Part of Speech TAGGER for MARATHI Language

Shubhangi Rathod*, Sharvari Govilkar** and Sagar Kulkarni***
*Department of Computer Engg. PIIT, New Panvel, India
rathod.shubhangi30@gmail.com
**Department of Computer Engg. PIIT, New Panvel, India
sgovilkar@mes.ac.in
***Department of Computer Engg. PIIT, New Panvel, India
skulkarni@mes.ac.in

Abstract: A part of speech (POS) tagging is one of the most well studied problems in the field of Natural Language Processing (NLP). Part-of-Speech Tagging (POS) tagging means assigning grammatical classes i.e. appropriate parts of speech tags like noun, adjective, verb, adverb etc to each word in a natural language sentence/word. The main challenge in POS tagging is to resolving the ambiguity in possible POS tags for a word so disambiguation rules and Tagset is vital parts of POS tagger. 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 find root word using morphological analyzer and compare the root word with corpus to assign appropriate tag. If word has assigned more than one tags then by using grammar rules ambiguity is removed. Meaningful rules are provided to improve the performance of the system.

Keywords: Part of Speech (POS), Tagset, Tokenizer, Stemmer, Morphological analyzer, Disambiguation.

Introduction

The work on Part-of-Speech (POS) tagging has begun in the early 1960s [2]. Part-of-Speech Tagging (POS) tagging means assigning grammatical classes i.e. appropriate parts of speech tags to each word in a natural language sentence/word. Assigning a POS tag to each word of an un-annotated text by hand is very time consuming, which results in the existence of various approaches to automate the job [3].The significance of these is the large amount of information they give about a word and its neighbours.

POS tagger is a necessary pre-processing module and extremely powerful as well as accurate tool [1] used in any application that deals with natural language processing. The tagging performance totally depends on tag dictionary. The large numbers of POS tagger available for English language which has got satisfactory performance but cannot be applied to Marathi language. Part-of-speech tagging in Marathi language is a very complex task as Marathi is highly inflectional in nature & morphologically rich language. The main challenge in POS tagging is to resolving the ambiguity in possible POS tags for a word [3].

Taggers can be classified as supervised or unsupervised: Supervised taggers are based on pre-tagged corpora, whereas unsupervised taggers automatically assign tags to words [6]. Furthermore, taggers divide into three types: (i) Rule Base Taggers: The rule based POS tagging approach that uses a set of hand constructed rules. (ii) Stochastic Taggers: A stochastic approach assigns a tag to word using frequency, probability or statistics [6]. It required vast stored contextual information because many high frequency words of POS are ambiguous. (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 [7].

Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken. This is not rare in natural languages such as Marathi language that a large percentage of word-forms are ambiguous. For example: In the sentence, **"पुजा देवं पुजा कर."** it is clear that the word **"पुजा"** is occurred two times in a sentence but the meaning of the word is different at both the places. The sentence given here contains ambiguity in the word which must be resolved before assigning tags to it. Designed system recognizes that the word **"पुजा"** has two different tags because of disambiguity rule for noun and verb. Thus the resultant tags assigned to the words in the sentence are

"पुजा::NNP देवं::NN पुजा::VM कर::VM .::RD_PUNC".

In these tags, the word **"पुजा"** has been assigned with two different tags, one act as 'Proper Noun' and other can be 'Verb Main'.

The paper presents, part of speech tagger for Marathi language. In section 2, related work is discussed in detail. Working of system is mentioned in detail in section 3. Section 4 explores accuracy obtained by POS tagger. Finally, paper is concluded in section 5.

Related Work

In this section we cite the relevant past literature that use the various pos tagging techniques. In the last few years the several approaches have been developed for English and other foreign languages. Most of the researchers concentrate on rule base rather than statistical approach for POS tagging. The small set of the meaningful rules of this tagger provides the better improvements over statistical tagger.

Jyoti Singh, et.al. [1] Proposed a Development of Marathi Part of Speech Tagger Using Statistical Approach. They used statistical tagger using Unigram, Bigram, Trigram and HMM Methods. To achieve higher accuracy they use set of Hand coded rules, it include frequency and probability. They use most frequently used tag for a specific word from the annotated training data and use this information to tag that word in the annotated text. They train and test their model by calculating frequency and probability of words of given corpus.

H.B. Patil, et.al. [2] Proposed a Part-of-Speech Tagger for Marathi Language using Limited Training Corpora. It is also a rule based technique. Here sentence taken as an input generated tokens. Once token generated apply the stemming process to remove all possible affix and reduce the word to stem. SRR used to convert stem word to root word. The root-words that are identified are then given to morphological analyzer. The morphological analysis is carried out by dictionary lookup and morpheme analysis rules.

Pallavi Bagul, et.al. [3] Proposed a Rule Based POS Tagger for Marathi Text. Which will assign part of speech to the words in a sentence given as an input and used a corpus which is based on tourism domain. 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. The ambiguity is resolved using Marathi grammar rules.

Jyoti Singh, et.al. [4] Proposed a Part of speech tagging of Marathi text using Trigram method. The main concept of Trigram is to explore the most likely POS for a token based on given information of previous two tags by calculating the transition probabilities between the tags and helps to capture the context of the sentence. The probability of a sequence is just the product of conditional probabilities of its trigrams. Each tag transition probability is computed by calculating the frequency count of two tags which come together in the corpus divided by the frequency count of the previous two tags coming in the corpus.

Nidhi Mishra, et.al. [5] Proposed Part of Speech Tagging for Hindi Corpus. The system scans the Hindi (Unicode) corpus and then extracts the Sentences and words from the given Hindi corpus. Finally Display the tag of each Hindi word like noun tag, adjective tag, number tag, verb tag etc. and search tag pattern from database.

Namrata Tapaswi, Suresh Jain [6] proposed a Treebank Based Deep Grammar Acquisition and Part-Of-Speech Tagging for Sanskrit Sentences. In the Sanskrit morphology meaning of the word is remain same. When affixes are added to the stem, words are differentiated at database level directly. The input is one sentence per line, split the sentence into words called lexeme .read each word to find longest suffix, and eliminated the suffix until the word length is 2. Apply the lexical rules and assign the tag. Remove the disambiguity using context sensitive rules.

Javed Ahmed MAHAR, Ghulam Qadir MEMON [7], proposed a system for “Rule Based Part of Speech Tagging of Sindhi Language”. Take input text, and generate token. Once token generated search and compare selected word from lexicon (SWL) .If word is found one or more times, then store associated tag and if not found add that word into lexicon by generating linguistic rule for new word.

Proposed System

We have designed 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. Rule-based part-of-speech tagging is the most powerful approach that uses manually written rules for tagging. Rule based tagger 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. The proposed approach consists of following phases:

1. Pre-processing
2. Stemmer
3. Morphological analyzer.
4. Tag Generator
5. Disambiguation.

Preprocessing

Validation of Input document

The input document may contain some words or sentences in other script or language. So, validation of Input document is very important stage because the resultant information is totally depends on the language and nature of query supplied to the system. Here we are analyzing whether the input document is valid in Devanagari script or not. The words which are not valid to Devanagari script are simply removed from further processing. To perform this operation we have used Unicode values called UTF-8 for Devanagari script document. The aim of this phase is to maintain pure Devanagari script document as an input to Morphological Analyzer.

Tokenization

This Tokenization is the process of separating word/tokens from input text. The division of input text into tokens is important for POS tagging. This tokenization task is possible by searching spaces between the words. The words separated from sentence and treat as single token so, we can deal with each word separately.

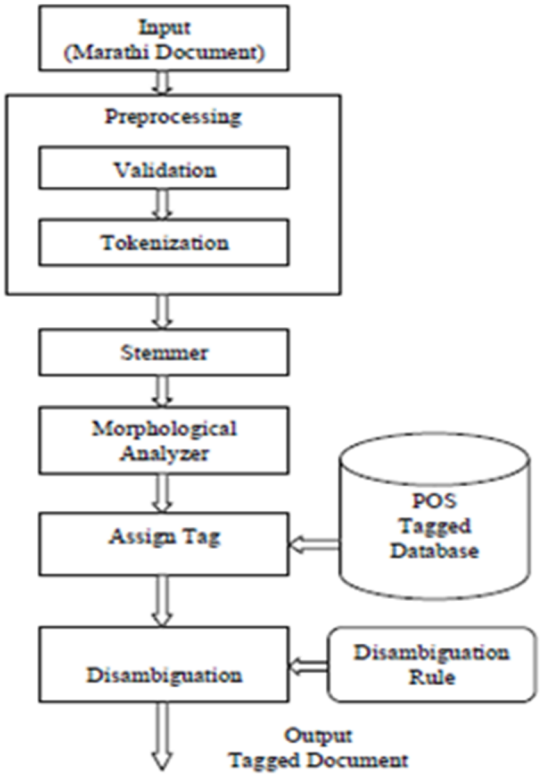

Fig 1. Proposed System

Stemmer

Stemming is important in the system, which uses a suffix list to remove suffixes from words and thus reduces the word to its stem. To remove suffixes from input document the Corpus is used consist of 1059 suffixes which frequently occur in Marathi language. The result of stemming is stem of word that can be given as input to Morphological Analyzer for further processing. The stem word contains inflections. The inflections in the stem word cannot be removed using simple stemming operation.

Morphological analyzer

The aim of morphological analysis is to recognize the inner structure of the word. The words after stemming are analyzed to check whether they are inflected or not. If stem word is inflected 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. There is need to design some standard rules called inflection rule which will enable the system to process the stem of words and find the actual Root word.

Tag Generator

Corpus linguistics is the study of language as expressed in samples (corpora) of "real world" text. Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based [9]. This phase assigns corresponding part of speech tags to the words and we have used tagset developed by IIIT Hyderabad [9] [10]. A well-chosen tagset is important to represents parts of speech. The language tagset represents parts of speech and consist on syntactic classes [8].

Algorithm for POS tagging System:

1) Take input text and generate a token.
2) Use tokens to generate stem of word.
3) Use rule to generate root word using morphological analyzer and stored them.
4) Select each word one by one and compare with corpus.
5) If word is found one or more times, then store associated tag or tags of word and else display "the word is not found" add this new word into corpus.
6) If one tag is stored, then display word with associated tag as an output.
7) Else apply rule to select most appropriate tag for word

Table 1. POS Tag list

No.	Name	Tag	Description	Example
1	NOUN	NN	Common Nouns	मुलगा, साखर, मंडळी, चांगुलपणा
		NNP	Proper Nouns (name of person)	मोहन, राम, सुरेश
		ABN	Abstract noun	गर्व,कौशल्य, क्रोध,चपळाई
2	PRONOUN	PPN	Personal pronoun	मी,आम्ही,तुम्ही
		PPS	Possessive pronoun	माझा,माझी, तुझा,तुझी,त्याचा
		PDM	Demonstrative pronoun	तो, ती, ते, हा, ही
		PRF	Reflexive pronoun	आपण,आम्ही, तुम्ही,तुम्हाला
		PRC	Reciprocal pronoun	एकमेकांचा, एकमेकाला
3	ADJECTIVE	JJ	Modifier of Noun	उत्साही, श्रेष्ठ,बळवान
4	VERB	VM	Verb Main (Finite or infinite)	बसणे, दिसणे, लिहिणे,पडला
		VAUX	Verb Auxiliary	नाही, नको, करणे,हवे, नये
5	ADVERB	RB	(Modifier of Verb)	आता, काल, कधी, नेहमी, लवकर
6	CONJUNCTION	CC	Coordinating and Subordinating	आणि,पण, जर, तर
7	POSTPOSITION	PSP	Postposition	आणि, वर, कडे,जवळ
8	INTERJECTION	INJ	Interjection	आहा, छान, अगो, हाय

9	NUMERAL(NUM)	NUM	Number	१,२,३,४
		NUMCD	Cardinal Numeral	एक, दोन, तीन
		NUMO	Ordinal Numeral	पहिला,दु सरा, तिसरा
10	RESIDUAL	RDS	Symbol residual	$, &, *, (,)
		RD_PUNC	Punctuation	? , ; : !
11	REDUPLICATION	RDP	Reduplications	जवळजवळ-
12	NEGATIVE	NEG	Negative	नाही,नको
13	DETERMINER	QF	Quantifiers	किती,पुष्कळ,खूप,भरपूर, बरेच
14	QUESTION WORDS	WQ	Question Words	काय, कधी, कु ठे
15	INTENSIFIER	INTF	Intensifier	खूप,फार,बराच,अतिशय
16	PARTICLES	RP	Particles	तर,ओहो
17	PHRASE	PHR	Phrase	नमस्कार, अभिनंदन,खेद आहे
18	ECHO	ECH	Echo Word	जेवणबिवण, डोकेबिके
19	QUATATIVE	UT	Quatative word	म्हणजे

Disambiguation

The Natural language has the ambiguity issues as the single word has different tags. To overcome the ambiguity issues and assigning a "correct" tag in particular contexts, disambiguation rules are required. Word Sense Disambiguation (WSD) is the process of identifying the sense of a polysemic word. In modern WSD systems, the senses of a word are typically taken from some specified dictionary. Disambiguation is based on contextual information or word/tag sequences. The ambiguity which is identified in the tagging module is resolved using the Marathi grammar rules.

Following example demonstrates processing of our system:

Input Query: मराठी भाषा हे महाराष्ट्राचे वैभव आहे. It is given in history of Marathi Language

Validation: मराठी भाषा हे महाराष्ट्राचे वैभव आहे.

Tokenization:

मराठी
भाषा
हे
महाराष्ट्राचे
वैभव
आहे
.

Stemmer: मराठी भाषा हे महाराष्ट्रा वैभव आहे.

Morphological Analyzer: मराठी भाषा हे महाराष्ट्र वैभव आहे.

POS Tagging Output: मराठी\\NNP भाषा\\NN हे\\PDM महाराष्ट्र\\NNP वैभव\\JJ आहे\\VAUX . \\RD_PUNC

Performance of System

We have developed our own corpus consisting of 17197 unique words, tagset consist 29 tags and we have developed 141 rules for disambiguation for Marathi languages. The performance of the system is measured for multiple documents as shown in Table. We have used randomly selected Marathi document as input to NLTK and our designed Tagger. While recording correctness of both the system we focused on the strength of both taggers to handle WSD of the words in the sentences. Many times we found that our designed Tagger performs well for both Tagging and handling WSD as compared with NLTK tagger. The designed tagger system is compared with other existing systems such as NLTK and Shallow Parser. The Table 5.1 shows the details of the testing results for ten Marathi language documents. It is found that the efficiency of designed POS tagger to assign correct tags to words in the document is better than that of NLTK and Shallow Parser.

Table 2. Table Performance of the System based on Document analysis

Doc. No.	Document Name	No. of words	Words correctly tagged by NLTK	Performance of NLTK (%)	Words correctly tagged by Shallow Parser	Performance of Shallow Parser (%)	Words correctly tagged by Designed Tagger	Performance of Designed Tagger (%)
1	Marathi Bhasha	106	62	58.49	73	68.87	102	96.23
2	Disambiguity Text	61	18	29.51	36	59.02	59	96.72
3	Agriculture text	130	64	49.23	99	76.15	127	97.69
4	Ramayana	44	20	45.45	34	77.27	42	95.45
5	Shivaji Maharaj	123	74	60.16	94	76.42	114	92.68
6	Panvel info	87	38	43.68	64	73.56	85	97.70
7	Aai text	58	33	56.90	46	79.31	53	91.38
8	Marathi Grammar text	101	38	37.62	70	69.31	95	94.06
9	Modi	101	58	57.43	70	69.31	96	95.05
10	Mumbai info	77	47	61.04	53	68.83	72	93.51

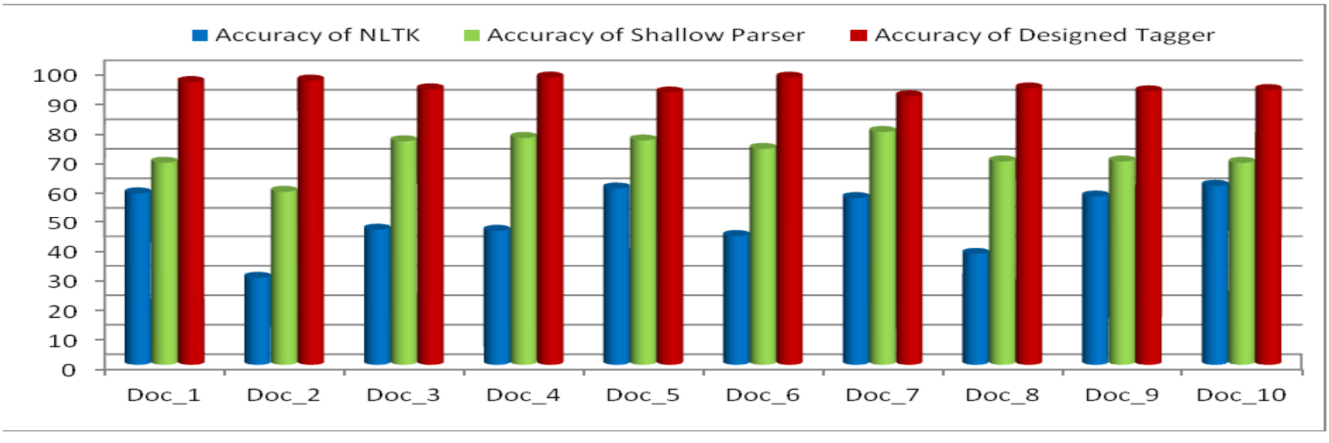

Figure 2. Graphical representation of performance analysis of Designed Tagger

The Figure shows the graphical representation of performance of the designed Tagger, NLTK and Shallow Parser. It is clearly observed that the designed tagger system gives higher performance result to tag the words correctly in the document than NLTK and Shallow Parser.

Table 3. System Testing based on Collection of Documents testing based on Collection of Documents

Overall Analysis on	Total No of Words	NLTK		SHALLOW PARSER		DESIGNED TAGGER SYSTM	
		Correctly tagged words	Accuracy (%)	Correctly tagged words	Accuracy (%)	Correctly tagged words	Accuracy (%)
Collection of 10 Documents	811	347	49.95	516	71.81	677	95.05

We have taken ten randomly selected Marathi documents to analyze the result of our designed tagger and it is compared with the existing systems such as NLTK and Shallow Parser. As shown in Table it can be easily observed that out of total number words (811) the designed POS tagger system gives efficiency of 95.05% i.e. it tags 677 words correctly, whereas NLTK and Shallow Parser gives the efficiency up to 49.95% and 71.81% respectively.

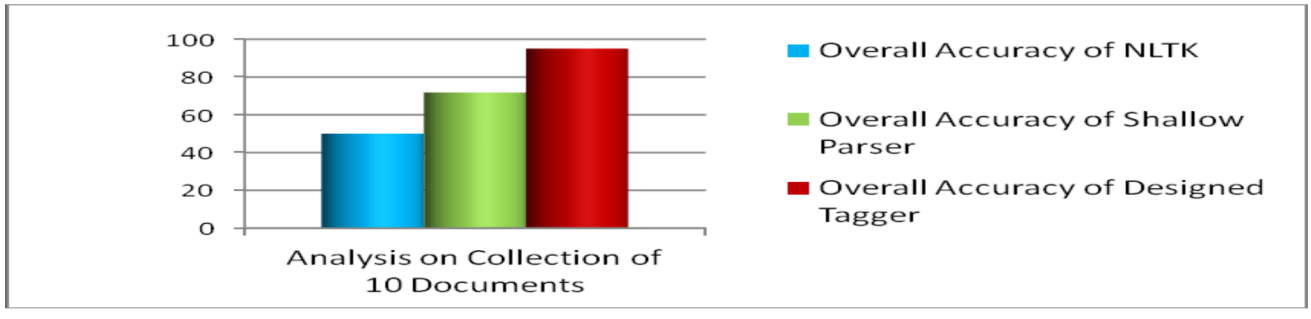

Figure 3. Graphical representation of testing based on collection of documents

The Figure shows that the efficiency of the designed POS tagger is higher than that of NLTK and Shallow Parser. The overall accuracy of the system is 95.05%.

Conclusion

The task of POS tagging is quite complex for Marathi language as the language is morphologically rich in script. There are some issues still present in tagging the words effectively because if stemming and morphology is not performed well then the root form is not generated correctly and thus the tag assigned to such incorrect words are not always correct. The POS tagger we designed for Marathi language uses Rule-based tagging approach which assigns all possible tags to word and WSD uses context rules to disambiguate the tags so that the accuracy is enhanced. It has been proved that the designed POS tagger for Marathi language gives relevant and acceptable performance up to **95.05%**.There is large scope to enhance the rule set of POS tagging as well as WSD context rules to improve the accuracy of the designed system up to maximum extent.

Acknowledgment

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

References

[1] Jyoti Singh, Nisheeth Joshi, Iti Mathur, "Development of Marathi Part of Speech Tagger Using Statistical Approach", international conference on Advance in Computing, Communications and Informatics (ICACCI), IEEE DOI 10.1109/ICACCI.2013.6637411,2013,1554-1559.

[2] H.B. Patil, A.S. Patil, B.V. Pawar "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora", International Journal of Computer Applications (0975 –8887) Recent Advances in Information Technology,2014.

[3] Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar, "Rule Based POS Tagger for Marathi Text" 2014 in proceeding of: International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1322-1326.
[4] Jyoti Singh, Nisheeth Joshi, Iti Mathur "PART OF SPEECH TAGGING OF MARATHI TEXT USING TRIGRAM METHOD", 2013 International Journal of Advanced Information Technology (IJAIT) Vol. 3, No.2, DOI: 10.5121/ijait2013.3203.
[5] Nidhi Mishra, Amit Mishra,"Part of Speech Tagging for Hindi Corpus" 2011 in proceeding of : International Conference on Communication Systems and Network Technologies, 978-0-7695-44373/11, 2011 IEEE DOI 10.1109/CSNT.2011.118.
[6] Namrata Tapaswi, Suresh Jain, "Treebank Based Deep Grammar Acquisition and Part- Of-Speech Tagging for Sanskrit Sentences", Software Engineering (CONSEG), 2012 CSI Sixth International Conference on, 971-1-4673-2174-7 IEEE DOI 10.1109/CONSEG.2012.6349476.
[7] Javed Ahmed MAHAR, Ghulam Qadir MEMON, "Rule Based Part of Speech Tagging of Sindhi Language " 2010 proceeding of International Conference on Signal Acquisition and Processing 978-0-7695-3960-7/10,2010 IEEE DOI 10.1109/ICSAP.2010.27.
[8] Sankaran Baskaran , Kalika Bali1, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, Rajendran S.5, Saravanan K.1, Sobha L.6, and KVS Subbarao, "A Common Parts-of-Speech Tagset Framework for Indian Languages".
[9] Kh Raju Singha Bipul Syam Purkayastha Kh Dhiren Singha "Part of Speech Tagging in Manipuri: A Rule-based Approach "International Journal of Computer Applications, (0975 – 8887) Volume 51– No.14, August 2012
[10] Bharati, A., Sharma, D.M., Bai, L., Sangal, R., "AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages",2006.