**SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE - 411043**
**Department of Computer Engineering**
**S.No.-27, Pune Satara Road, Dhankawadi, Pune-411043**

Laboratory Practice-VI (AY 2022-23)

Batch- Q3 Sem-VIII

Date-12-05-2023

| Names | Roll No |
|---|---|
| Rutuja Banginwar | 42303 |
| Vrushaket Chaudhari | 41305 |
| | |

Lab Teacher Name: Prof. P. S. Vidap

Title of project :- POS Taggers For Indian Languages

## Introduction

Hindi POS tagging is the process of labelling each word in a Hindi sentence with its corresponding part-of-speech (POS) tag. In other words, it involves identifying the grammatical function of each word in a sentence, such as noun, verb, adjective, adverb, etc. POS tagging is a fundamental task in natural language processing (NLP) and is used in many downstream NLP applications, such as machine translation, sentiment analysis, and information retrieval.

Hindi is an Indo-Aryan language spoken in India, Nepal, and other parts of the world. It is the fourth most spoken language in the world, with over 600 million speakers. Hindi has a rich morphological system with a large number of inflections, which makes POS tagging a challenging task.

There are various methods and tools available for Hindi POS tagging, including rule-based approaches, dictionary-based approaches, and machine learning-based approaches. Rule-based approaches involve the use of hand-crafted rules to identify the POS tags of words in a sentence. Dictionary-based approaches use a pre-defined dictionary of words and their corresponding POS tags to label new sentences. Machine learning-based approaches, on the other hand, involve the use of machine learning algorithms, such as hidden Markov models

(HMMs) and conditional random fields (CRFs), to learn the patterns in the training data and predict the POS tags of new sentences.

In recent years, deep learning-based approaches, such as neural networks, have also shown promising results for Hindi POS tagging. These approaches can learn complex features and patterns in the data and achieve state-of-the-art performance on Hindi POS tagging tasks.

Overall, Hindi POS tagging is an important task in NLP, and the development of accurate and efficient POS taggers for Hindi can greatly benefit various NLP applications.

## a) .Motivation

There are several motivations for undertaking a project on Hindi POS tagging:

1)Improve accuracy of NLP applications: Hindi is one of the most spoken languages in the world, and accurate Hindi POS tagging can greatly improve the accuracy of various NLP applications, such as machine translation, sentiment analysis, and information retrieval.

2)Address the challenges of Hindi morphology: Hindi has a rich morphological system with a large number of inflections, which makes POS tagging a challenging task. Developing a robust and accurate POS tagger for Hindi can help address these challenges and improve the overall quality of NLP applications for Hindi.

3)Develop better language resources: Building a Hindi POS tagger requires a large amount of annotated data, which can serve as a valuable resource for other NLP tasks. This project can contribute to the development of better language resources for Hindi NLP research.

4)Enable new research directions: Accurate Hindi POS tagging can open up new research directions in NLP, such as studying the syntactic and semantic structures of Hindi text and developing new algorithms and models for Hindi NLP.

Overall, the development of an accurate and efficient Hindi POS tagger can greatly benefit the field of NLP and contribute to the development of better NLP applications for Hindi.

## b). Objective/ Purpose

The purpose of this project on Hindi POS tagging is to develop an accurate and efficient POS tagger for Hindi language. The project aims to achieve the following objectives:

1)Data Collection: Collect a large corpus of annotated Hindi text to serve as the training data for developing the POS tagger.

2)Preprocessing: Perform necessary preprocessing steps, such as tokenization, stemming, and lemmatization, to prepare the text data for POS tagging.

3)Model Selection: Explore and compare various machine learning algorithms and deep learning models for Hindi POS tagging, such as hidden Markov models, conditional random fields, and neural networks, and select the most suitable model for the task.

4)Training and Evaluation: Train the selected model on the annotated training data and evaluate its performance on a separate test set to measure its accuracy and efficiency.

5)Improvement: Experiment with different feature engineering techniques, hyperparameter tuning, and data augmentation methods to improve the performance of the POS tagger.

The ultimate goal of this project is to develop a robust and accurate Hindi POS tagger that can be used in various NLP applications, such as machine translation, sentiment analysis, and information retrieval. The project also aims to contribute to the development of bett

## c). Scope of Project

The scope of this project on Hindi POS tagging includes:

1)Developing a Hindi POS tagging system: The main objective of the project is to develop an accurate and efficient POS tagger for Hindi language. The project will involve exploring and comparing different machine learning algorithms and deep learning models for Hindi POS tagging, and selecting the most suitable model for the task.

2)Collecting and preprocessing data: The project will involve collecting a large corpus of annotated Hindi text to serve as the training data for developing the POS tagger. The text data

will also need to be preprocessed, which includes steps such as tokenization, stemming, and lemmatization.

3)Training and evaluating the POS tagger: The selected model will be trained on the annotated training data and evaluated on a separate test set to measure its accuracy and efficiency. The project will involve experimenting with different feature engineering techniques, hyperparameter tuning, and data augmentation methods to improve the performance of the POS tagger.

4)Developing language resources: The project will also contribute to the development of better language resources for Hindi NLP research, such as annotated Hindi text data, and can be used as a foundation for future research in the field.

The scope of the project is limited to the development of a Hindi POS tagging system and the associated tasks required to train and evaluate it. The project does not include the development of NLP applications that use the POS tagger, although the POS tagger can be used as a component in various NLP applications

## d). Intended Audience (who all can use your project)

The intended audience for this project on Hindi POS tagging includes:

1)NLP researchers: The project can be of interest to NLP researchers who are interested in developing NLP applications for Hindi language. The project can also contribute to the development of better language resources for Hindi NLP research.

2)Language technology professionals: The project can be of interest to language technology professionals who are involved in developing language processing tools and applications for Hindi language.

3)Language educators: The project can be of interest to language educators who teach Hindi language, as it can help improve the quality of language processing tools and applications used in language education.

4)Students: The project can be of interest to students who are interested in learning about NLP and developing language processing tools and applications. The project can also serve as a learning resource for students who are studying Hindi language and want to learn more about its morphological system and syntactic structure.

5)General public: The project can be of interest to the general public who are interested in learning about language processing and its applications, and how technology can be used to analyze and understand natural language.

## 2. Overall Description

## a) Fundamental Requirements

The functional requirements for this project on Hindi POS tagging can include:

1)Data collection: The system should be able to collect a large corpus of annotated Hindi text for training and testing the POS tagger.

2)Preprocessing: The system should be able to perform necessary preprocessing steps, such as tokenization, stemming, and lemmatization, to prepare the text data for POS tagging.

3)Model selection: The system should be able to explore and compare different machine learning algorithms and deep learning models for Hindi POS tagging, and select the most suitable model for the task.

4)Training and evaluation: The system should be able to train the selected model on the annotated training data and evaluate its performance on a separate test set to measure its accuracy and efficiency.

5)Feature engineering: The system should be able to experiment with different feature engineering techniques, such as using word embeddings or character n-grams, to improve the performance of the POS tagger.

6)Hyperparameter tuning: The system should be able to tune the hyperparameters of the selected model, such as the learning rate or regularization strength, to optimize its performance.

7)Data augmentation: The system should be able to augment the training data using techniques such as random deletion, substitution, or permutation of words to improve the robustness of the POS tagger.

8)User interface: The system should provide a user interface that allows users to input text data and receive the corresponding POS tags as output.

9)Performance metrics: The system should be able to calculate various performance metrics, such as precision, recall, and F1 score, to evaluate the accuracy and efficiency of the POS tagger.

10)Output formats: The system should be able to output the POS tags in various formats, such as text, XML, or JSON, depending on the needs of the user

## b. Operating Environment (Database, S/W and H/W requirement)

The operating environment for this project on Hindi POS tagging will depend on the specific implementation and the chosen programming language, libraries, and frameworks. However, some general requirements for the operating environment can include:

1)Hardware requirements: The system should be able to run on a standard personal computer or server with sufficient memory, processing power, and storage capacity to handle the size of the data and the complexity of the models.

2)Software requirements: The system should be able to run on a modern operating system, such as Windows, Linux, or macOS, and support the required software dependencies, such as Python, Anaconda, or TensorFlow.

3)Programming language: The system should be implemented in a programming language that supports the required machine learning and deep learning libraries, such as Python, Java, or C++.

4)Libraries and frameworks: The system should use relevant libraries and frameworks for NLP and machine learning, such as NLTK, spaCy, Scikit-learn, or Keras, depending on the specific requirements of the project.

5)Database management: The system may require a database management system to store and manage the annotated text data and the trained models, such as MySQL, PostgreSQL, or MongoDB.

## 3. Conclusion

In conclusion, the project on Hindi POS tagging has the potential to contribute to the development of better NLP tools and applications for Hindi language. The project involves collecting a large corpus of annotated Hindi text, performing preprocessing, experimenting with various machine learning and deep learning models, feature engineering, hyperparameter tuning, and evaluating the performance of the POS tagger. The project can be useful for NLP researchers, language technology professionals, language educators, students, and the general public interested in natural language processing and its applications. The project's success will depend on the availability of high-quality annotated data, the selection of appropriate models and techniques, and the careful evaluation of the performance metrics. Overall, the project can provide valuable insights into the Hindi language's morphological and syntactic structures and contribute to the development of better language resources and tools for Hindi NLP research and applications.