CSE 4000: Thesis/ Project

DEPENDENCY PARSING FOR BANGLA TEXT

By

Shaikh Prothomaa Binte Minhaz

Roll: 1807010

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

February, 2024

Dependency Parsing for Bangla Text

By

Shaikh Prothomaa Binte Minhaz

Roll: 1807010

A thesis submitted in partial fulfillment of the requirements for the degree of

"Bachelor of Science in Computer Science and Engineering"

Supervisor:

Dr. K. M. Azharul Hasan

Professor

Department of Computer Science and Engineering

Khulna University of Engineering & Technology, Khulna

Signature

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

February, 2024

Acknowledgment

All the praise to God, whose blessing and mercy succeeded me to complete this thesis work fairly. I would like to express my sincere gratitude and appreciation to all those who have supported me throughout the course of this research endeavor. First and foremost, I am deeply indebted to my thesis supervisor, Dr. K. M. Azharul Hasan, Professor, Department of Computer Science and Engineering, Khulna University of Engineering & Technology, for his unwavering guidance, insightful feedback, and continuous encouragement. His expertise and mentorship has been invaluable in shaping the direction of this research. My sincere thanks go to the Department of Computer Science and Engineering at Khulna University of Engineering and Technology (KUET) for providing me with the necessary resources, facilities, and opportunities to conduct this research. I am grateful for the stimulating academic environment that has nurtured my intellectual growth. Furthermore, I extend my gratitude to my friends and family for their unwavering belief in my abilities and their constant encouragement. Their emotional support has been my pillar of strength. In conclusion, this research would not have been possible without the collective efforts, support, and contributions of all those mentioned above and many others who have left an indelible mark on this journey. Thank you for being a part of this significant chapter in my academic and personal growth.

Author

Abstract

A fundamental task in natural language processing (NLP) called dependency parsing involves analyzing the grammatical structure of sentences by detecting the syntactic connections between words. This study examines the difficulties and methods for performing dependency parsing for Bangla (Bengali) text, a language with complex morphology and distinctive syntactic properties. The article addresses the value of dependency parsing in capturing the linguistic subtleties of Bangla sentences and their applications in various NLP tasks. In this paper, a method is proposed for developing dependency parsing on Bangla text by using a graph-based approach. Here the parsing tree is generated from a directed graph of Bangla text input. This method helps to enrich the Bangla language linguistic resources and annotated corpora that help to use language globally.

Contents

		PAGE
Title Page		i
Acknowledgment		ii
Abstract		iii
Contents		iv
List of Tables		vi
List of Figures		vii
CHAPTER I	**Introduction**	**1**
	1.1 Introduction	1
	1.2 Background	1
	1.3 Objectives	2
	1.4 Scope	2
	1.5 Unfamiliarity of the Problem	2
	1.6 Project Planning	3
	1.7 Applications of the work	4
	1.8 Organization of the thesis	4
CHAPTER II	**Literature Review**	**5**
	2.1 Introduction	5
	2.2 Literature Review	6
	2.3 Discussion of research gap solution	7
CHAPTER III	**Dependency Parser Generation for Bangla Text**	**8**
	3.1 Introduction	8
	3.2 Definition of terms	9
	3.3 Proposed algorithm	14
CHAPTER IV	**Implementation, Results and Discussions**	**17**

	4.1 Experimental Setup	17
	4.2 Evaluation Metrics	17
	4.3 Dataset	18
	4.4 Implementation and Results	19
	4.5 Objective Achieved	22
	4.6 Financial Analyses and Budget	22
CHAPTER V	**Societal, Health, Environment, Safety, Ethical, Legal and Cultural Issues**	23
	5.1 Intellectual Property Considerations	23
	5.2 Ethical Considerations	23
	5.3 Safety Considerations	23
	5.4 Legal Considerations	24
	5.5 Impact of the Project on Societal, Health, and Cultural Issues	24
	5.6 Impact of Project on the Environment and Sustainability	25
CHAPTER VI	**Addressing Complex Engineering Problems and Activities**	26
	6.1 Complex engineering problems associated with the current thesis	26
	6.2 Complex engineering activities associated with the current thesis	27
CHAPTER VII	**Conclusions**	**28**
	7.1 Summary	28
	7.2 Limitations	28
	7.3 Recommendations and Future Works	29
Appendix		30
References		31

List of Tables

Table no.	Description	Page
1.1	Gantt chart showing progress	3
2.1	summary of other papers	7
3.1	Graph after assigning weight	11
3.2	Edmonds algorithm	15
4.1	Accuracy of training dataset	19
4.2	Dependencies of English text	20
4.3	Dependencies of Bangla text	21
4.4	Specifications of budgeting	22

List of Figures

Figure no.	Description	Page
3.1	Graph by tokens	10
3.2	Graphical representation of dependencies in Bangla sentence	13
3.3	Input page	13
3.4	Output page	14
3.5	Proposed Methodology	16
4.1	Graphical representation of dependencies in English sentence	20

CHAPTER I

Introduction

1.1 Introduction

Dependency parsing is a crucial component of natural language processing (NLP) that helps understand the syntactic structure of sentences by identifying word relationships. However, its application to languages with complex morphologies like Bangla presents unique challenges and opportunities. Bangla, spoken by millions worldwide, has a rich linguistic heritage with intricate grammatical rules and flexible word order, making it an intriguing subject for computational linguistics research. The importance of dependency parsing in Bangla text processing is significant, as it forms the basis for various NLP applications such as machine translation, sentiment analysis, information retrieval, and question answering. However, the development of efficient and accurate dependency parsers tailored specifically for Bangla text has been hindered by the lack of annotated corpora, linguistic resources, and research focus compared to more widely studied languages. This thesis work aims to contribute to this emerging field by providing a comprehensive analysis and implementation of dependency parsing techniques for Bangla text. By leveraging advances in computational linguistics and machine learning, the researchers aim to develop robust and accurate dependency parsers that can effectively handle the complexities of Bangla syntax. This research not only advances dependency parsing for Bangla but also paves the way for broader applications of NLP in the Bangla-speaking world.

1.2 Background

Due to the importance of comprehending the syntactic links between words in this rich and complicated language, Dependency Parsing (DP) in Bangla (Bengali), literature has drawn more and more interest from researchers and practitioners. Analyzing the grammatical structure of sentences by determining how words depend on one another is the fundamental

goal of dependency parsing in natural language processing (NLP). Bangla language has extensive syntactic and morphological elements, such as flexible word order, intricate verb structures, and a range of noun inflections. This complexity makes it difficult for NLP tasks like dependency parsing, which try to effectively capture these linguistic nuances. Developing Bangla-specific tools and models became more and more important as NLP research and technology developed, according to experts and linguists. Due to its importance for comprehending the language's syntax and semantics, dependency parsing has become a crucial topic of research. Sentiment analysis, information extraction, machine translation, and other NLP applications can all benefit from accurate dependency parsing in Bangla. It improves language comprehension and makes advanced language technology possible.

1.3 Objectives

- To find out the syntactic relationships between words in Bangla sentences.
- To create the annotated corpora and linguistic resources for Bangla
- To create a dependency tree of Bangla input text
- To find out the parts of speech of each word in the input

1.4 Scope

Dependency parsing enables an in-depth analysis of the syntactic relationships and structure within Bangla sentences, advancing linguistic study and illuminating the peculiarities of the language. Several NLP applications, including text summarization, question answering, speech recognition, and chatbots, benefit from dependency parsing by enhancing accuracy and efficiency. Dependency parsing can be applied to activities that are specialized to a given domain, boosting information extraction in areas like legal texts, medical records, and technical manuals in Bangla.

1.5 Unfamiliarity of the Problem

This proposed method is implementing dependency parsing for Bangla text is quite different from other methods. Hereafter parts of speech tagging, the work implements the maximum spanning tree of the given Bangla text from the directed graph using the Edmonds algorithm, transition and emission probabilities. This approach enhances the linguistic resources of the

Bangla language and its global usage. This approach helps the computer system to enrich its knowledge base and generate human-like output.

1.6 Project planning

Here is a timeline that illustrates the key stages of this research journey:

Table 1.1: Gantt chart showing progress

Task Name	**1st Term**						**2nd Term**					
	1-2	3-4	5-6	7-9	10-12	13	1-2	3-4	5-6	7-10	11-12	13
Topic Selection												
Thesis Planning												
Literature Review												
Paper Selection												
Data Gathering												
Progress Defense Preparation												
Feature Extraction												
Feature Fusion												
Model Training and Evaluation												
Thesis Report Manuscript												
Thesis Defense												
Final Manuscript												

1.7 Applications of the Work

- This work helps to improve the accuracy of machine translation systems by providing insights into the syntactic structure of sentences.
- It contributes to understanding the syntactic and semantic structure of questions and passages, facilitating better matching of questions with relevant answers.
- This can reveal the syntactic structure for improving sentiment analysis and opinion.
- It supports linguistic research by revealing syntactic relationships, grammatical patterns and language structures in Bangla text.
- This can aid in aligning speech with text, improving the accuracy of speech recognition systems. It enables more intuitive and accurate interactions with computers through natural language interfaces, facilitating voice assistants and chatbots.

1.8 Organization of the Report

This report has seven chapters. Each chapter provides a detailed demonstration of this thesis work.

Chapter I illustrates the background and reason for selecting this topic for the thesis. Also represents the planning for progress, objectives and application in real life of this method and the unfamiliarity of this work.

Chapter II interprets the related works of this thesis work that helped to establish this model and also the uniformity of the thesis topic.

Chapter III describes the proposed methodology of the work in detail.

Chapter IV describes about experimental setup, evaluation, dataset, implementation and result. Also, this chapter includes the achieved objective and financial budget for this thesis.

Chapter V contains societal, health, environment, safety, ethical, legal and cultural issues.

Chapter VI contains complex engineering problems and activities.

Chapter VII contains a summary of this thesis work and possible future works.

CHAPTER II

Literature Review

2.1 Introduction

In natural language processing (NLP), dependency parsing is a fundamental process that includes analyzing the syntactic structure of sentences by determining the relationships between words. These relationships are commonly shown as directed connections between words, where the syntactic function of each word within the sentence depends on the function of the others. By determining which words are heads and which are dependents, dependency parsing aims to show the hierarchical structure of sentences.
In a directed graph, the nodes represent words and edges for the relationships among them, It can be used to illustrate the dependence structure of a phrase. Every word in the sentence has a single headword that determines its syntactic function. Subject-verb, modifier, and object relationships are among the relationships that dependency parsing detects.

Dependency parsing is essential to many NLP applications such as question answering, information retrieval, machine translation, sentiment analysis, and syntactic and semantic analysis. These applications perform better in tasks requiring deep language comprehension because accurate dependency parsing helps them to better understand the relationships between words in a phrase. Dependency parsing has been an important topic of research and development in the field of natural language processing (NLP) because of its current performance standards over different languages and domains. Dependency parsing is fundamental to many NLP applications, and research is still being done to increase its precision, effectiveness, and suitability in a variety of language settings.

This chapter includes a literature review and discussion of the research gap solution.

2.2 Literature Review

Utpal Garain et al. (2014) have described a Grammar-driven dependency parsing for Bangla. Bangla language has a free-word order nature which makes parsing a sentence difficult. The Paninian grammatical model is used to solve this difficulty by simplifying the complex and compound sentences and then parsing the simple sentences by satisfying the Karaka demands of the Demand Groups (Verb Groups), and finally rejoining such parsed structures with appropriate links and Karaka labels. The parser has been trained with a Treebank of 1000 annotated sentences and then evaluated with un-annotated test data of 150 sentences.

Arnab Dhar et al. (2012) described a two-stage dependency parser for Bangla. In the first stage, authors build a model using a Bangla dependency Treebank released in ICON 2009 and subsequently, this model is used to build a data-driven Bangla parser [1]. In the second stage, constraint-based parsing has been used to modify the output of the data-driven parser. This implements the Bangla-specific constraints with the help of demand frames of Bangla verbs. In the data-driven module, authors use Covington's algorithm as implemented in MaltParser by Nivre (2006, 2007, 2009) for statistically annotating the dependency relations in Bangla sentences.

Urmi Ghosh et al. (2019) develop a code-mixing (CM) NLP system that has significantly gained importance in recent times due to an upsurge in the usage of CM data by multilingual speakers [5]. The authors present a rule-based system to computationally generate a synthetic code-mixing treebank for Bengali and English (Syn-BE) which is used to further improve the accuracy of dependency parser. They use a dataset of 500 Bengali-English tweets annotated under the Universal Dependencies scheme.

Sanjay Chatterji et al. in the paper "Grammar Driven Rules for Hybrid Bengali Dependency Parsing" describe a hybrid approach for parsing Bengali sentences based on the dependency tagset and Treebank released in the ICON 2009 tool contest [3]. A data-driven dependency parser is considered a baseline system. Some handcrafted rules are identified based on the error patterns in the output of the baseline system.

Phani Gadde et al. present a data-driven dependency parsing strategy that makes use of sentence-specific clause information to enhance the parser efficiency. A partial parser is used to automatically incorporate the clausal information. We use a modified version of MSTParser to demonstrate the experiments on Hindi, a morphologically rich free-word-

order language. We conducted all of our studies using the parsing contest data from ICON 2009. Our unlabeled attachment and labeled attachment accuracy improvements over the baseline parsing accuracy were 0.87% and 0.77%, respectively.

2.3 Discussion of Research Gap Solution

In this proposed method, dependency parsing is implemented for Bangla text using a graph-based approach which is not done yet in other research papers that are shown in Table 2.1.

Table 2.1: summary of other papers

Paper name	Author	Methodology	Dataset
A Hybrid Dependency Parser for Bangla	Arnab Dhar, Sanjay Chatterji, Sudeshna Sarkar, Anupam Basu	A two-stage dependency parser for Bangla they are a Data-driven module and a Grammar-driven module	Bangla Treebank was released in ICON 2009
Evaluation of Two Bengali Dependency Parsers	Arjun Das, Arabinda Shee, Utpal Garain	two dependency parsers for a free-word order for Bangla. Grammar-based parser and Maltparser	ICON NLP Tool Contest data and Dataset-II (developed by authors)
Non-projective Dependency Parsing using Spanning Tree Algorithms	Ryan McDonald, Fernando Pereira, Kiril Ribarov, Jan Hajic	Formalize weighted dependency parsing as searching for maximum spanning trees (MSTs) in directed graphs.	Prague Dependency Treebank

CHAPTER III

Dependency Parser Generation for Bangla Text

3.1 Introduction

The technique of a dependency parsing algorithm describes the sequential steps the algorithm takes to examine a sentence and build a dependency tree that represents the syntactic relationships between words. Dependency parsing algorithms can be categorized into transition-based and graph-based approaches. Here is a general description of the proposed methodology of the graph-based approach:

Step 1: Take user input as a paragraph or a sentence.

Step 2: If the input is a paragraph then sentence separation occurs according to ending symbols.

Step 3: Each word of a sentence converts into tokens and identifies the POS of each token in a sentence.

Step 4: Create a fully connected directed graph using the tokens with a root node that has only outgoing edges to each node of the graph.

Step 5: Assign weights to each edge using transition and emission probabilities added with some factor.

Step 6: Execute Edmond's algorithm to get the max spanning tree. It provides nodes with a single parent.

Step 7: According to the transition probability dictionary calculate the accuracy of each sentence.

3.2 Definition of terms

Separation of sentences: In this work, the user can input a paragraph. So each sentence needed to be separated. Sentences are separated by using "।", "!", "?" symbols for Bangla text and for English text ".", "!", "?" symbols are used. For example,

Input paragraph: প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ। এই দেশে পরিচিত অপরিচিত অনেক পর্যটক-আকর্ষক স্থান আছে। এর মধ্যে প্রত্নতাত্ত্বিক নিদর্শন, ঐতিহাসিক মসজিদ এবং মিনার, পৃথিবীর দীর্ঘতম প্রাকৃতিক সমুদ্র সৈকত, পাহাড়, অরণ্য ইত্যাদি অন্যতম। এদেশের প্রাকৃতিক সৌন্দর্য পর্যটকদের মুগ্ধ করে। বাংলাদেশের প্রত্যেকটি এলাকা বিভিন্ন স্বতন্ত্র বৈশিষ্ট্যে বিশেষায়িত ।

After separating into sentence,

প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ।

এই দেশে পরিচিত অপরিচিত অনেক পর্যটক-আকর্ষক স্থান আছে।

এর মধ্যে প্রত্নতাত্ত্বিক নিদর্শন, ঐতিহাসিক মসজিদ এবং মিনার, পৃথিবীর দীর্ঘতম প্রাকৃতিক সমুদ্র সৈকত, পাহাড়, অরণ্য ইত্যাদি অন্যতম।

এদেশের প্রাকৃতিক সৌন্দর্য পর্যটকদের মুগ্ধ করে।

বাংলাদেশের প্রত্যেকটি এলাকা বিভিন্ন স্বতন্ত্র বৈশিষ্ট্যে বিশেষায়িত ।

Lexical analysis: Lexical analysis, also known as tokenization is a crucial phase because it divides the input phrase into individual words (tokens) and gives each one linguistic information that will be used to establish the syntactic relationships between them in the dependency tree.

Parts of speech (POS) detection: Define the POS of each word in a sentence. Bangla language is a rich morphological language. So to detect the POS of each word while considering morphological analysis. For Bangla text POS tagging and tokenization, BNLP is used. For English text POS tagging and tokenization, NLTK is used.

The tokenization and POS detection are done using BNLP. For example,

Sentence 1: প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ।

After tokenization and POS detection: [('প্রাকৃতিক', 'JJ'), ('রূপবৈচিত্র্যে', 'NC'), ('ভরা', 'NC'), ('আমাদের', 'PPR'), ('এই', 'DAB'), ('বাংলাদেশ', 'NC'), ('।', 'PU')]

Edge labeling: Make a fully connected directed graph with words as the nodes and edges G = (V, E) standing in possible word dependency relationships. Assign the weights to the edges based on linguistic characteristics, transition and emission probabilities which are from training annotated corpus. The transition probability refers to the probability of transitioning from one hidden state to another. The emission probability refers to the probability of observing a particular output (emission) symbol from a given hidden state. Here also add a 'root' node which only has outgoing edges to all nodes. In figure-3.1, the edges are shown for 'root' and 'রূপবৈচিত্র্যে' nodes. Like this, all other edges for all nodes need to be created for generating a fully connected graph.

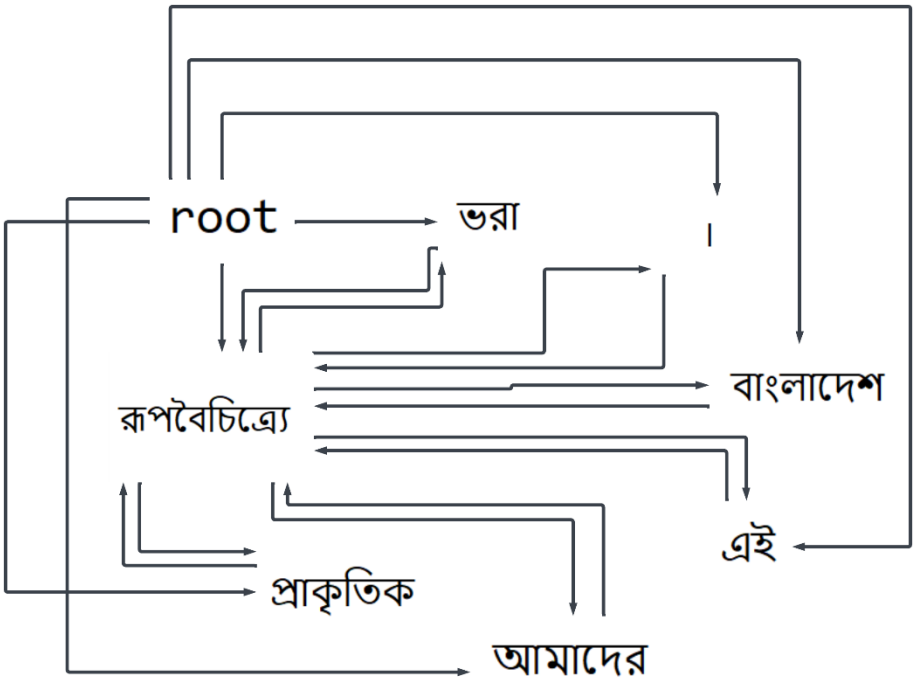


Figure 3.1: Graph by tokens

Table 3.1: Graph after assigning weight

From	To	Weight
root	প্রাকৃতিক	0.0784313725490196
	রূপবৈচিত্র্যে	0.7209302325581395
	ভরা	0.31794871794871793
	আমাদের	7.564102564102565
	এই	0.275
	বাংলাদেশ	0.625
	।	0.6377171215880894
প্রাকৃতিক	রূপবৈচিত্র্যে	360.4651162790697
	ভরা	52.99145299145299
	আমাদের	8.404558404558404
	এই	5.092592592592593
	বাংলাদেশ	3.858024691358024
	।	0.6377171215880894
রূপবৈচিত্র্যে	প্রাকৃতিক	0.0784313725490196
	ভরা	0.6542154690302836
	আমাদের	0.10375998030319016
	এই	0.06287151348879742
	বাংলাদেশ	0.047629934461210166
	।	0.6377171215880894
ভরা	প্রাকৃতিক	0.0784313725490196
	রূপবৈচিত্র্যে	0.018313525188186235
	আমাদের	0.001280987411150496
	এই	0.0007761915245530546
	বাংলাদেশ	0.0005880238822371626
	।	0.6377171215880894

আমাদের	প্রাকৃতিক	0.0784313725490196
	ভরা	0.00022609290355785473
	রূপবৈচিত্র্যে	3.323758924098377e-05,
	এই	9.582611414235244e-06
	বাংলাদেশ	7.259554101693365e-06
	।	0.6377171215880894
এই	প্রাকৃতিক	0.0784313725490196
	ভরা	2.791270414294503e-06
	আমাদের	4.1034060791337996e-07
	রূপবৈচিত্র্যে	6.508090286798232e-08
	বাংলাদেশ	8.962412471226377e-08
	।	0.6377171215880894
বাংলাদেশ	প্রাকৃতিক	0.0784313725490196
	ভরা	3.4460128571537076e-08
	আমাদের	5.065933431029382e-09
	এই	8.03467936641757e-10
	রূপবৈচিত্র্যে	4.868470972024205e-10
	।	0.6377171215880894
।	প্রাকৃতিক	0.0784313725490196
	ভরা	4.25433686068359e-10
	আমাদের	6.254238803739978e-11
	এই	9.919357242490827e-12
	রূপবৈচিত্র্যে	6.01045799015334e-12
	বাংলাদেশ	4.5533772652676815e-12

Tree generation: Implement the maximum spanning tree (MST) from the directed graph. To implement MST, we use the Chu-Liu-Edmonds algorithm.

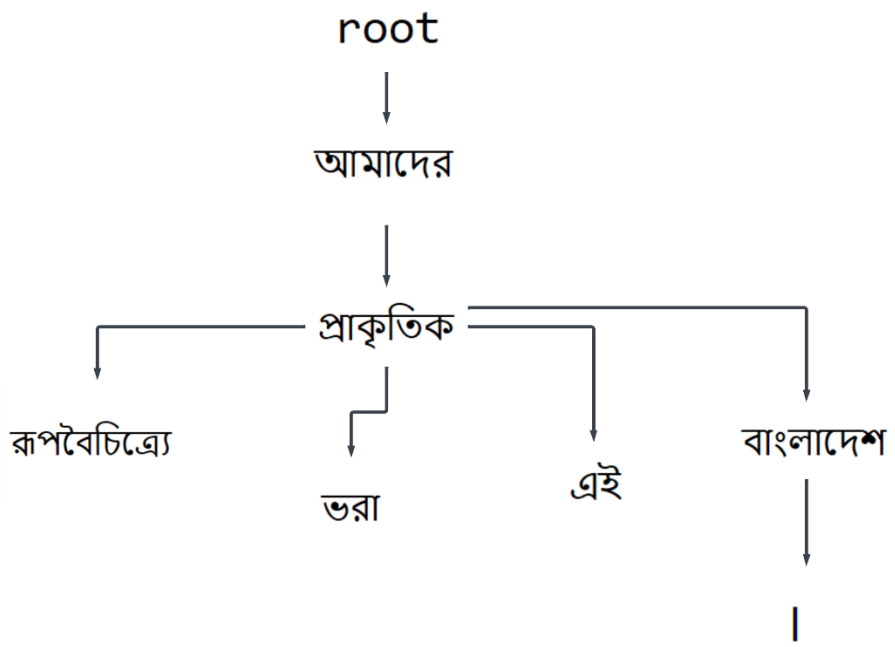


Figure 3.2: Graphical representation of dependencies in Bangla sentence

Flask application framework: Flask is a lightweight WSGI web application framework in Python. It is used for user input and generates output for users.

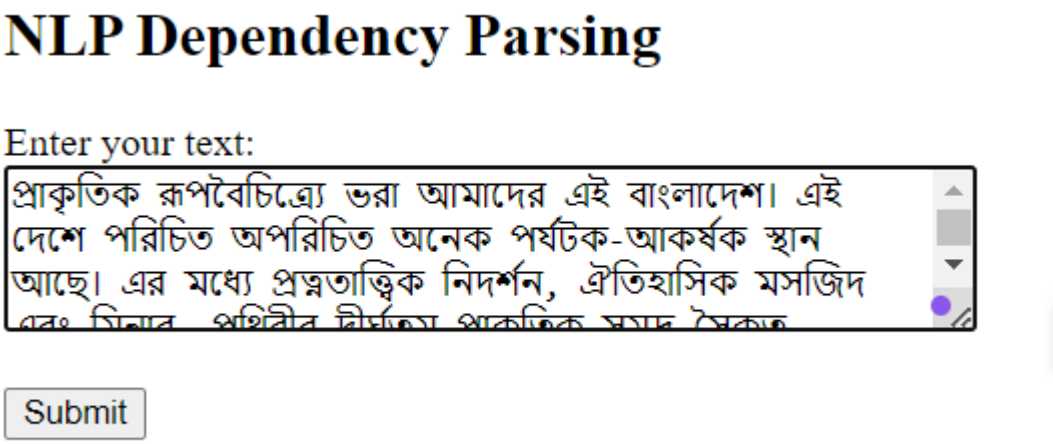
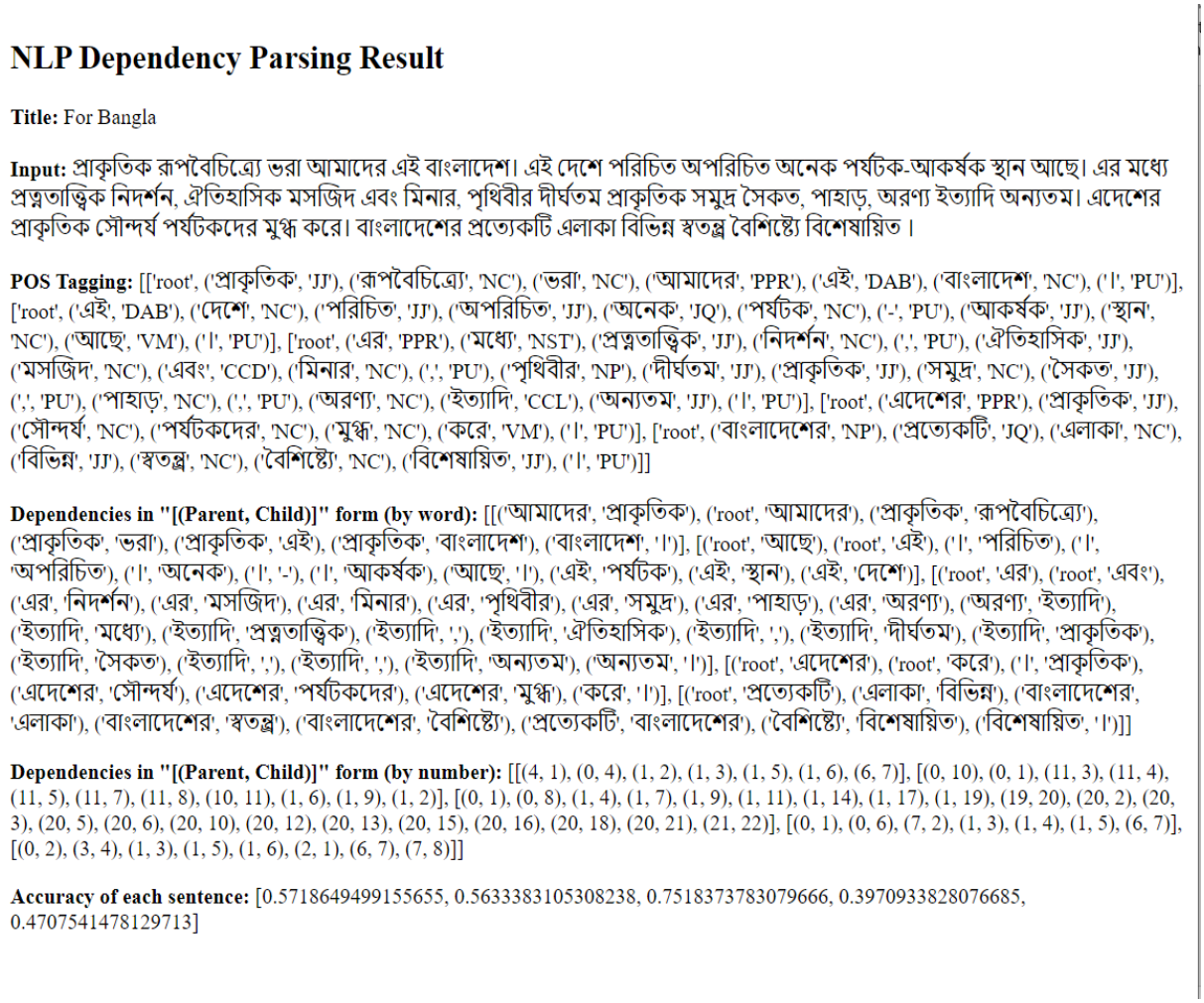


Figure 3.3: Input page



NLP Dependency Parsing Result

Title: For Bangla

Input: প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ। এই দেশে পরিচিত অপরিচিত অনেক পর্যটক-আকর্ষক স্থান আছে। এর মধ্যে প্রত্নতাত্ত্বিক নিদর্শন, ঐতিহাসিক মসজিদ এবং মিনার, পৃথিবীর দীর্ঘতম প্রাকৃতিক সমুদ্র সৈকত, পাহাড়, অরণ্য ইত্যাদি অন্যতম। এদেশের প্রাকৃতিক সৌন্দর্য পর্যটকদের মুগ্ধ করে। বাংলাদেশের প্রত্যেকটি এলাকা বিভিন্ন স্বতন্ত্র বৈশিষ্ট্যে বিশেষায়িত ।

POS Tagging: [['root', ('প্রাকৃতিক', 'JJ'), ('রূপবৈচিত্র্যে', 'NC'), ('ভরা', 'NC'), ('আমাদের', 'PPR'), ('এই', 'DAB'), ('বাংলাদেশ', 'NC'), ('।', 'PU')], ['root', ('এই', 'DAB'), ('দেশে', 'NC'), ('পরিচিত', 'JJ'), ('অপরিচিত', 'JJ'), ('অনেক', 'JQ'), ('পর্যটক', 'NC'), ('-', 'PU'), ('আকর্ষক', 'JJ'), ('স্থান', 'NC'), ('আছে', 'VM'), ('।', 'PU')], ['root', ('এর', 'PPR'), ('মধ্যে', 'NST'), ('প্রত্নতাত্ত্বিক', 'JJ'), ('নিদর্শন', 'NC'), (',', 'PU'), ('ঐতিহাসিক', 'JJ'), ('মসজিদ', 'NC'), ('এবং', 'CCD'), ('মিনার', 'NC'), (',', 'PU'), ('পৃথিবীর', 'NP'), ('দীর্ঘতম', 'JJ'), ('প্রাকৃতিক', 'JJ'), ('সমুদ্র', 'NC'), ('সৈকত', 'JJ'), (',', 'PU'), ('পাহাড়', 'NC'), (',', 'PU'), ('অরণ্য', 'NC'), ('ইত্যাদি', 'CCL'), ('অন্যতম', 'JJ'), ('।', 'PU')], ['root', ('এদেশের', 'PPR'), ('প্রাকৃতিক', 'JJ'), ('সৌন্দর্য', 'NC'), ('পর্যটকদের', 'NC'), ('মুগ্ধ', 'NC'), ('করে', 'VM'), ('।', 'PU')], ['root', ('বাংলাদেশের', 'NP'), ('প্রত্যেকটি', 'JQ'), ('এলাকা', 'NC'), ('বিভিন্ন', 'JJ'), ('স্বতন্ত্র', 'NC'), ('বৈশিষ্ট্যে', 'NC'), ('বিশেষায়িত', 'JJ'), ('।', 'PU')]]

Dependencies in "[(Parent, Child)]" form (by word): [[('আমাদের', 'প্রাকৃতিক'), ('root', 'আমাদের'), ('প্রাকৃতিক', 'রূপবৈচিত্র্যে'), ('প্রাকৃতিক', 'ভরা'), ('প্রাকৃতিক', 'এই'), ('প্রাকৃতিক', 'বাংলাদেশ'), ('বাংলাদেশ', '।')], [('root', 'আছে'), ('root', 'এই'), ('।', 'পরিচিত'), ('।', 'অপরিচিত'), ('।', 'অনেক'), ('।', '-'), ('।', 'আকর্ষক'), ('আছে', '।'), ('এই', 'পর্যটক'), ('এই', 'স্থান'), ('এই', 'দেশে')], [('root', 'এর'), ('root', 'এবং'), ('এর', 'নিদর্শন'), ('এর', 'মসজিদ'), ('এর', 'মিনার'), ('এর', 'পৃথিবীর'), ('এর', 'সমুদ্র'), ('এর', 'পাহাড়'), ('এর', 'অরণ্য'), ('অরণ্য', 'ইত্যাদি'), ('ইত্যাদি', 'মধ্যে'), ('ইত্যাদি', 'প্রত্নতাত্ত্বিক'), ('ইত্যাদি', ','), ('ইত্যাদি', 'ঐতিহাসিক'), ('ইত্যাদি', ','), ('ইত্যাদি', 'দীর্ঘতম'), ('ইত্যাদি', 'প্রাকৃতিক'), ('ইত্যাদি', 'সৈকত'), ('ইত্যাদি', ','), ('ইত্যাদি', ','), ('ইত্যাদি', 'অন্যতম'), ('অন্যতম', '।')], [('root', 'এদেশের'), ('root', 'করে'), ('।', 'প্রাকৃতিক'), ('এদেশের', 'সৌন্দর্য'), ('এদেশের', 'পর্যটকদের'), ('এদেশের', 'মুগ্ধ'), ('করে', '।')], [('root', 'প্রত্যেকটি'), ('এলাকা', 'বিভিন্ন'), ('বাংলাদেশের', 'এলাকা'), ('বাংলাদেশের', 'স্বতন্ত্র'), ('বাংলাদেশের', 'বৈশিষ্ট্যে'), ('প্রত্যেকটি', 'বাংলাদেশের'), ('বৈশিষ্ট্যে', 'বিশেষায়িত'), ('বিশেষায়িত', '।')]]

Dependencies in "[(Parent, Child)]" form (by number): [[(4, 1), (0, 4), (1, 2), (1, 3), (1, 5), (1, 6), (6, 7)], [(0, 10), (0, 1), (11, 3), (11, 4), (11, 5), (11, 7), (11, 8), (10, 11), (1, 6), (1, 9), (1, 2)], [(0, 1), (0, 8), (1, 4), (1, 7), (1, 9), (1, 11), (1, 14), (1, 17), (1, 19), (19, 20), (20, 2), (20, 3), (20, 5), (20, 6), (20, 10), (20, 12), (20, 13), (20, 15), (20, 16), (20, 18), (20, 21), (21, 22)], [(0, 1), (0, 6), (7, 2), (1, 3), (1, 4), (1, 5), (6, 7)], [(0, 2), (3, 4), (1, 3), (1, 5), (1, 6), (2, 1), (6, 7), (7, 8)]]

Accuracy of each sentence: [0.5718649499155655, 0.5633383105308238, 0.7518373783079666, 0.3970933828076685, 0.4707541478129713]

Figure 3.4: Output page

3.3 Proposed algorithm

The proposed algorithm is Edmond's algorithm which is used for finding the maximum spanning tree in a fully connected directed graph $G = (V, E)$ where V is the set of vertices (nodes) representing words in a sentence and E is the set of directed edges representing possible syntactic dependencies between words. This graph has a root node with only outgoing edges. Each edge in the graph is associated with a weight that represents the likelihood or cost of the dependency relation between the connected words. In this work, the weight is determined using transition and emission probability with some factor multiplication. The output of the Edmonds algorithm is the maximum spanning tree (MST) of the input graph, which represents the most likely syntactic structure or dependency parse tree of the sentence.

Table 3.2: Edmonds algorithm

Algorithm 1 Chu-Liu-Edmonds Algorithm
1: function Max_Spanning_Tree (G = (V, E), root, score)
2: F ← []
3: T ′ ← []
4: score′ ← []
5: for each v ∈ V do
6: bestInEdge ← $argmax_{e=(u,v)\in E}$ score[e]
7: F ← F ∪ bestInEdge
8: for each e = (u, v) ∈ E do
9: score′[e] ← score[e] − score[bestInEdge]
10: if T = (V, F) is a spanning tree then return T
11: else
12: C ← a cycle in F
13: G′ ← Contract (G, C)
14: T ′ ← MaxSpanningTree (G′, root, score′)
15: T ← Expand (T ′, C)
16: return T
17: function Contract (G, C) return contracted graph
18: function EXPAND (T, C) return expanded graph

In Table 3.1, the Edmonds algorithm is shown which is used to implement the maximum spanning tree.

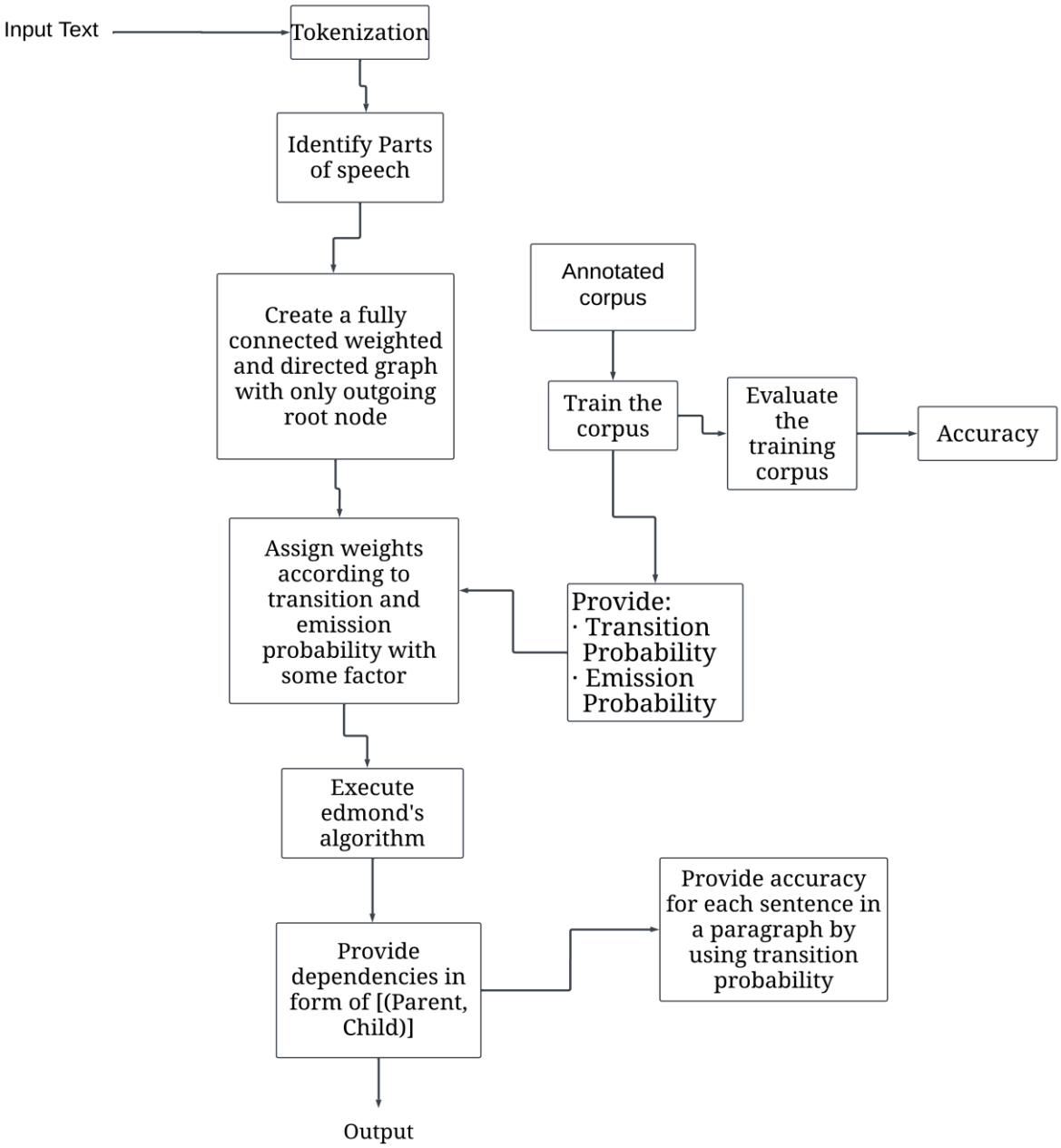


Figure 3.5: Proposed Methodology

CHAPTER IV

Implementation, Results and Discussions

This section contains experimental setup, evaluation metrics, dataset, implementation and other discussions.

4.1 Experimental Setup

The computer is installed Windows 10 OS. Python programming language is used to implement Edmond's algorithm, greedy decoding. Jupyter Notebook is used as IDE. Python module version is python 3.10.4. For the NLP task NLTK is used and for Bangla POS tagging BNLP is used. Flask which is a Python web framework used to create web applications. Both user input and output trees were shown using Flask. Flask 3.0.2 is used.

4.2 Evaluation

Evaluation is the measurement that is used to assess the performance or effectiveness of a system, model, algorithm or process. For the evaluation of this thesis work, the estimated accuracy of each generated tree is measured. For this computation, the transition probability is used. At first sort the transition probabilities in descending order of the same parent node that occurs in a current sentence of the paragraph. Then accuracy of each edge is summed up to get the accuracy of a sentence. If any transition probability is absent in the dataset that is held in the current sentence, this system adds a factor for the accuracy of that edge.

For English paragraph,

Input: The Padma is a major river in Bangladesh. It is the main distributary of the Ganges, flowing generally southeast for 356 kilometres to its confluence with the Meghna River near the Bay of Bengal. The city of Rajshahi is situated on the banks of the river.

Estimated accuracy: [0.5915291671951165, 0.6384512702148489, 0.27069518973579243]

For Bangla paragraph,

Input: প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ। এই দেশে পরিচিত অপরিচিত অনেক পর্যটক-আকর্ষক স্থান আছে। এর মধ্যে প্রত্নতাত্ত্বিক নিদর্শন, ঐতিহাসিক মসজিদ এবং মিনার, পৃথিবীর দীর্ঘতম প্রাকৃতিক সমুদ্র সৈকত, পাহাড়, অরণ্য ইত্যাদি অন্যতম। এদেশের প্রাকৃতিক সৌন্দর্য পর্যটকদের মুগ্ধ করে। বাংলাদেশের প্রত্যেকটি এলাকা বিভিন্ন স্বতন্ত্র বৈশিষ্ট্যে বিশেষায়িত ।

Estimated accuracy from method 1: [0.5718649499155655, 0.5633383105308238, 0.7518373783079666, 0.3970933828076685, 0.4707541478129713]

In method 2, estimated accuracy for Bangla is [0.7142857142857143, 0.9090909090909091, 0.4090909090909091, 1.0, 0.75]. In this method, for which edge the transition probability is absent counted as the wrong edge. Then take all the correct edges and divided by number of tokens in sentence.

In figure-3.3, ('আমাদের', 'প্রাকৃতিক'), ('root', 'আমাদের'), ('প্রাকৃতিক', 'রূপবৈচিত্র্যে'), ('বাংলাদেশ', '।') these 4 edges are correct and ('প্রাকৃতিক', 'ভরা'), ('প্রাকৃতিক', 'এই'), ('প্রাকৃতিক', 'বাংলাদেশ') these 3 edges are incorrect. So method 1 (0.5718649499155655) accuracy calculation is more feasible than method 2 (0.7142857142857143)

4.3 Dataset

In this thesis work, the annotated corpus is used. For English, there are one hundred unique sentences with 42 unique states. States are 'IN', 'DT', 'NN', ',', 'NNS', 'VBD', '``', 'PRP', 'VBP', 'PDT', '.', '"', 'JJ', 'NNP', 'VBZ', 'WP', 'VBN', 'RB', 'CC', 'VBG', 'RP', 'EX', 'MD', 'VB', 'RBR', 'TO', 'CD', 'PRP$', ':', 'NNPS', 'JJR', 'POS', '-LRB-', '-RRB-', 'WDT', 'WRB', 'RBS', 'JJS', '$', 'WP$', 'SYM', 'LS'.

For Bangla, there are fifty unique sentences with 33 unique states. States are 'RDF', 'NC', 'JJ', 'CCD', 'NV', 'PU', 'NP', 'VM', 'JQ', 'PP', 'PPR', 'VAUX', 'NST', 'PRL', 'AMN', 'CCL', 'CSB', 'DAB', 'CX', 'EX', 'MD', 'VB', 'NNS', 'VBG', 'WP', 'PRP', 'VBD', '.', '"', 'ALC', 'PWH', 'PRF', 'NNP'.

Accuracy of training corpus is shown in Table 4.1.

Table 4.1: Accuracy of training dataset

Accuracy Of Dataset	Bangla	English
	0.7692307692307693	0.7027027027027027

4.4 Implementation and Results

In this thesis work, users can enter a paragraph as input. The input will be separated into each sentence. Each word of a sentence is tokenized and tagged the POS. This work will generate a fully weighted-directed connected graph with an extra “root” node which has only outgoing edges to each word of that sentence. For assigning weights of directed edges of the graph, this system uses transition and emission probability with some factor. This transition and emission probabilities are acquired from training corpus. Then the graph passes to the Edmonds algorithm that provides the maximum spanning tree of the graph where the sum of weights of all edges is maximum than other spanning trees of the graph. The resultant tree indicates the dependencies of the words in sentence. The output segment shows the dependencies as a parent-child relationship in list i.e. [(Parent, Child)].

For English paragraph,

Input text: “The Padma is a major river in Bangladesh. It is the main distributary of the Ganges, flowing generally southeast for 356 kilometres to its confluence with the Meghna River near the Bay of Bengal. The city of Rajshahi is situated on the banks of the river.”

Output dependencies: [[('root', 'is'), ('is', 'a'), ('a', 'The'), ('a', 'Padma'), ('a', 'major'), ('major', 'river'), ('river', 'in'), ('in', 'Bangladesh'), ('Bangladesh', '.')], [('root', 'is'), ('root', 'southeast'), ('southeast', 'for'), ('for', 'It'), ('for', 'the'), ('for', 'main'), ('for', 'distributary'), ('for', 'of'), ('for', 'the'), ('for', 'Ganges'), ('for', ','), ('for', 'flowing'), ('for', 'generally'), ('for', '356'), ('356', 'kilometres'), ('kilometres', 'to'), ('to', 'its'), ('its', 'confluence'), ('confluence', 'with'), ('with', 'the'), ('the', 'Meghna'), ('Meghna', 'River'), ('River', 'near'), ('near', 'the'), ('the', 'Bay'), ('Bay', 'of'), ('of', 'Bengal'), ('Bengal', '.')], [('root', 'is'), ('root', 'situated'), ('situated', 'on'), ('on', 'The'), ('on', 'city'), ('on', 'of'), ('on', 'Rajshahi'), ('on', 'the'), ('the', 'banks'), ('banks', 'of'), ('of', 'the'), ('the', 'river'), ('river', '.')]]

Table 4.2: Dependencies of the sentence "*The Padma is a major river in Bangladesh.*"

Token	Head	Children
The	A	[]
Padma	A	[]
Is	Root	[a]
A	Is	[the, Padma, major]
Major	A	[River]
River	Major	[in]
in	River	[Bangladesh]
Bangladesh	In	[.]
.	Bangladesh	[]

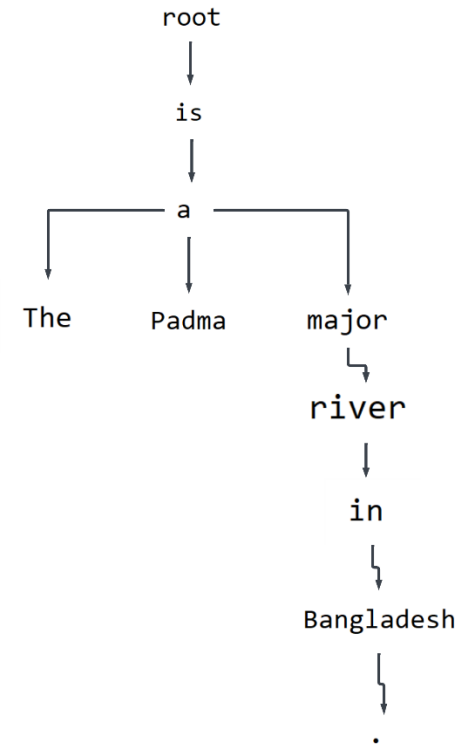


Figure 4.1: Graphical representation of dependencies in English sentence

For Bangla paragraph,

Input text: “প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ। এই দেশে পরিচিত অপরিচিত অনেক পর্যটক-আকর্ষক স্থান আছে। এর মধ্যে প্রত্নতাত্ত্বিক নিদর্শন, ঐতিহাসিক মসজিদ এবং মিনার, পৃথিবীর দীর্ঘতম প্রাকৃতিক সমুদ্র সৈকত, পাহাড়, অরণ্য ইত্যাদি অন্যতম। এদেশের প্রাকৃতিক সৌন্দর্য পর্যটকদের মুগ্ধ করে। বাংলাদেশের প্রত্যেকটি এলাকা বিভিন্ন স্বতন্ত্র বৈশিষ্ট্যে বিশেষায়িত।”

Output dependencies: [[('আমাদের', 'প্রাকৃতিক'), ('root', 'আমাদের'), ('প্রাকৃতিক', 'রূপবৈচিত্র্যে'), ('প্রাকৃতিক', 'ভরা'), ('প্রাকৃতিক', 'এই'), ('প্রাকৃতিক', 'বাংলাদেশ'), ('বাংলাদেশ', '।')], [('root', 'আছে'), ('root', 'এই'), ('।', 'পরিচিত'), ('।', 'অপরিচিত'), ('।', 'অনেক'), ('।', '-'), ('।', 'আকর্ষক'), ('আছে', '।'), ('এই', 'পর্যটক'), ('এই', 'স্থান'), ('এই', 'দেশে')], [('root', 'এর'), ('root', 'এবং'), ('এর', 'নিদর্শন'), ('এর', 'মসজিদ'), ('এর', 'মিনার'), ('এর', 'পৃথিবীর'), ('এর', 'সমুদ্র'), ('এর', 'পাহাড়'), ('এর', 'অরণ্য'), ('অরণ্য', 'ইত্যাদি'), ('ইত্যাদি', 'মধ্যে'), ('ইত্যাদি', 'প্রত্নতাত্ত্বিক'), ('ইত্যাদি', ','), ('ইত্যাদি', 'ঐতিহাসিক'), ('ইত্যাদি', ','), ('ইত্যাদি', 'দীর্ঘতম'), ('ইত্যাদি', 'প্রাকৃতিক'), ('ইত্যাদি', 'সৈকত'), ('ইত্যাদি', ','), ('ইত্যাদি', ','), ('ইত্যাদি', 'অন্যতম'), ('অন্যতম', '।')], [('root', 'এদেশের'), ('root', 'করে'), ('।', 'প্রাকৃতিক'), ('এদেশের', 'সৌন্দর্য'), ('এদেশের', 'পর্যটকদের'), ('এদেশের', 'মুগ্ধ'), ('করে', '।')], [('root', 'প্রত্যেকটি'), ('এলাকা', 'বিভিন্ন'), ('বাংলাদেশের', 'এলাকা'), ('বাংলাদেশের', 'স্বতন্ত্র'), ('বাংলাদেশের', 'বৈশিষ্ট্যে'), ('প্রত্যেকটি', 'বাংলাদেশের'), ('বৈশিষ্ট্যে', 'বিশেষায়িত'), ('বিশেষায়িত', '।')]]

Table 4.3: Dependencies of the sentence “*প্রাকৃতিক রূপবৈচিত্র্যে ভরা আমাদের এই বাংলাদেশ।*”

Token	Head	Children
প্রাকৃতিক	আমাদের	[রূপবৈচিত্র্যে, ভরা, এই, বাংলাদেশ]
রূপবৈচিত্র্যে	প্রাকৃতিক	[]
ভরা	প্রাকৃতিক	[]
আমাদের	root	[প্রাকৃতিক]
এই	প্রাকৃতিক	[]

বাংলাদেশ	প্রাকৃতিক	[।]
।	বাংলাদেশ	[]

4.5 Objective Achieved

In this paper, the proposed method represents the implementation technique of dependency parsing for Bangla text from a directed graph of input. In this method, finding out the POS tag of input words. Also successfully generating a dependency parsing that tree which is the maximum spanning tree for Bangla. This work also provides an estimated accuracy for each sentence in a paragraph. So the objective is successfully achieved.

4.6 Financial Analyses and Budget

Creating a detailed financial analysis and budget for a dependency parsing project for Bangla text involves estimating costs associated with various aspects of the project. Costs will vary based on the project's scope, goals, team size, resources, and other factors. Scope of costs includes software installment, creating annotated corpus, resources and other paid contents. Also expenses for testing and evaluating the dependency parsing models on representative Bangla text samples.

Table 4.4: Specifications of budgeting

Scope of expenses	**Estimated Cost (৳)**
Software installments	৳7,000
Create annotated corpus	৳8,000
Resources	৳3,000
Paid contents	৳10,000
Total estimated budget	৳28,000

Chapter V

Societal, Health, Environment, Safety, Ethical, Legal and Cultural Issues

5.1 Intellectual Property Considerations

This thesis work builds upon existing knowledge. It should be protected by copyright laws, patentable, open-source, and attribution. Collaborations with other researchers should be established, and a commercialization strategy should be developed. Agreements for technology transfer and licensing can support the management of commercialization and intellectual property rights.

5.2 Ethical Considerations

When parsing is used in educational settings, ethical issues must be taken into account. Without impairing linguistic diversity or students' linguistic comprehension, parsing should improve educational experiences. Parsing algorithms should be created to avoid harming marginalized communities or maintaining discrimination against them. Ethics related to intellectual property and fair use must be taken into account whenever dependency parsing is used on copyrighted content. Parsing applied to analyze or evaluate language skills without taking into account any drawbacks or restrictions raises ethical questions. Processing sensitive or private text in Bangla may be required for dependency parsing. It becomes essential to safeguard individual privacy and adhere to data protection laws.

5.3 Safety Considerations

To protect the safety of researchers, participants, and the larger community, great attention to safety is needed in the dependency parsing research for the Bangla language. Data security, ethical data use, bias and fairness, algorithmic safety, accessibility, environmental

effect, collaborative safety, and community involvement are important safety factors. Data security is the ethical use of data, which includes informed authorization, privacy and confidentiality preservation, and compliance with regulations governing research involving individuals. To ensure fairness, varied datasets must be used, potential biases in data or algorithms must be addressed, and social impacts must be taken into account. Dependency parsing models' robustness and reliability are evaluated using algorithmic safety to make sure they operate correctly and securely in real-world scenarios. Accessibility guarantees that study results are understandable to a wide range of people, including those with special needs or disabilities. By implementing sustainable methods, minimizing on energy use.

5.4 Legal Considerations

This work must take legal considerations to ensure relevant laws, regulations, and ethical standards.

- **Data Protection and Privacy Laws:** It should Ensure compliance with data protection and privacy laws When collecting, storing, or processing data, obtain informed consent from participants and implement appropriate security measures to safeguard data privacy.
- **Intellectual Property Rights**: While using pre-existing datasets, language resources and software tools, respect copyrights, patents, trademarks, and trade secrets. While collecting any required consents, licenses, or agreements from the owners of the rights, and properly acknowledge the sources according to relevant copyright laws.
- **Data Sovereignty and Jurisdiction:** It should take legal concerns and data sovereignty into account when processing or storing data on cloud-based or outside platforms. Recognize the rules and laws on data residency, international data transfers, and government access to data across borders.

5.5 Impact of the Project on Societal, Health, and Cultural Issues

The impact of a project on societal, health, and cultural issues can be significant. Here is the impact of these issues:

- **Societal Impact:** This thesis work aims to increase language accessibility by developing dependency parsers for Bangla text. It enables Bangla speakers with

limited English proficiency to access online information and services. Additionally, it enhances educational materials and language-learning software, advances literacy and protects the Bangla language and culture.

- **Health Impact:** Dependency parsing tools can improve health information access in Bangla. It promotes preventive healthcare practices. Improved language processing can support telemedicine platforms and health chatbots, enhancing patient communication. Natural language processing techniques can also be used for public health surveillance, monitoring trends and detecting disease outbreaks.
- **Cultural Impact:** The work aims to preserve and promote the Bangla language and culture through advanced language technologies, including annotated corpora and linguistic tools. It also promotes cultural expression through natural language processing tools, facilitating the creation and dissemination of Bangla-language content online. This promotes digital inclusion.

5.6 Impact of Project on the Environment and Sustainability

The thesis work aims to improve energy efficiency, promote digital accessibility, reduce paper consumption, and support sustainable development by developing advanced natural language processing models for Bangla text. This will enable marginalized communities to access technology and information in their native language, fostering socio-economic development and reducing inequalities. Additionally, the work can help reduce paper consumption and environmental impacts by facilitating digital communication and analyzing Bangla-language data for environmental monitoring and conservation.

Chapter VI

Addressing Complex Engineering Problems and Activities

This chapter discusses about the complex engineering problems and activities associated with this thesis work.

6.1 Complex engineering problems associated with the current thesis

Several complex engineering problems occurred in this work due to the unique characteristics of the language. Some of these challenges are:

- **Lack of annotated data**: this work requires large amounts of annotated data for training, but for the Bangla language, such annotated datasets may be limited. Building high-quality annotated datasets for Bangla poses a significant challenge.
- **Morphological complexity**: Bangla is a rich morphological language with a variety of prefixes, suffixes, and compound words. Parsing Bangla text requires handling complex morphological variations.
- **Syntactic ambiguity**: Bangla sentences often contain syntactic ambiguities where the same sequence of words can have multiple valid dependency structures. Resolving such ambiguities requires advanced parsing techniques and linguistic knowledge that makes dependency parsing more challenging.
- **Out-of-vocabulary words**: This model encounters out-of-vocabulary words. Handling such words and effectively generalizing to unseen vocabulary items is crucial for robust parsing performance.
- **Word segmentation**: This work needs to accurately tokenize Bangla text into individual tokens and identify POS before parsing, which can be challenging due to the presence of compound words and agglutinative morphology.
- **Resource constraints**: Limited dataset, personnel, and other resources constrained the scope and scale of the research project.

6.2 Complex engineering activities associated with the current thesis

This thesis work involves several complex engineering activities. Some of these activities include:

- **Data Collection and Annotation**: Acquiring large amounts of annotated Bangla text data is crucial for training dependency parsing models. This involves collecting diverse text sources, annotating them and ensuring high annotation quality.
- **Preprocessing and Tokenization**: Bangla text often lacks explicit word boundaries which makes tokenization challenging. Robust tokenization method is used that accurately segment Bangla text into individual tokens or words, considering morphological variations, compound words, and other linguistic phenomena.
- **Feature Engineering**: It includes effective features to represent Bangla linguistic properties that are essential for generating accurate parsing models. It may extract various features from the input text, such as part-of-speech tags to improve parsing accuracy.
- **Model Selection and Architecture Design**: An appropriate parsing algorithm is chosen. It may explore a graph-based parsing model.
- **Apply Edmond's Algorithm:** Edmonds algorithm is applied to the directed graph to find the maximum spanning tree. The maximum spanning tree represents the most probable syntactic structure of the sentence, where each word is connected to exactly one parent based on the predefined dependency relations.
- **Evaluation and Fine-tuning:** The performance evaluation of the dependency parsing system on annotated test data to assess its accuracy, coverage, and robustness. Fine-tune the parsing model and algorithm parameters based on the transition and emission probabilities and iteratively improve the parsing system.
- **Error Analysis**: Evaluating parsing models on annotated test datasets is essential for assessing their performance. It conducts comprehensive error analysis to identify errors.

CHAPTER VII

Conclusions

7.1 Summary

In conclusion, the intricate details of adapting dependency parsing techniques to the distinctive linguistic features of the Bangla language have been examined in the thesis research on "Dependency Parsing for Bangla text". In addition to revealing word relationships and improving our comprehension of the grammar and syntax of the Bangla text, the study has made significant contributions to the syntactic analysis of the Bangla language. The results of this study provide insight into the importance of dependency parsing in Bangla and its wider implications for linguistics and natural language processing. The development of annotated corpora and linguistic resources intended specifically for Bangla dependency parsing has been aided by the thesis research. These resources provide a foundation for future research endeavors and facilitate the advancement of Bangla language technology.

The research's outcomes place a strong emphasis on the usefulness of dependency parsing in Bangla. The proper study of syntactic links improves the performance of several natural language processing tasks, including sentiment analysis and machine translation.

7.2 Limitations

Bangla has fewer linguistic resources than languages like English, which can make training and assessing dependency parsing models more difficult. There aren't many comprehensive treebanks, part-of-speech taggers, or annotated corpora for Bangla. Designing dependency parsing models that faithfully represent the relationships between words in Bangla is difficult due to its rich morphological features and flexible word order. Parsing errors can result from ambiguities in Bangla sentences for polysemy, homonymy, and idiomatic idioms. Due to linguistic eccentricity and a lack of qualified annotators, manually annotating

Bangla phrases for dependency parsing might be difficult. It takes a significant amount of memory and computing power to create accurate dependency parsing models. It can take a lot of time to train models and conduct reviews.

7.3 Recommendations and Future Works

- This work can be expanded by detecting appropriate parts of speech of words in a sentence by considering more linguistic features
- Assign more accurate weight by considering more morphological information
- By using large and enriched corpus, solving ambiguity and accuracy of output with the help of semantics of the sentence
- Use this mechanism to enrich knowledge-base of computer
- Generate automated human-like Bangla text

APPENDIX

Appendix 1: Bangla dataset

REFERENCES

References

[1] Dhar, Arnab, et al. "A hybrid dependency parser for Bangla." Proceedings of the 10th Workshop on Asian Language Resources. 2012.

[2] Das, Arjun, Arabinda Shee, and Utpal Garain. "Evaluation of two bengali dependency parsers." Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages. 2012.

[3] Chatterji, Sanjay, et al. "Grammar driven rules for hybrid bengali dependency parsing." ICON09 NLP TOOLS CONTEST: INDIAN LANGUAGE DEPENDENCY PARSING (2009): 38.

[4] McDonald, Ryan, et al. "Non-projective dependency parsing using spanning tree algorithms." Proceedings of human language technology conference and conference on empirical methods in natural language processing. 2005.

[5] Ghosh, Urmi, Dipti Misra Sharma, and Simran Khanuja. "Dependency parser for bengali-english code-mixed data enhanced with a synthetic treebank." Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019). 2019.

[6] Nivre, Joakim. "Parsing indian languages with maltparser." Proceedings of the ICON09 NLP Tools Contest: Indian Language Dependency Parsing (2009): 12-18.

[7] Mannem, Prashanth. "Bidirectional dependency parser for hindi, telugu and bangla." ICON09 NLP TOOLS CONTEST: INDIAN LANGUAGE DEPENDENCY PARSING (2009): 49.

[8] Husain, Samar. "Dependency parsers for indian languages." Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing (2009).

[9] Kosaraju, Prudhvi, et al. "Experiments on indian language dependency parsing." Proceedings of the ICON10 NLP Tools Contest: Indian Language Dependency Parsing (2010): 40-45.

[10] Husain, Samar, et al. "The ICON-2010 tools contest on Indian language dependency parsing." Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing, ICON 10 (2010): 1-8.

[11] Das, Amitava, and Sivaji Bandyopadhyay. "Phrase-level Polarity Identification for Bangla." Int. J. Comput. Linguistics Appl. 1.1-2 (2010): 169-182.

[12] Khatun, Ayesha, and Mohammed Moshiul Hoque. "Statistical parsing of Bangla sentences by CYK algorithm." 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2017.

[13] Abhilash, Aswarth, and Prashanth Mannem. "Bidirectional dependency parser for indian languages." Proceedings of the ICON10 NLP Tools Contest: Indian Language Dependency Parsing (2010).