Part of Speech Tagger for Marathi Language

Sharvari Govilkar
Department of Information Technology
TSEC, Bandra, Mumbai, India

Bakal J. W
SJCOE , Dombivli, Thane
India

Shubhangi Rathod
Department of Computer Engineering
PIIT, New Panvel, India

ABSTRACT

A part of speech (POS) tagging is one of the most well studied problem in the field of Natural Language Processing (NLP). A POS Tagger is the process of assigning correct tag like noun, adjective, verb, adverb etc to each word of the input sentence. 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.

General Terms

Part of Speech, Marathi

Keywords

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

1. INTRODUCTION

A part of speech (POS) tagging is one of the important problem in the field of Natural Language Processing (NLP) that has begun in the early 1960s. A POS Tagging is the process of assigning exact tag like noun, verb, adverb, adjective etc 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. It is an extremely powerful and accurate tool [1] used in any application that deals with natural language processing. 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 [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].

POS tagging is a necessary pre-module and building block for various NLP application. POS tagging is difficult for Marathi language due to unavailability of corpus for computational processing. The rule based part of speech tagger that assigns all possible tags to word 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 search words in corpus and second, it assigns a tag by resolving ambiguity. The small set of meaningful rules helps to remove disambiguity of words. The paper presents a rule based part of speech tagger for Marathi language. Related work and past literature is discussed in section 2. Basic working of POS tagger is discussed in section 3. Finally, paper concludes in section 4.

2. LITERATURE SURVEY

In this section we cite the relevant past literature that use the various pos tagging techniques. 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.

3. 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. If word has more than one tag then it is an instance of ambiguity , so such a word can be disambiguate by using small set of context rule. Finally word is displayed with single associated tag as an output. The rule-based tagger which we are going to develop has many advantages over other taggers. The proposed approach consists of following phases:

1. Preprocessing

2. Stemmer

3. Morphological analyzer.

4. Tag Generator

5. Disambiguation.

3.1 Preprocessing

3.1.1 Validation of Input document:

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. The input document may contain some words or sentences in other script or language. 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. The aim of this phase is to maintain pure Devanagari script document as an input to Morphological Analyzer.

3.1.2 Tokenization:

This Tokenization is the process of separating 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. Information retrieval applications requires list of root words for the purpose of easy searching, maintain index of words, calculating term frequency of words etc.

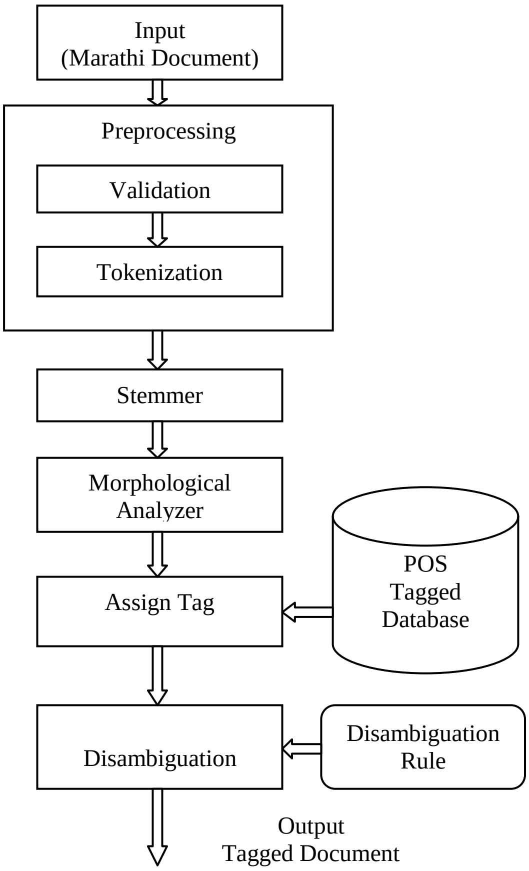

Figure 1. Proposed System

3.2 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. 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. In Marathi language the infected words are the words, which belong to Noun, Pronoun, Adjective, or Verb. So the Suffix Replacements Rules (SRRs) are used for these categories words only.

3.3. 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.

3.4. 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 from array 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.

3.5. 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. 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\\NN भाषा\\NN हे\\NNP महाराष्ट्र\\NN वैभव\\JJ आहे\\VM . \\RD_PUNC

Table 1. Tagset

Name	**Tag**	**Description**	**Example**
NOUN	NN	Common Nouns	मुलगा, साखर
	NNP	Proper Nouns	मोहन, राम, सुरेश
	NV	Verbal noun	समर्पण, कुरवंडी
	NST	Noun Denoting Spatial and Temporal Expressions	मागे, पुढे, वर, खाली
PRONOUN	PPN	Personal pronoun	मी, आम्ही, तुम्ही
	PPS	Possessive pronoun	माझा, माझी, तुझा
	PDM/DEM	Demonstrative pronoun	तो, ती, ते, हा, ही
	PRF	Reflexive pronoun	आपण, आम्ही, तुम्ही, तुम्हाला
	PRC	Reciprocal pronoun	एकमेकांचा, एकमेकाला, आपल्याला
	PIN	Interrogative pronoun	काय,कसे,कोण,का
ADJECTIVE	JJ	Modifier of Noun	उत्साही, श्रेष्ठ,बळवान
	MQ	Quantifier	किती,पुष्कळ,खूप,भरपूर,चिक्कार
VERB	VM	Verb Main	बसणे, दिसणे, लिहिणे,पडला,
	VAX	Verb Auxiliary	नाही, नको, करणे, हवे, नये
ADVERB	ADV/RB	(Modifier of Verb)	आता, काल, कधी, नेहमी, लवकर
CONJUNCTION	CC	Coordinating and Subordinating	आणि, पण, जर, तर
POSTPOSITION	PSP	Postposition	आणि, वर, कडे, जवळ
INTERJECTION	INJ	Interjection	आहा, छान, अगो

NUMERAL	NUMCD	Cardinal Numeral	एक, दोन, तीन
	NUMO	Ordinal Numeral	पहिला,दुसरा
RESIDUAL	RDF	Foreign word residual	English, गुजराती
	RDS	Symbol residual	$, &, *, (,)
	RD_PUNC	Punctuation	? , ; : !
REDUPLICATION	RDP	Reduplications	जवळ-जवळ
COMPOUNDS	XC	Compounds	काळेमांजर- काळमांजर, तेलपाणी- तेलपाणी
NEGATIVE	NEG	Negative	नाही,नको
DEMONSTRATIVE	QF	Quantifiers	बहुत, थोडा, कम
QUESTION WORDS	WQ	Question Words	काय, कधी, कु ठे
INTENSIFIER	INTF	Intensifier	खूप, फार,बराच,अतिशय
PARTICLES	RP	Particles	तर,ओहो
PHRASE	PHR	Phrase	नमस्कार, अभिनंदन,खेद आहे
ECHO	ECH	Echo Word	जेवणबिवण, डोकेबिके,पडणबिडण

4. CONCLUSION

POS tagger is very useful in the context of Information Extraction and Question Answering systems. A Rule-based POS tagger simply assigns all possible tags to word and then uses context rules to remove disambiguation associated with words. Rules based POS tagger would yield better accuracy when there is large set of meaningful rules to remove disambiguation. Our proposed approach handles large set of rules for Marathi POS tagger.

5. REFERENCES

[1] Jyoti Singh Nisheeth Joshi Iti Mathur "Development of Marathi Part of Speech Tagger Using Statistical Approach", Advances in Computing, Communications and Informatics (ICACCI), 2013.

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

[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.

[5] Nidhi Mishra, Amit Mishra, "Part of Speech Tagging for Hindi Corpus" 2011 in proceeding of : International Conference on Communication Systems and Network Technologies.

[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.

[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.

[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).