ISSN:0975-9646

Rule Based POS Tagger for Marathi Text

Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar

Department of Computer Technology, YCCE
Nagpur- 441110, Maharashtra, India

Abstract **- Part-of-Speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, adjective, adverb, or other lexical class marker to each word in a sentence. This paper presents a POS Tagger for Marathi language text using Rule based approach, which will assign part of speech to the words in a sentence given as an input. We describe our system as the one which tokenizes the string into tokens and then comparing tokens with the WordNet to assign their particular tags. There are many ambiguous words in Marathi language and we resolve the ambiguity of these words using Marathi grammar rules.**

Keywords- **POS-Part Of Speech, WordNet, Tagset, Corpus.**

I. INTRODUCTION

Part-of-Speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, adjective, adverb, or other lexical class marker to each word in a sentence. POS tagging is a necessary pre-module to other natural language processing tasks like natural language parsing, semantic analyzer, information extraction and information retrieval. A word can occur with different lexical class tags in different contexts. The main challenge in POS tagging involves resolving this ambiguity in possible POS tags for a word. We developed a POS tagger which will assign part of speech to the word in a sentence provided as input to the system. Here we have assigned five tags only viz. noun, adverb, adjective, verb and pronoun. Several approaches have been proposed and successfully implemented for POS tagging for different languages.

There are various approaches of POS tagging, which can be divided into three categories; rule based tagging, statistical tagging and hybrid tagging.

A. Rule based approach:

The rule based POS tagging model requires a set of hand written rules and uses contextual information to assign POS tags to words. The main drawback of rule based system is that it fails when the text is unknown, because the unknown word would not be present in the WordNet. Therefore the rule based system cannot predict the appropriate tags. Hence for achieving higher accuracy in this system we need to have an exhaustive set of hand coded rules.

B. Statistical approach:

A statistical approach includes frequency and probability. The simplest statistical approach finds out the most frequently used tag for a specific word from the annotated training data and uses this information to tag that word in the unannotated text. These systems are having more efficiency than the rule based approach. The problem with this approach is that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules of a language.

C. Hybrid approach:

A hybrid approach may perform better than statistical or rule based approaches. The POS tagger which is implemented using hybrid approach is having higher accuracy than the individual rule based or statistical approach. The hybrid approach first uses the set of hand coded language rules and then applies the probabilistic features of the statistical method.

Most common POS taggers use a POS dictionary, which is also known as WordNet, having words tagged with a small set of possible output tags.

In this paper we are presenting the POS Tagger for Marathi Language .The main problem in part of speech tagging is ambiguous words. The Marathi Language is full of ambiguous words. There may be many words which can have more than one tag. To solve this problem we consider the context instead of taking single word.

For example-

ते फूल लाल आहे .

The given sentence is ambiguous because 'लाल ' can be used as an adverb as well as an adjective but by conventional Marathi grammar rules, 'लाल ' should be a adverb because it is coming before a verb but it is an adjective.

II. LITERATURE SURVEY

Considerable amount of work has already been done in the field of POS tagging for English and other foreign languages. Different approaches like the rule based approach, the stochastic approach and the transformation based learning approach along with modifications have been tried and implemented. However, if we look at the same scenario for South-Asian languages such as Marathi and Hindi, we find out that not much work has been done [5]. The main reason for this is the unavailability of a considerable amount of annotated corpora of sound quality, on which the tagging models could train to generate rules for the rule based and transformation based models and probability distributions for the stochastic models. In the following sections, we describe some POS tagging models that have been implemented for Indian languages along with their performances.

We have found that most of the research on POS tagging on the South-Asian languages has been done using statistical approaches like HMM, MEM etc.HMM i.e. Hidden Markov model based tagger is described in [2], reporting a performance of 76.49% accuracy on training and test data having about 25000 and 6000 words, respectively. This tagger uses HMM in combination with probability models of certain contextual features for POS tagging.

In 2007, Asif Ekbal [6] proposed a HMM based POS tagger for Hindi, Bengali and Telugu. Here they make use of pre-tagged corpus and HMM. Handling of unknown words is based on suffixes. It reported accuracy of 90.90% for Bengali, 82.05% for Hindi and 63.93% for Telugu.

In the year 2006, Pranjal Awasthi [7] proposed an approach to POS tagging using a combination of HMM and error driven learning. They have used Conditional Random Fields (CRF), TnT, and TnT with Transformation Based Learning (TBL) approaches and have reported accuracy of 69.4%, 78.94%, and 80.74% respectively for the three approaches for Hindi.

Sankaran Baskaran [8] in the year 2006 used HMM based approach for tagging and chunking. He achieved a Precision of 76.49% for tagging and 55.54% for chunking using the tag-set developed in IIIT-Hyderabad.

In 2006, Himanshu Agrawal and Anirudh Mani [3] presented a CRF based POS tagger and chunker for Hindi. Various experiments were carried out with various sets and combinations of features which mark a gradual increase in the performance of the system. A Morphological analyzer was used to provide extra information such as root word and possible POS tags for training. Training on 21,000 words, they could achieve an accuracy of 82.67%.

Pattabhi R K Rao [4] in the year 2007 proposed a hybrid POS tagger for Indian languages. Handling of unknown words is based on lexical rules. Precision and Recall for Telugu were 58.2% and 58.2% respectively.

For the Telugu language, Sudheer K. in [9] reported the performances of various approaches of POS tagging. Here the pre-annotated training corpora are the training data released for the NLPAI Machine Learning Competition 2006, consisting of 27336 words. The size of the testing data used is around 5662 tokens. Using the above data, the HMM based approach demonstrates an accuracy of 82.47% whereas the MEM based approach displays 82.27% which are very similar.

III. METHODOLOGY

A .Databases used

1) WordNet:

WordNet is an electronic database which contains parts of speech of all the words which are stored in it. It is trained from the corpus for higher performance and efficiency.

2) Corpus:

For correct POS tagging, training the tagger well is very important, which requires the use of well annotated corpora. Annotation of corpora can be done at various levels which include POS, phrase or clause level, dependency level etc. For POS Tagging in Marathi we are using a corpus which is based on tourism domain. It is an annotated corpus. As not much work done on Marathi language, we had to start with the unannotated corpus we took a small part of it and manually tag it.

3) Tagset

Apart from corpora, a well-chosen tagset is also important. For deciding upon a tagset, we should consider the following properties:

- Fineness Vs coarseness

  When choosing the tagset for a POS tagger, we have to decide whether the tags will allow for precise distinction of the various features of POS of the language i.e. whether features like plurality, gender and other information should also be available or whether the tagger would only provide the different lexical categories.

- Syntactic function Vs lexical category

  The lexical category of a word can be different than the POS of the word in a sentence, and the tagset should be able to represent both.

  E.g. लाल – Noun, Adjective (lexical category)

  ते फूल लाल आहे – adjective (syntactic category)

- New tags Vs tags from a standard tagger

  It has to be decided whether an existing tagset should be used, or a new tagset should be applied according to the specifics of the language on which the tagger will work.

  In Marathi POS tagger we use Marathi WordNet as a tagset which will be working as our database. The record in the tagset consists of two parts, first is the word along with its intended tag and second is the root word for the corresponding word. The tag representation consists of 4 bits which represents Noun, Adjective, Adverb, and Verb.

- When the first bit is 1 i.e. 1000 the word is a noun.
- When the second bit is 1 i.e. 0100 the word is an Adjective.
- When the third bit is 1 i.e. 0010 the word is an Adverb.
- When the fourth bit is 1 i.e. 0001 the word is a verb.
- We also have combinations like 1100 for ambiguous words that can be used both as a noun and as an Adjective.
- Another combination which we have for ambiguous words is 0110. This means that the specified word can be used both as an Adjective and as an Adverb.

  For pronouns we are using a separate database which contains all the possible pronouns which can be used in Marathi Language.

B. Details of identified modules

The Marathi sentence that is to be analyzed is given as an input by the user. The input is then sent to tokenizing function.

1) Tokenizer

This module generates the tokens of the given input sentence and the delimiter that is used for tokenizing is space followed by dot(.) . It also calls the other modules when required. The tokens of the sentence are basically stored in a String array for further processing.

2) Tagging

The tagging module assigns tags to tokens and also search for ambiguous words and according to their type assign some special symbols to them. If we encounter words which are not present in the WordNet they are treated as unidentified. These unidentified tokens are compared with the pronoun database if these tokens are present in the database then they are treated as pronouns. The ambiguous words are those words which act as a noun and adjective or adjective and adverb according to different context.

3) Resolving Ambiguity

The ambiguity which is identified in the tagging module is resolved using the Marathi grammar rules. These rules are:

Rule 1:

If we have a token which is assigned notation as 0110 signifies that it can be used as an adjective as well as an adverb, then such ambiguity is resolved as:

- if the next token is a noun or an adjective then the ambiguous token becomes an adjective.
- if the next token is a verb then the ambiguous token becomes an adverb.

Rule 2:

If we have a token which is assigned notation as 1100 signifies that it can be used as a noun as well as an adjective, then such ambiguity is resolved as:

- if the next token is a noun and the previous token is not a noun then the ambiguous word becomes an adjective
- otherwise it becomes an adverb.

Rule 3:

If we have a token which is assigned notation as 1100 signifies that it can be used as a noun as well as a adjective, then such ambiguity is resolved as:

- if the previous token is a noun then the ambiguous word becomes an adjective, even if the next token is a verb.
- otherwise it becomes an adverb.

4) Displaying results

This module will be displaying the final result. The tokens i.e. words in the sentences are shown with their corresponding parts of speech.

*C. **Flowchart***

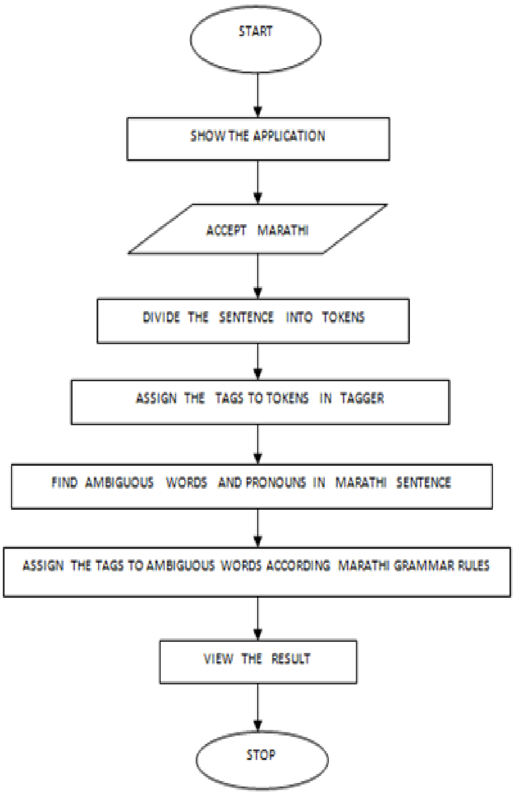

START
SHOW THE APPLICATION
ACCEPT MARATHI
DIVIDE THE SENTENCE INTO TOKENS
ASSIGN THE TAGS TO TOKENS IN TAGGER
FIND AMBIGUOUS WORDS AND PRONOUNS IN MARATHI SENTENCE
ASSIGN THE TAGS TO AMBIGUOUS WORDS ACCORDING MARATHI GRAMMAR RULES
VIEW THE RESULT
STOP

Fig.1 Flowchart

*D. **Process Overview***

The Marathi sentence is taken as input from the user, then the tokens are created i.e. each word is separated. Then tagging is done by comparing with the words in the WordNet along with this, ambiguous words and pronouns are found out. The ambiguous words are those words which can act as a noun and adjective in certain context, or act as an adjective and adverb in certain context. Then their ambiguity is resolved using Marathi grammar rules as stated.

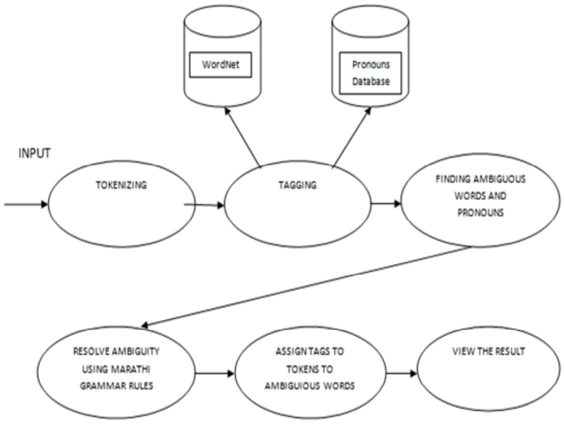

WordNet
Pronouns Database
INPUT
TOKENIZING
TAGGING
FINDING AMBIGUOUS WORDS AND PRONOUNS
RESOLVE AMBIGUITY USING MARATHI GRAMMAR RULES
ASSIGN TAGS TO TOKENS TO AMBIGUOUS WORDS
VIEW THE RESULT

Fig.2 Process overview

IV Result

Case 1: Normal sentences

1. मेहनती शेतकरी खूप पिके काढतात .

The above sentence is given as an input to the POS tagger and we get the output as:

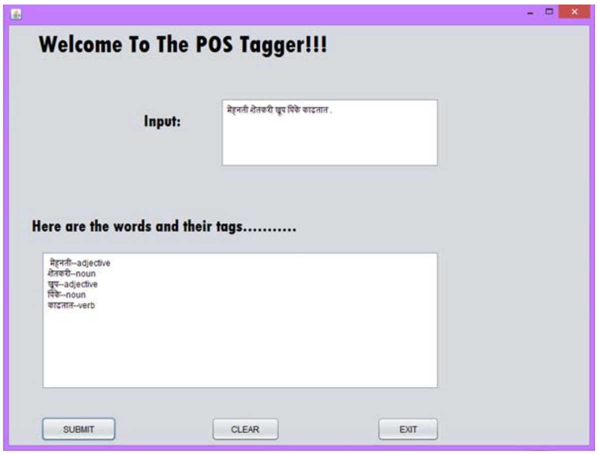
मेहनती -adjective

शेतकरी -noun

खूप -adjective

पिके -noun

काढतात -verb



Welcome To The POS Tagger!!!

Input: मेहनती शेतकरी खूप पिके काढतात .

Here are the words and their tags...........

मेहनती--adjective
शेतकरी--noun
खूप--adjective
पिके--noun
काढतात--verb

Fig.3 Example 1

when we compare the tokens with the words in the WordNet, we find entries of मेहनती as an adjective (0100), शेतकरी as a noun (1000), खूप as an adjective (0100), पिके as a noun (1000) and काढतात as a verb (0001).

Case 2: ambiguity of adjective and adverb

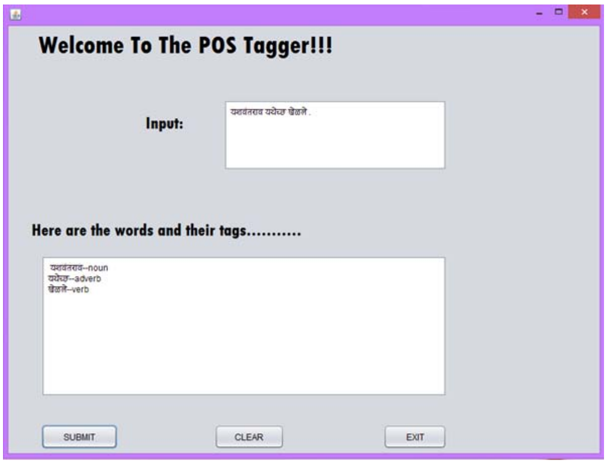
यशवंतराव यथेच्छ खेळले .

The above sentence is given as an input to the POS tagger and we get the output as:

यशवंतराव -noun

यथेच्छ -adverb

खेळले –verb



Welcome To The POS Tagger!!!

Input: यशवंतराव यथेच्छ खेळले .

Here are the words and their tags..........

यशवंतराव--noun
यथेच्छ--adverb
खेळले--verb

Fig.4 Example 2

when we compare the tokens with the words in the WordNet, we find entries of यशवंतराव as a noun (1000), यथेच्छ as 0110 and खेळले as a verb (0001).the ambiguity for the word ' यथेच्छ ' is resolved using Marathi grammar rule 1 as stated above.

Case 3: ambiguity of adjective and noun

यशवंतराव यथेच्छ भोजन् करीत आहे .

The above sentence is given as an input to the POS tagger and we get the output as:

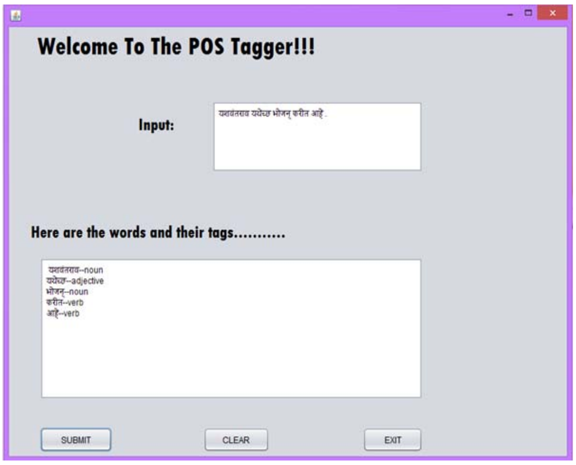
यशवंतराव -noun

यथेच्छ -adjective

भोजन् -noun

करीत –verb

आहे –verb



Welcome To The POS Tagger!!!

Input: यशवंतराव यथेच्छ भोजन् करीत आहे .

Here are the words and their tags..........

यशवंतराव--noun
यथेच्छ--adjective
भोजन्--noun
करीत--verb
आहे--verb

Fig.5 Example 3

when we compare the tokens with the words in the WordNet, we find entries of यशवंतराव as a noun (1000), यथेच्छ as 0110 , भोजन् as a noun (1000),and करीत as a verb (0001), आहे as a verb. The ambiguity for the word ' यथेच्छ ' is resolved using Marathi grammar rule 2 as stated above.

Case 4: ambiguity of adjective and noun in special case

ते फूल लाल आहे .

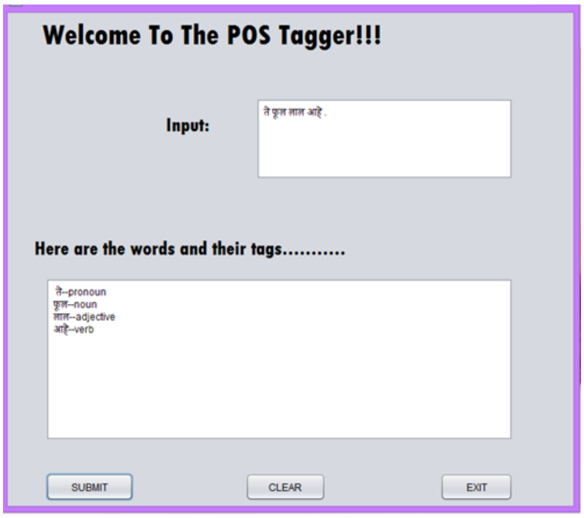
The above sentence is given as an input to the POS tagger and we get the output as:

ते -pronoun

फूल -noun

लाल -adjective

आहे –verb



Welcome To The POS Tagger!!!

Input: ते फूल लाल आहे .

Here are the words and their tags..........

ते--pronoun
फूल--noun
लाल--adjective
आहे--verb

Fig. 6 Example 4

The given case is special because by conventional Marathi grammar rules, 'लाल ' should be a adverb because it is coming before a verb but it is an adjective. When we compare the tokens with the words in the WordNet and pronoun database, we find entries of ते as a pronoun, फूल as noun (0110) , लाल as a 1100,and आहे as a verb (0001).the ambiguity for the word 'लाल' is resolved using Marathi grammar rule 3 as stated above.

V. Conclusion

Part of Speech Tagging is playing a vital role in most of the natural language processing applications. Since Marathi an ambiguous language, it is hard for tagging. The rule based POS tagger described here is resolving ambiguity and assigning the tags to the ambiguous words using Marathi grammar rules. It provides correct tag for all the words that are present in the WordNet. The range of words for which the POS tagger can be used, can be raised by updating the WordNet.

VI. REFERENCES

1. Jyoti Singh, Nisheeth Joshi, Iti Mathur,"Development of Marathi Part of Speech Tagger Using Statistical Approach".
2. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, pages: 257–286.
3. Agarwal, H., Mani, "Part of Speech Tagging and Chunking with Conditional Random Fields. "In Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad,India (2006).
4. Pattabhi, R.K.R., SundarRam, R.V., Krishna, R.V., Sobha, L.,"A Text Chunker and Hybrid POS Tagger for Indian Languages"In Proceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages, IIIT Hyderabad, Hyderabad, India (2007).
5. Fahim Muhammad Hasan," Comparison Of Different Pos Tagging Techniques", Brac University, Dhaka, Bangladesh, , pages: 13,2006.
6. Ekbal, A., Mandal, S.: POS Tagging using HMM and Rule based Chunking. In: Proceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages, IIIT Hyderabad, Hyderabad, India (2007).
7. Awasthi, P., DelipRao, Ravindran, B.: Part of Speech Tagging and Chunking with HMM and CRF. In: Proceedings of NLPAI Machine LearningWorkshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006).
8. Baskaran, S.: Hindi Part of Speech Tagging and Chunking. In: Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006).
9. Karthik Kumar G, Sudheer K, Avinesh Pvs, "Comparative Study of Various Machine Learning Methods For Telugu Part of Speech Tagging", In Proceedings of the NLPAI Machine Learning 2006 Competition.