Computer Engineering Department

Government College of Engineering & Research Avasari Khurd

Savitribai Phule Pune University

Final Year of Computer Engineering

2022-23 Semester I

Project Synopsis

- Project Group ID: **18.**
- Title of the Project: **Syntactical Parsing Of Marathi Text Based On CYK Algorithm**
- Team Members :

Sr.	Roll No	Name	Email	Mobile	Sign
1	19121056	Pranav Pradip Ranaware	pranav.ranaware122019@gcoeara.ac.in	7517490959	
2	20221082	Vaishnavi Nilkamal Swami	vaishnavi.swami122020@gcoeara.ac.in	8208492240	
3	19121017	Kalyani Mahadev kadam	kalyani.kadam122019@gcoeara.ac.in	7020984647	

Date:

Prof. S.G. Farkade
Guide

Dr. S.A. Thorat
Project Coordinator

Abstract

Marathi is an Indo-Aryan Language and forms the official language of the state of Maharashtra. It is ranked as the 4th most spoken language in India and the 15th most spoken language in the world. When Computational linguistics is concerned, writing grammar production for a language is a bit difficult because of different gender and number forms. This is an effort to write context-free grammar for Marathi sentences. CFGs are very much suitable for expressing natural languages as well as programming languages and hence form a major part of the field of natural language processing and pattern recognition.

Computational linguistics is one of the attractive research topics in the Natural Language Processing (NLP) and Artificial Intelligence domains present a particular syntactical parsing technique for Marathi texts which is one of the Indian languages. Cocke–Younger–Kasami (CYK) parsing technique has been adopted to parse Marathi sentences and identify their grammatical structure. Currently, very less NLP tools are available to parse several Indian languages. Hence, an effort has been made by us to efficiently parse the structure of the complex sentences in Marathi text using the CYK algorithm. It is a bottom-up dynamic programming approach that functions only with the grammar in Chomsky's normal form (CNF).

Keywords:

Natural language processing, CYK algorithm, Syntactical parsing, Chomsky normal form, Production rules, Rule-based grammar.

Background and Motivation

Background:

India is a multilingual country. Indian constitution lists 22 languages, referred to as scheduled languages. These languages are given status, recognition and official encouragement. Of the entire population, barely 10% Indians use English to transact and most prefer regional languages, which have evolved over centuries. As there is diversity in languages, language processing applications are a boon to the people for their day-to-day transactions. However, understanding and generation of these natural languages i.e. processing of these natural languages by machine is complex. Therefore, we review the work carried out by researchers on various techniques developed for processing Indian Regional Languages.

Motivation:

A very little attempt has been made to develop a syntactical parser on Indian languages including Marathi. Marathi is a spoken as well as a written language in Maharashtra state and it is basically Used all over Maharashtra. We made an effort to develop a systematic syntax analyzer to parse all types of Marathi texts, and the same has been presented in the paper. Python programming language has been used and the NLTK tree module has also been imported to serve the purpose. NLTK tree module has been adopted to display the parser tree for every valid Marathi sentence. Syntax analysis is the process of checking the grammatical correctness of a sentence and identifying the relationship between the words in an input sentence. It is very simple to write the production rules or grammar for fixed order spoke languages compared to other free order languages.

Problem Definition and Objectives

Problem Definition:

- Currently, very less NLP tools are available to parse several Indian languages. Hence, an effort has been made by us to efficiently parse the structure of the complex sentences in Marathi text using the CYK algorithm.
- As already parser are implemented in English, Hindi, and Kannad we are overcoming this in Marathi.

Objectives:

- To create a parsing technique based on Marathi.
- To display the parser tree for every valid Marathi sentence NLTK tree module is used .
- To develop efficient dynamic programming technique and suitable for analyzing complex sentences CYK Algorithm is used.
- To separates a series of text based on Marathi grammar rules NLP is used .

Literature Survey

"Syntactic Parsing in Kannada Text/Natural Language Processing". This paper presents a particular syntactical parsing technique on Kannada texts which is one of the South Indian languages. Cocke–Younger–Kasami (CYK) parsing technique has been adopted to parse Kannada sentences and identify their grammatical structure. It is a bottom-up dynamic programming approach which functions only with the grammar in Chomsky normal form (CNF) [1].

"A Comprehensive Survey On Indian Regional Language". This paper presents there are different native languages existing in various parts of the world,each with its own alphabets ,sign,grammer. It is comparitvely easy for computers to process the data as represented in the English language through standard ASCII codes than in other in natural languages [2].

"Parsing for Natural Language in Odia: a novel study". In this paper a simple parsing for Odia Language using Context-Free Grammar (CFG) is shown. Parsing is a technique used for building a sentence automatically in the phrases of grammar as well as in lexicon using syntactic analysis. It includes semantic analysis and syntactic analysis that basically attention on parsing for Odia Language based on Context Free Grammar along with top down approach. All the things are being represented as simple tree primarily formalism in context free grammar. [3].

Methodology

An input sentence is processed by a parser according to the productions of a grammar and builds one or more constituent structures that satisfy the grammar. A parser is a procedural interpretation of the grammar. It permits a grammar to be evaluated against a collection of test sentences, helping linguists to find mistakes in their grammatical analysis.

The CYK parser model proposed in this paper reads CFG grammar and converts into CNF format if required. The Maharshtrian dataset which contains different kinds like imperative, declarative, assertive, compound sentences, and interrogative sentences has been given as input to the parser model. Each word is treated as a terminal symbol. These terminals are matched with the production rules in the CNF grammar and try to fetch the corresponding productions by RHS.

This continues to identify the respective LHS of the productions recursively through the bottom-up approach until it gets the start symbol (S). Finally, if the start symbol of the grammar is not obtained, then the input sentence given is not accepted by the grammar. If the start symbol is obtained, it outputs with the NLTK parse tree. For some sentences, it may create more than one parse tree when parsing technique gets the same start symbol through different production rules. The simple parsers suffer from limitations in both completeness and efficiency. In order to overcome these, dynamic programming technique has been applied here to resolve the parsing problem.

Dynamic parsing algorithm accumulates the intermediate results and reuses them when relevant, achieving efficiency gains. This technique can be applied to syntactical parsing, allowing us to store partial solutions during the parsing task and then look them up as necessary in order to efficiently arrive at a complete solution. This approach to parsing is known as chart parsing. A typical sentence has been taken as an example to present the proposed technique.

It is very difficult to design a grammar for the entire Marathi language. For the sake of simplicity, this paper considers writing very simple grammatical productions with CFG that will generate a subset of Marathi sentences .

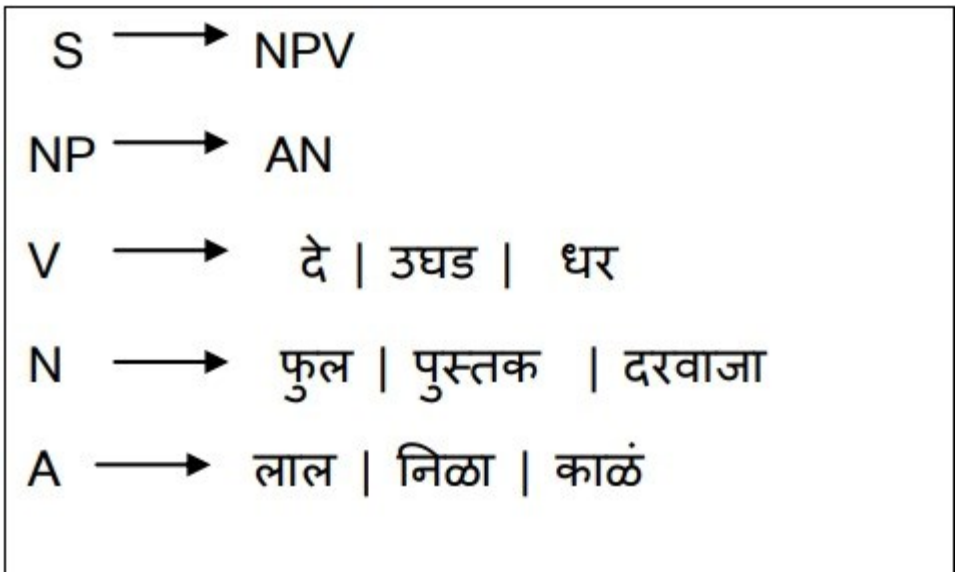

Fig: Marathi Grammatical Production

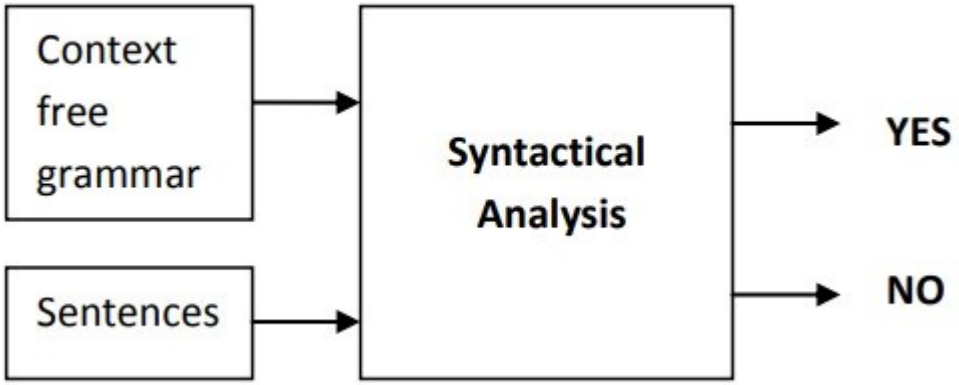

Fig: Syntactical analysis

Parsing or syntactic analysis is the process of analyzing a string of symbols, according to the rules of a formal grammar. It is the task of analyzing the grammatical structure of natural language . A parser forms separate units like subject, verb, and object and determines the relations between these units.

The linguistic rules provided are mostly in context-free grammar form. This form basically provides a simple and mathematically precise mechanism for describing the methods by which phrases in some natural language are built from smaller blocks . It also gives the basic recursive structure of sentences, the way in which clauses nest inside other clauses, and the way in which lists of adjectives and adverbs are swallowed by nouns and verbs. Context-free grammars are simple enough to allow the construction of efficient parsing algorithms.

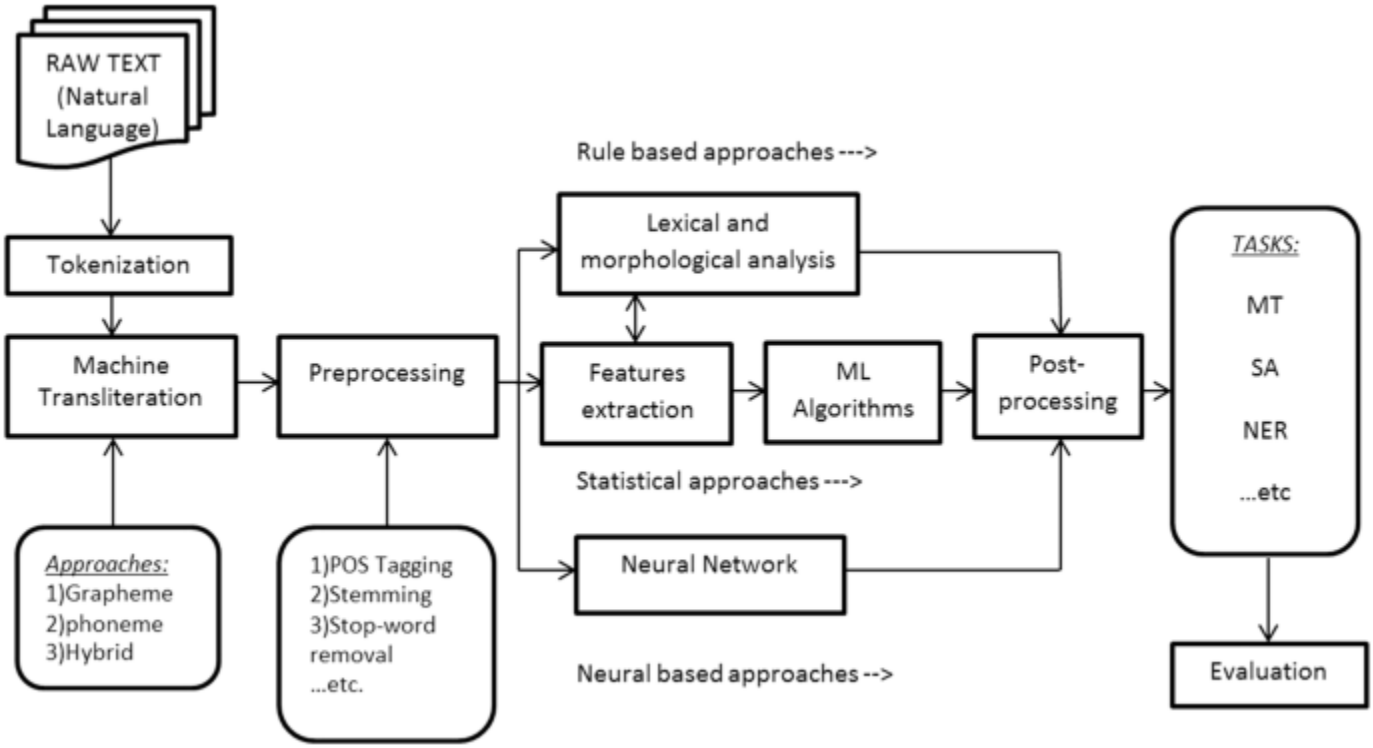

Fig: Generic block diagram

Software and Hardware Requirements

Software Requirement

- Python
- NLP

Hardware Requirement

- Windows 7 or higher
- I3 processor system or higher
- 4 GB RAM or higher
- 100 GB ROM or higher

References

[1] M. Rajani Shree, "Syntactic Parsing in Kannada Text/Natural Language Processing", Third International Conference on Intellegent computing information and control systems ICICCS, March 2022.

[2] Harish BS, Rangan RK , "A comprehensive survey on Indian regional language processing" , Proceedings of the ınternational conference [ACCTA-2010] on Special Issue of IJCCT, June 2020.

[3] Das BR, Singh D, Bhoi PC, Priyadarshini P "Parsing for Natural Language in Odia: a novel study" , International Conference on Computer Science, Engineering and Applications (ICCSEA), March 2020.

[4] Teodora Dordevic and Suzana Stojkovic , "Different Approaches in Serbian Language Parsing Using Context free Grammers" , Proceedings of the second ınternational conference on data science, E-learning and ınformation systems, Sep2020.

[5] Denis Eka Cahyani, Langlang Gumilar, Ajie pangestu "Indonesian Parsing Using Probabilistic Context-Free Grammer and CYK ", Third International Seminar On Research Of Technology and Intelligent System, Dec 2020.