© 2018 IJSRST | Volume 4 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Rule Based Part-of-Speech Tagger for Marathi Language

Gaikwad Deepali K.*, Naik Ramesh R. , C. Namrata Mahender
Department C.S. and I.T, Dr. Babasaheb Ambedkar Marathawada University, Aurangabad, Maharashtra, India

ABSTRACT

A part of speech (POS) tagging is one of the best studied problems in the field of Natural Language Processing (NLP). POS tagging is the process of assigning a part-of-speech like noun, verb, adjective, adverb to each word in a sentence. In this paper we present a Marathi part of speech tagger. It is morphologically rich language. It is spoken by the native people of Maharashtra. POS tagging is difficult for Marathi language due to unavailability of corpus for computational processing. In this paper, a POS Tagger for Marathi language using Rule based technique is presented. Our proposed system which tokenizes the string into tokens, find root word using morphological analyzer and compare the root word with the WordNet to assign appropriate tag. If word has assigned more than one tags then by using Marathi grammar rules ambiguity is removed. Meaningful rules are provided to improve the performance of the system.

Keywords: Part of Speech (POS), Tokenization, Stemmer, Morphological Analyzer, Tag Generation.

I. INTRODUCTION

One of the most important activities in natural language processing is Part of Speech Tagging that has begun in the early 1960s. A POS tagging is the process of assigning exact tag like nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories to each word in the input sentence, based on its classification as well as its relationship with adjacent and related words in a phrase, sentence, or paragraph. POS Tagging is a primary stage of linguistics, text analysis like information retrieval, machine translator, text to speech synthesis, information extraction etc. It is a basic form of morphological analysis where it only deals with assigning an appropriate POS tag to the word, while morphological analysis is used to find the internal structure of the word. Indian languages are morphologically rich. Unavailability of corpus for computational processing due to this tagging of Indian languages i.e, Marathi language is difficult [1]. It is an extremely powerful and accurate tool used in any application that deals with natural language processing [2]. The tagging performance totally depends on tag dictionary. It is also called grammatical tagging or word-category disambiguation.

Taggers can be classified as supervised or unsupervised: Supervised taggers are based on pre-tagged corpora, whereas unsupervised taggers automatically assign tags to words. The approaches of POS tagging, which can be further divided into three categories; rule based tagging, stochastic tagging and hybrid tagging.

i. Rule Based Tagging: The rule based POS tagging approach that uses a set of hand crafted rules. 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 or corpus.
ii. Stochastic Taggers: A stochastic approach assigns a tag to word using frequency, probability or statistics. 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.

iii. Hybrid Taggers: The hybrid approach, assign tag to the word using statistical approach after that, if wrong tag is found then by applying some rules tagger tries to change it. The hybrid approach first uses the set of hand coded language rules and then applies the probabilistic features of the statistical method POS tagging is needed as a pre-processing module for NLP application. But POS tagging is difficult for Marathi language because corpus is not available for computational processing. The rule based part of speech tagger that assigns all possible tags to words in a sentence given as an input and uses a set of hand written rules. Dictionary plays an important role to assign appropriate tag to each word. The tagger divided into two stages. First, it searches words in corpus, if word found in corpus then it assigns a tag. Otherwise, stems and morphologically analyse the word [3].

In this paper we are presenting the POS Tagger for Marathi Language. The POS tagging consists of three stages: Tokenization, Stemming and Morphological Analysis.

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. However, if we look at the same development for Marathi, we find out that not much work has been done. The main reason for this is the unavailability of annotated quality corpora, 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 Table 1, we describe some POS tagging models that have been implemented for Indian languages along with their performances.

Table 1. Survey Of Pos Tagger For Different Indian Languages

Authors	Approaches	Languages	Performance
Pranjal Awasthi (2006)	HMM and error driven learning using Conditional Random Fields (CRF), TnT, and TnT with Transformation Based Learning (TBL) approaches	Hindi	69.4%, 78.94%, and 80.74% [4]
Sankaran Baskaran (2006)	HMM used tagging and chunking	-	76.49% for tagging and 55.54% for chunking [5]
Himanshu Agrawal & Anirudh Mani (2006)	Conditional Random Fields (CRF)	Hindi	82.67% [6]
Pattabhi R.K. Rao (2007)	Hybrid POS tagger	Telgu	Precision 58.2% and Recall 58.2% [7]
Hasan (2007)	Unigram, Bigram, HMM and Brill's POS Tagging Approaches	Bangla, Hindi and Telugu	-[8]

Kumar (2007)	statistical part-of-speech tagger (maximum entropy Markov model)	Hindi	94.89% [9]
Asif Ekbal (2007)	HMM based POS tagger	Hindi, Bengali and Telugu	90.90% for Bengali, 82.05% for Hindi and 63.93% for Telugu [10, 20]
Patel (2008)	Machine Learning using CRF	Gujarati	92% [11]
Singh (2008)	Conditional Random Field (CFR) and Support Vector Machine (SVM)	Manipuri	72.04% for CRF and 74.38% for SVM. [12]
Dhanalakshmi V (2009)	SVM based machine learning	Tamil	95.64% [13]
Ekbal and Sivaji (2008)	POS taggers using Hidden Markov Model (HMM) and Support Vector Machine (SVM)	Bengali	85.56% for HMM and 91.23% for SVM [14]
Manju K. (2009)	Hidden Markov Model (HMM)	Malayalam	90%[15]

[1, 16].

III. PROPOSED SYSTEM

We proposed a rule based part of speech tagger that assigns parts of speech to each word, such as noun, verb, adjective, adverb etc in a sentence. The proposed approach consists of following phases:

A. Tokenization
B. Stemmer
C. Morphological Analyzer
D. Tag Generation

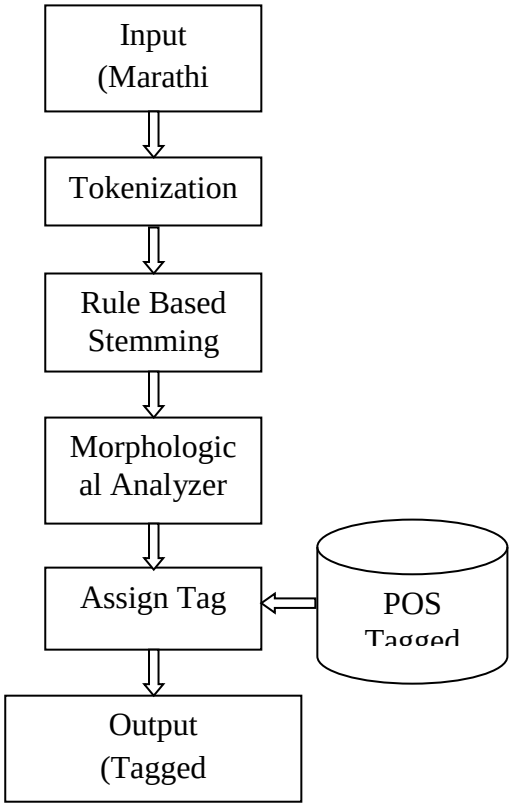

Figure 1. Proposed System

A. Tokenization

Tokenization is the process of separating tokens from input text. Each word is separated from sentence by white space or punctuation marks and treat as single token and then deal with each word individually. The split up of input text into tokens is important for POS tagging.

B. Stemming

A stemming is a process of converting morphologically identical words to root word d affixes without applying morphological analysis of that term. Stemming techniques are divided into two categories: Language Specific (Rule -Based) and Statistical (Corpus-Based) techniques.

i. Language Specific (Rule-Based) Stemmer: Language Specific or Rule-Based stemmer makes use of certain pre-define rule according to language to map the morphological alternative of the word to its base form. This language related rules are created manually by the linguists. Rule-based stemming methods are further divided into three categories: Table Lookup, Affix Stripping, and Morphological.

ii. Statistical (Corpus-Based) Stemmer: It based on unsupervised learning of the language by analyzing the lexicon or finding the co-occurrence or context of the words in the corpus. These are also called corpus-based techniques. These algorithms also perform suffix stripping but after performing some statistical analysis on the corpus. The major advantage of statistical techniques is that it does not require any prior knowledge of the language [17].

Stemming is important in the system, which uses a suffix and affix list to remove suffixes and affixes from words and thus reduces the word to its stem. The result of stemming is stem of word that can be given as input to Morphological Analyzer for further processing [18]. We use one of them is Affix stripping Approach of Rule Based Stemmer for stemming. Affix that is, prefix or suffix of the word. Affix removal algorithms delete suffix and/or prefix of the word according to specific rules or suffix list. Most of the work has been done on suffix stripping as compare to prefix [3]. We have designed the rule based stemmer, which is discussed in detailed in below section:

The common morphological patterns found in Marathi are:

i. <Original word>= <stem/root words + plain suffixes>

   e. g. भारती = भारत + ी

ii. <Original word>=<stem/root words + plain suffixes + complex suffixes>

   e. g. झाडावर = झाड + ा+ वर

iii. <Original word>= <stem/root words + plain suffix + join word + complex suffixes>

   e. g.घराच्यासाठी = घर + ा + च्या + साठी

The suffix stripping rules for the rule-based stemmer are based on these patterns. According to our present research work, we only stem noun words

C. Morphological Analyzer

The morphological analysis is used to identify the inner structure of the word. It analyzed the stem word to check whether they are twisted or not. If stem word is twisted then the root word is formed by addition of replacement characters with stem word. A morphological analyzer is expected to produce root words for a given input document and it is carried out by dictionary lookup and morpheme analysis rules [2].

D. Tag Generation

Tag generation is final stage of POS tagging. It assigns appropriate tag to morphological analyzed word or root word [19].

Algorithm for POS tagging System:

1) Take input text and generate a token.
2) Use tokens to generate stem word.
3) Use rule to generate root word from stem word using morphological analyzer.
4) Select each word one by one and compare with WordNet or corpus.

5) Assign appropriate tag to each word.

IV. RESULT

In this research paper, we collect 1364 word for testing the performance of our system. This system gives the 100% accuracy.

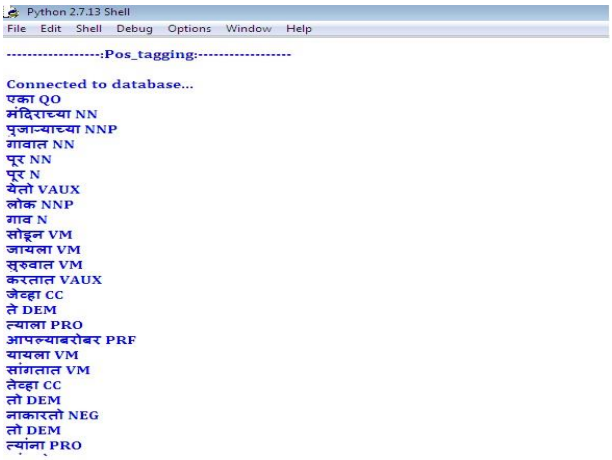

Figure 2. Result of Rule based POS Tagger for Marathi Language

The accuracy was calculated using the formula:

$$\text{Accuracy} = \frac{\text{No. of Correctly tagged}}{\text{No. of correctly tagged + No. of Incorrectly tagged}} * 100$$

V. CONCLUSION

Part of speech tagging is very useful in natural language processing applications like Information extraction, Text summarization, Question Answering System, etc. In this paper, we discuss about Marathi part of speech tagger but Marathi is an ambiguous language, it is hard for tagging. The rule based part of speech tagger design that resolving ambiguity and assigning the tags to the words using Marathi grammar rules. It assigns all possible tag for all the words that are present in the Marathi WordNet or corpus. If word is not present in Marathi WordNet, that word is converting into stem word and stem word is passed as input to morphological analyzer for converting into root word. Again search this word in Marathi WordNet and then assign tag to it. The result of our Rule Based Part of Speech Tagger is 100%. .

VI. REFERENCES

[1]. Singh Jyoti. Joshi Nisheeth and Mathur Iti. 2013. Part of Speech Tagging of Marathi Text using Trigram Method. International Journal of Advanced Information Technology (IJAIT). Vol. 3. No.2.

[2]. Govilkar Sharvari. Bakal J. W and Rathod Shubhangi. 2015. Part of Speech Tagger for Marathi Language. International Journal of Computer Applications. Volume 119-No.18.

[3]. Gaikwad Deepali. 2017. Rule Based Text Summarization for Marathi Text. M.Phil. Thesis. Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. India.

[4]. Awasthi P. Delip Rao and RAvindran B. 2006. Part of Speech Tagging and Chunking with HMM and CRF. In Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian Languages. IIIT Hyderabad, India.

[5]. Baskaran S. 2006. 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, India.

[6]. Agrawal H. And Mani. 2006. 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, India.

[7]. Pattabhi RKR. SundarRam RV. Krishna RV And Sobha L. 2007. A Text Chunker and Hybrid POS Tagger for Indian Laguages. In Prceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages. IIIT Hyderabad, India.

[8]. Hasan Fahim Muhammad. Zaman Naushad Uz and Khan Mumit. 2007. Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages. In proceeding of Center for Research on Bangla Language Processing.

[9]. Dalal Aniket. Kumar Nagraj. Sawant Uma. Shelke Sandeep and Bhattacharyya Pushpak, 2007. Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi. In Proceedings of International Conference on Natural Language Processing (ICON).

[10]. Ekbal A. and Mandal S. 2007. POS Tagging using HMM and Rule based Chunking. In Prceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages. IIIT Hyderabad, India.

[11]. Patel Chirag and Gali Karthik. 2008. Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. pp 117-122.

[12]. Singh Thoudam Doren and Bandyopadhyay Sivaji. 2008. Morphology Driven Manipuri POS Tagger. Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. pages 91–98. Hyderabad, India.

[13]. Dhanalakshmi V. Anandkumar M. Rajendran S and Soman K P. 2009. Tamil POS Tagging using Linear Programming. in proceeding of International Journal of Recent Trends in Engineering. Vol. 1. No. 2.

[14]. Ekbal Asif and Bandyopadhyay Shivaji. 2008. Web-based Bengali News Corpus for Lexicon Development and POS Tagging. In Proceeding of Language Resource and Evaluation.

[15]. Manju K. Soumya S. and Idicul S.M. 2009. A Development of A POS Tagger for Malayalam-An Experience. In Proceedings of International Conference on Advances in Recent Techno logies inCommunication and Computing.

[16]. J Antony P. And P Soman K. 2011. Parts of Speech Tagging for Indian Laguages: A Literature Survey. International Journal of Computer Applications. Vol. 34- No. 8.

[17]. Gaikwad Deepali K. Sawane Deepali and C. Namrata Mahender. 2017. Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer IOSR Journal of Computer Engineering (IOSR-JCE). Volume 3.pp 51-54.

[18]. Patil H.B. Patil A.S and Pawar B.V. 2014. Part-of-Speech Tagger for Marathi Language using Limited Training Corpora. International Journal of Computer Applications.

[19]. Bagul Pallavi. Mishra Archana. et.al. 2014. Rule Based POS Tagger for Marathi Text. (IJCSIT) International Journal of Computer Science and Information Technologies. Vol. 5 (2). 1322-1326.

[20]. Joshi Nisheeth. Darbari Hemant and Mathur Iti. 2013. HMM based POS Tagger for Hindi. In Proceeding of 2013 International Conference on Artificial Intelligence and Soft Computing.