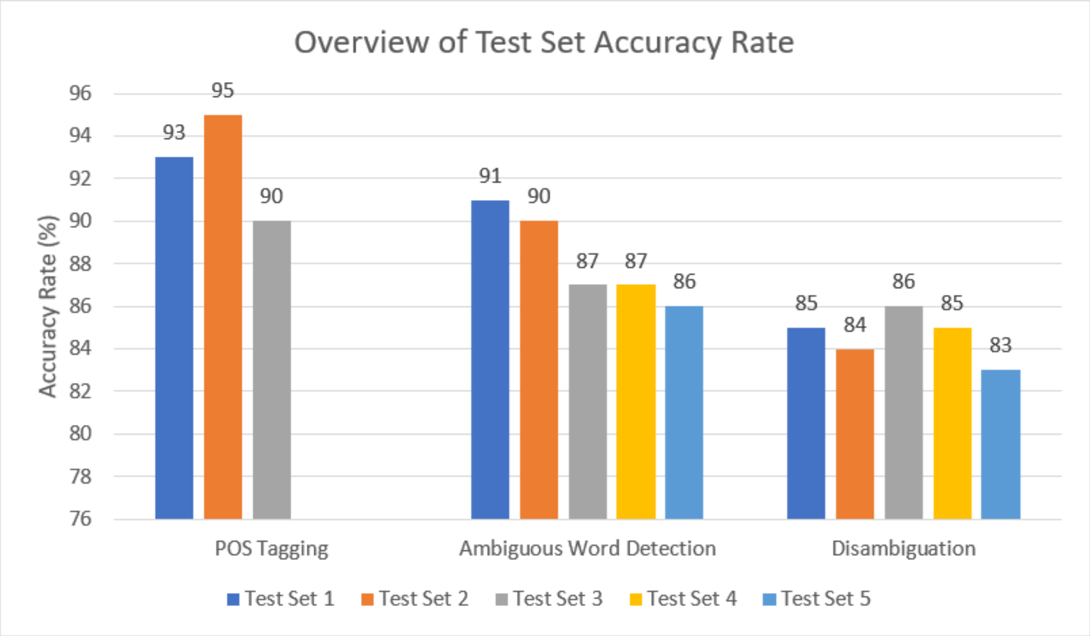


Figure 4.6: Graph of POS Tagging, Ambiguous Word Detection and Disambiguation Accuracy

Disambiguation Systems along with their linear relations are shown in the graph: 4.7.

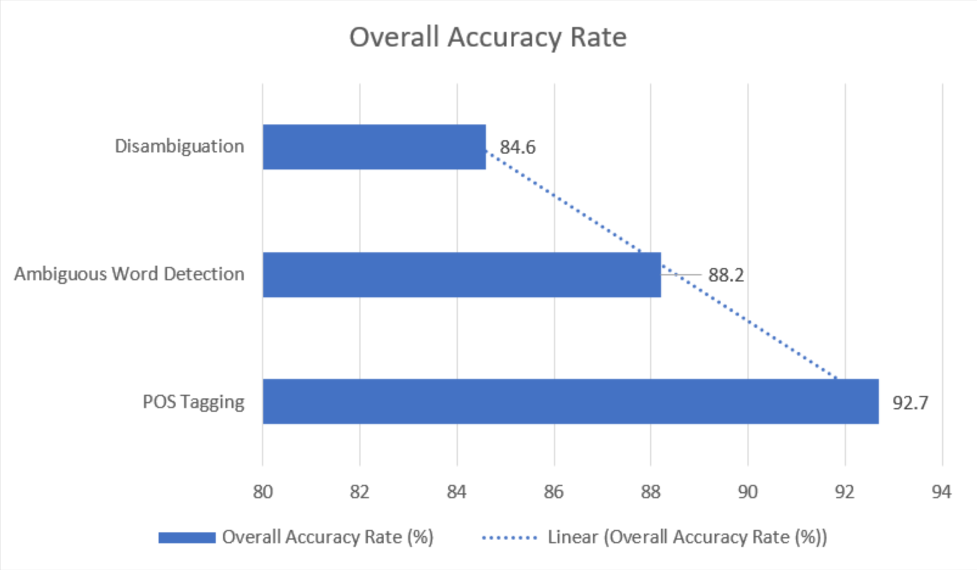


Figure 4.7: Graph of POS Tagging, Ambiguous Word Detection and Disambiguation Overall Accuracy

4.4.5 Error in the System and Its Limitation

Everything system has more or less error in it. A system with lesser error gives more accurate results and better performance. From sections 4.4.1, 4.4.2 and 4.4.3, it has been found that they have errors of 7.3%, 11.8% and 15.4% respectively.

Here some errors are obvious and hard to solve. For instance,

তোমার এতো মাথা ব্যথার কারণ কি?

We know that **"মাথা"** is an ambiguous word and **"ব্যথা"** is a feature word. Here **"মাথা"** could mean **headache** or **concern** which still gives us dual meaning that means we cannot disambiguate this. Because sometimes the meaning of the word depends on the tone of the voice of the speaker rather than feature words which are hard to disambiguate even for humans themselves.

4.5 Conclusion

In this chapter, we have discussed the dataset we used for our project, our system framework, and its components, and lastly, the performance and evaluation of the produced results as well as error. We didn't discuss any social and environmental or ethical impact in this chapter because our developed system has no impact on those. The ending of this chapter brings us to the conclusion which will be discussed in the next chapter.

Chapter 5

Conclusion

5.1 Conclusion

Development of Bangla Word Sense Disambiguation is not a one-time development rather it's a continuous process. The more we work with it, the better result we will get. However, we have tried to do some improvements in this field within our time, resources and limit.

According to the objectives, we made during Final Project Proposal, we have made contributions to -

- Bangla Lexicon database by adding 1,17,042 words.
- Bangla Corpus Database by adding 64,760,021 (about 65 million) sentences and 854,589,384 (about 1 billion) words
- Bangla POS Tagging which provides 92.7% accuracy
- Bangla Ambiguous Word Detection which provides 88.2% accuracy
- Bangla Word Sense Disambiguation which provides 84.6% accuracy
- Bangla Natural Language Toolkits by developing it.

5.2 Future Work

There is a lot to develop on Bangla Word Sense Disambiguation. With the advancement of science, more and more methodologies are being developed. We already mentioned that the unavailability of enough Bangla Lexicon and Corpus database are the biggest problem to work with Bangla. Continuous contributions to Bangla will definitely help to reduce that problem. We will also try to enlarge

our existing Bangla Communities and make new as much as possible.
Sometimes the change in methodology could give us more accurate results and so we will try different methodologies as well.

I deployed [23] and published this repository [24] online so that, anyone can visit, reuse, make changes and contribute to this project for further improvement.

Gitlab Ripo: **https://gitlab.com/tareqas/bwsd**
Online version of BWSD: **https://bwsd.herokuapp.com**

References

[1] C. P. Singh *et al.*, 'A review on word sense disambiguation emphasizing the data resources on wordnet and corpus,' *INFORMATION TECHNOLOGY IN INDUSTRY*, vol. 9, no. 2, pp. 996–1016, 2021 (cit. on p. 8).

[2] B. M. Kumar and M. Siddappa, 'Kannada word sense disambiguation using semantic relations,' in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1767, 2021, p. 012 025 (cit. on p. 8).

[3] A. Haque and M. M. Hoque, 'Bangla word sense disambiguation system using dictionary based approach,' *ICAICT, Bangladesh*, 2016 (cit. on p. 8).

[4] A. R. Pal, D. Saha and S. K. Naskar, 'Word sense disambiguation in bengali: A knowledge based approach using bengali wordnet,' in *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2017, pp. 1–5. DOI: 10.1109/ICECCT.2017.8117900 (cit. on p. 8).

[5] A. R. Pal, D. Saha and N. S. Dash, 'Automatic classification of bengali sentences based on sense definitions present in bengali wordnet,' *arXiv preprint arXiv:1508.01349*, 2015 (cit. on p. 8).

[6] N. Islam, N. Nawaz, M. Tanzil and D. M. Ali, 'Supervised information theoretic bangla word sense disambiguation using lexeme list,' in *Conference on Language and Technology (CLT07), University of Peshawar, Pakistan*, 2007 (cit. on p. 8).

[7] R. Pandit and S. K. Naskar, 'A memory based approach to word sense disambiguation in bengali using k-nn method,' in *2015 IEEE 2nd international conference on recent trends in information systems (ReTIS)*, IEEE, 2015, pp. 383–386 (cit. on p. 9).

[8] N. T. T. Aung, K. M. Soe and N. L. Thein, 'A word sense disambiguation system using nave bayesian algorithm for myanmar language,' *International Journal of Scientific & Engineering Research*, vol. 2, no. 9, pp. 1–7, 2011 (cit. on p. 9).

[9] D. Yarowsky, 'Unsupervised word sense disambiguation rivaling supervised methods,' in *33rd annual meeting of the association for computational linguistics*, 1995, pp. 189–196 (cit. on p. 9).

[10] S. Nazah, M. M. Hoque and M. R. Hossain, 'Word sense disambiguation of bangla sentences using statistical approach,' in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, IEEE, 2017, pp. 1–6 (cit. on p. 9).

[11] A. Ranjan Pal, A. Kundu, A. Singh, R. Shekhar and K. Sinha, 'A hybrid approach to word sense disambiguation combining supervised and unsupervised learning,' *arXiv e-prints*, arXiv–1611, 2015 (cit. on p. 9).

[12] *Patrika read latest bengali news from west bengal's leading newspaper.* [Online]. Available: `https://www.anandabazar.com/` (cit. on p. 12).

[13] *News, behind the news.* [Online]. Available: `https://www.banglatribune.com/` (cit. on p. 12).

[14] *Bdnews24.com.* [Online]. Available: `https://bangla.bdnews24.com/` (cit. on p. 12).

[15] Banglanews24, *Bangla news and entertainment 24x7.* [Online]. Available: `https://www.banglanews24.com/` (cit. on p. 13).

[16] বাংলাপিডিয়ায় স্বাগতম! [Online]. Available: `http://bn.banglapedia.org/index.php` (cit. on p. 13).

[17] দৈনিক ইত্তেফাক. [Online]. Available: `https://www.ittefaq.com.bd/` (cit. on p. 13).

[18] *Most popular bangla news: Entertainment: Breaking news.* [Online]. Available: `https://www.jugantor.com/` (cit. on p. 13).

[19] K. Kantho. [Online]. Available: `https://www.kalerkantho.com/` (cit. on p. 13).

[20] প্রথম আলো: বাংলা নিউজ পেপার. [Online]. Available: `https://www.prothomalo.com/` (cit. on p. 13).

[21] *Get the latest online bangla news !* [Online]. Available: `https://samakal.com/` (cit. on p. 13).

[22] *World's largest bangla blog community.* সামহোয়্যার ইন ব্লগ - বাঁধ ভাঙার আওয়াজ । বাংলা ব্লগ|. [Online]. Available: `https://www.somewhereinblog.net/` (cit. on p. 13).

[23] *Online bwsd by tareqas.* [Online]. Available: `https://bwsd.herokuapp.com` (cit. on p. 31).

[24] *Developing a supervised system using lexicon database for bangla word sense disambiguation.* [Online]. Available: `https://gitlab.com/tareqas/bwsd` (cit. on p. 31).