

|  |
| --- |
|  |
| A Mini Project Report on |
| **Rule-Based POS Tagger**  **(Hindi)** |
| *Submitted in partial fulfillment of the CIE for the subject*  **Natural Language Processing (ISEA2)** |
| by |
| **Ahmed Uddin Khalid Syed (1MS17IS008)**  **Ayush Agrawal (1MS17IS031)**  **Nida Rais (1MS17IS073)** |
|  |
|  |
| *Under the guidance*  *of* |
| **Ms. Rajeshwari S B** |
| Assistant Professor |
|  |
| Department of ISE, MSRIT |
|  |
| Bengaluru-560054 |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **1.** | **Introduction** | **3** |
| **2.** | **Problem Statement** | **5** |
| **3.** | **Source Code** | **6** |
| **4.** | **Testing, Results and Discussion** | **8** |
| **5.** | **Conclusion and Future Work** | 12 |
|  |  |  |
|  |  |  |

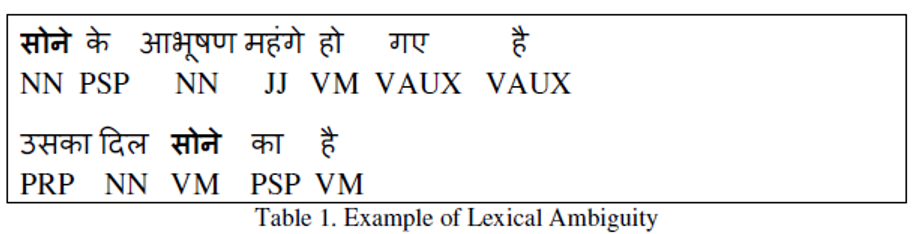
**Introduction**

Natural language processing is a field of computer science, artificial intelligence (also called machine learning) and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output.

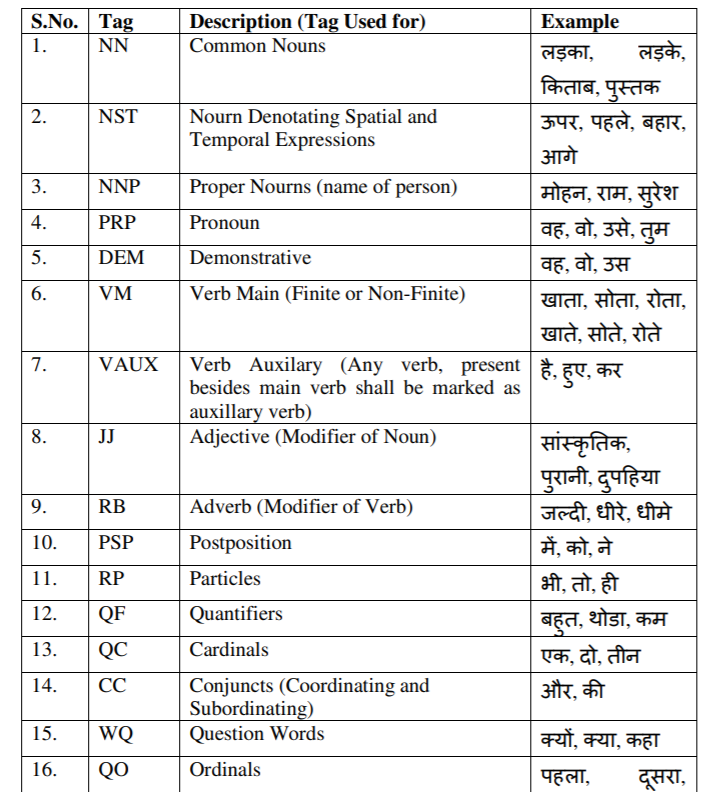
Part of Speech tagger is an important application of natural language processing. Part of speech tagging is the process of assigning a part of speech like noun, verb, preposition, pronoun, adverb, adjective or other lexical class marker to each word in a sentence.

There are a number of approaches to implement part of speech tagger, i.e. Rule Based approach, Statistical approach and Hybrid approach. Rule-based tagger use linguistic rules to assign the correct tags to the words in the sentence or file. Statistical Part of Speech tagger is based on the probabilities of occurrences of words for a particular tag. Hybrid based Part of Speech tagger is combination of Rule based approach and Statistical approach.

Our project implements rule-based tagging in Hindi.



POS tags for Hindi words:



**Problem Statement**

Though there are obviously many approaches to POS tagging, tagging of Indian Languages still poses a challenge. This is due to the morphological richness of Indian Languages.

Rule-based part-of-speech tagging is the oldest approach that uses hand-written rules for tagging. Rule based taggers depends on dictionary or lexicon to get possible tags for each word to be tagged. Hand-written rules are used to identify the correct tag when a word has more than one possible tag.

**Source Code**

import nltk

from nltk.tag import brill

from nltk.corpus import indian

nltk.corpus.indian.words('hindi.pos')

from nltk.tag import untag

word\_patterns = [

(r'^-?[0-9]+(.[0-9]+)?$', 'CD'),

(r'.\*देश$', ' NN'),

(r'.\*है$', ' VAUX'),

(r'.\*सबसे$', ' JJS'),

(r'.\*और$', ' CC'),

(r'.\*ड़ा$', ' JJ'),

(r'.\*वो$', ' PRP'),

(r'.\*रा$', ' PP'),

(r'.\*लो$', ' VB'),

(r'.\*ता$', ' VM'),

(r'.\*गे$', ' VB'),

(r'.\*ना$', ' VB'),

(r'.\*था$', ' VBD'),

(r'.\*थे$', ' VBD'),

(r'.\*कर$', ' VB'),

(r'.\*ह$', ' PRP'),

(r'.\*के$', ' VB'),

(r'.\*आॅं$', ' NN'),

(r'.\*का$', ' NN'),

(r'.\*हाॅं$', ' PREP'),

(r'.\*दार$', ' JJ'),

(r'.\*आ$', ' JJ'),

(r'.\*धर$', ' PREP'),

(r'.\*म$', ' NNC')

]

regexp\_tagger = nltk.RegexpTagger(word\_patterns)

sentences = list(indian.tagged\_sents('hindi.pos'))

print(len(sentences))

training\_data = sentences[:50]

gold\_data = sentences[50:57]

testing\_data = [untag(s) for s in gold\_data]

training\_data [0]

brill.Template.\_cleartemplates()

templates = brill.fntbl37()

trainer = nltk.tag.brill\_trainer.BrillTaggerTrainer(initial\_tagger=regexp\_tagger,templates=templates, trace=3,deterministic=True)

tagger1 = trainer.train(training\_data, max\_rules=10000)

Rules = tagger1.rules()

text = " वह खाना खाता है ।"

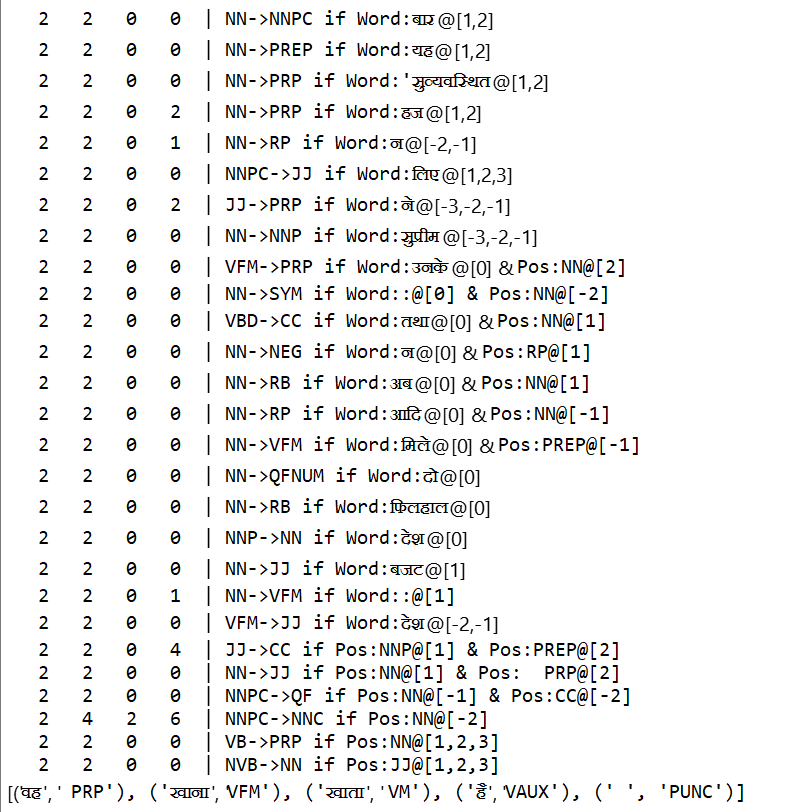
model = tagger1

new\_tagged = (model.tag(nltk.word\_tokenize(text)))

print(new\_tagged)

print()

**Testing, Results and Discussion**

****

**Conclusion and future work**

Automatic part of speech tagging is an area of natural language processing where statistical techniques have been more successful than rule- based methods. In this paper, we present a simple rule-based part of speech tagger. The rule-based tagger has many advantages over these taggers, including: a vast reduction in stored information required, the perspicuity of a small set of meaningful rules, ease of finding and implementing improvements to the tagger, and better portability from one tag set, corpus genre or language to another. Perhaps the biggest contribution of this work is in demonstrating that the stochastic method is not the only viable method for part of speech tagging. The fact that a simple rule-based tagger that can perform so well should offer encouragement for researchers to further explore rule-based tagging, searching for a better and more expressive set of rule templates and other variations.