2018 International Conference on Communication, Information & Computing Technology (ICCICT), Feb. 2-3, Mumbai, India

A Hybrid Part-of-Speech Tagger for Marathi Sentences

Madhuri M. Deshpande
Department of Computer Science, Savitribai Phule Pune University, Pune, India.
madhuri.deshp@gmail.com

Dr. Sharad D. Gore
Ex-Professor and Ex-Head, Department of Statistics, Savitribai Phule Pune University, Pune, India.
sharaddgore@gmail.com

***Abstract*— With thousands of languages in the world, and the increasing speed and quantity of information being distributed across the world, automatic translation between languages by computers, Machine Translation, has become an increasingly important area of research. For a machine to translate text in one natural language to target text in another language, it requires an understanding of the language, its grammar (syntax), its meaning (semantics) and the ability to use this knowledge for making inferences. Words have definite meaning(s) making them deterministic and finite. Words are not ambiguous in their meaning. Context dependency arises when a word is used with a group of words (bag-of-words) in a specific way that causes its meaning to be dependent on the group of words. In this paper, we present a hybrid, multi-pass Part-Of-Speech (POS) tagger developed for Marathi sentences which builds feature vector for each word in a sentence by referring to the previous and next word preceding and succeeding the current word that is being tagged. The analysis of the Marathi input sentence is done first by tokenizing each word in the sentence and finding the *stem* for each token. Every token is analyzed for its POS tag, the tense, mood and aspect. This process is *POS tagging*. Ambiguities may arise in the process of tagging.**

Keywords*— *POS Ambiguity, Machine Translation, Marathi Grammar rules, Part-of-Speech (POS), Root/Stem of a Word, Stemming, Tagger, YASS.

I. INTRODUCTION

POS Tagging is "the process of associating parts of speech to each word in a sentence"[8]. Every word in a sentence will be assigned an unambiguous grammar tag such as noun, verb, adjective, and so on. The granularity of POS tags depends on the application. In the process of POS tagging, the tagger needs to remove the inflections and derive the root word or the stem. The process of extracting the stem by separating the inflections is called stemming. Separating the inflection from a word is important to retain the correct form of the word in Marathi. Words can be classified in two types depending on the Inflection as:

Inflectional Words: Noun, Pronoun, Adjective, Verb
Non-Inflectional Words: Adverb, Preposition, Interjection, Conjunction

The words are inflected on the basis of changing Gender (Masculine, Feminine, Neuter), Multiplicity,(Singular, Plural), Tense (Present, Past, Future), and Case (Nominative, Accusative, Instrumental, Dative, Ablative, Genitive, Locative, Vocative).

1) Noun Inflections in Marathi

Inflections with Nouns in Marathi tell us the gender, number and case. The inflection of a word can be determined from the word endings called Vibhakti pratayay (विभक्ती प्रत्यय). Table 1 describes the word endings and its inflections.

Table 1: Case Terminations for Nouns [23]

Case (**विभक्ती**)	Singular(**एकवचन**) Suffix(**प्रत्यय**)	Plural (**अनेकवचन**) Suffix(**प्रत्यय**)
प्रथमा (Nominative)	-------	-------
द्वितीया (Accusative)	**स, ला, ते**	**स, ला, ना, ते**
तृतीया (Instrumental)	**ने, ए, शी**	**नी, शी, ई, ही**
चतुर्थी (Dative)	**स, ला, ते**	**स, ला, ना, ते**
पंचमी (Ablative)	**ऊन, हून**	**ऊन, हून**
षष्ठी (Genitive)	**चा, ची, चे**	**चे, च्या, ची**
सप्तमी (Locative)	**त, ई, आ**	**त, ई, आ**
संबोधन (Vocative)	-------	**नो**

2) Verb Inflections in Marathi

Inflections with Verbs in Marathi express person, number and gender of the subject alone or that of both the subject the object of the verb. Inflected verbs also express tense, aspect and mood. When we translate the verb using Marathi Dictionary we get the gerundial form that is, it is given with the particle 'णे', for example, पाहणे – to see. For inflecting the verb, first we need to derive the verbal root (also known as धातू - Dhatu) and then add personal endings to it, to indicate its relation to the noun. We can get verbal roots by dropping the particle 'णे' from the gerundial form (example, पाह from पाहणे, आवड from आवडणे and so on). Inflection of the verb

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

depends upon the gender (लिंग) (Masculine, Feminine and Neuter), the number (वचन) (Singular, Plural), Person (पुरुष)(First, Second and Third) and tenses (काळ) (Present, Past and Future). Sometimes personal endings may also depend on moods (अर्थ), the constructions (प्रयोग), the participle and the verbal nouns (धातुसाधीते). The table for rules of verb inflection is given in Table 2 (Pre(व) indicates Present tense, Pt (भु) indicates Past tense and F (भ) indicates Future Tense).

Table 2: Rules for verb inflection [21]

पुरुष (Person)	पुल्लिंगी (Masculine)						स्त्रीलिंगी (Feminine)						नपुसक (Neutral)					
	एकवचन Singular			अनेकवचन Plural			एकवचन Singular			अनेकवचन Plural			एकवचन Singular			अनेकवचन Plural		
	काळ (Tense)			काळ (Tense)			काळ (Tense)			काळ (Tense)			काळ (Tense)			काळ (Tense)		
	व Pre	भु Pt	भ F	व Pre	भु Pt	भ F	व Pre	भु Pt	भ F	व Pre	भु Pt	भ F	व Pre	भु Pt	भ F	व Pre	भु Pt	भ F
प्रथम (First)	तो	लो	ईन एन	तो	लो	ऊ	ते	ले	ईन एन	तो	लो	ऊ	--	--	--	--	--	--
द्वितीय (Second)	तोस	लास लीस लेस ल्यास	शील	ता	लात लीत लेत ल्यात	आल	तेस	लास लीस लेस ल्यास	शील	ता	लात लीत लेत ल्यात	आल	--	--	--	--	--	--
तृतीय (Third)	तो	ला	ईल एल	तात	ले	तील	ते	ली	ईल एल	तात	ल्या	तील	ते	ले	ईल एल	तात	ली	ती ल

3) Adjective inflections in Marathi

Inflections associated with adjectives express the gender and number and this inflection depends upon gender, multiplicity, attachment of postpositions to the noun modified by such objective. When genitive case makers or some prepositions are attached to nouns, it produces an adjective.

4) Pronoun inflections in Marathi

Inflected pronouns give information about gender, number, case and person. A pronoun (such as तो, ती, ते, जो, जी, त्यांनी, आम्ही and many more) is a word that can be substituted in place of a noun. Pronoun inflection is similar to noun inflection but there are some special cases which need to handle separately. Table 3 gives a list of special cases of inflections in pronouns. Some sentences have the same structure along with the parse tree, gender, multiplicity, cases but have different translation of pronoun. Following cases gives a comparison of different cases where the same pronoun is used but has different inflections depending upon the verbs used in the sentence. In the sentence 'मी रोज चालते' the verb (चालते) indicates the gender of the pronoun (मी) as feminine.

However, in the succeeding sentences the gender is unknown.

1) मला पेरु आवडतो - The Verb (आवडतो) does not indicate the gender of the pronoun (मी) since, the pronoun can be masculine or feminine.

2) मला चिंच आवडते - The verb (आवडते) does not indicate the gender of the pronoun (मला) since, the pronoun can be masculine or feminine.

3) त्यांना पेरू आवडतात - Verb (आवडणे) does not indicate the gender of the pronoun (त्यांना).

Table 3: Special cases of Pronoun Inflections

Marathi Pronoun	Number	Corresponding English word
मला / माझ्या	Singular	I/Me
आम्ही / आम्हाला	Plural	We/Us
तुला	Singular	You
तुम्हाला	Plural	You
त्याला	Singular	Him
तिला	Singular	Her
त्यांना	Plural	They

4) The Lookup Lists used by the Hybrid Marathi tagger

The tagger developed and used in this study uses a number of lookup lists. The first one is the list of Marathi stopwords, the second is a suffix list (विभक्ती), list of pronouns (सर्वनाम), list of adjectival pronouns (सर्वनामिक विशेषण), a list of conjunctions (उभयान्वयी अव्यये). a list of prepositions (शब्दयोगी अव्यये), and a list of exclamation words (केवलप्रयोगी अव्यये).

The suffix list contains a list of suffixes with feature values like gender, number, person and other relevant morphological information. It deals only with inflectional suffixes not derivational. Stem does not necessarily correspond to linguistic root of a word. Stemming improves performance by reducing morphologically variants into same words. The tags set used in this study is shown in table 4.

II. The Design of the Hybrid POS Tagger

The tagger used in this study uses a variant of the "Yet Another Suffix Stripper" (YASS) stemmer that uses Marathi grammar rules for disambiguation of POS tags. The components of the tagger are depicted in figure 1. The YASS distance measures used by the tagger are D_1, D_2, D_3 and D_4 [17]. Majumder et al. consider two strings, X and Y, each of

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

length n and n' respectively, for which distance is to be computed. They use a Boolean function p_i (for penalty) which is used to compute the distances, D_1, D_2, D_3 and D_4. For any two strings $X = x_0\, x_1 \ldots x_n$ and $Y = y_0\, y_1 \ldots y_{n'}$, they define the penalty function p_i (a Boolean function) as follows:

$$p_i = \begin{cases} 0 & \text{if } x_i = y_i \quad 0 \le i \le \min(n, n') \\ 1 & \text{otherwise} \end{cases}$$

Table 4: POS tags used by the Hybrid Marathi POS tagger

Category	POS Tag	Category	POS Tag
Noun (NN)	NNP	Pronouns	PRP
	NNC	Adverb	RB
	NNPC	Conjunction	CC
	NJJ		DEM
Adjective (JJ)	JJ	Interjection	UH
	INTF	Preposition	PREP
	QFNUM	Question Word	QW
Verb (VRB)	VFM	Symbol	SYM
	VJJ	Unknown	UNK
	VAUX		
	VRB		
	VNN		

If there is a mismatch in the i^{th} position of X and Y then p_i is 1. The distance D_1 is defined as follows:

$$D_1(X, Y) = \sum_{i=0}^{n} \frac{1}{2^i} p_i$$

The distances D_2, D_3, D_4 are defined as follows [17]:

$$D_2(X, Y) = \frac{1}{m} \times \sum_{i=m}^{n} \frac{1}{2^{i-m}} \quad \text{if } m > 0, \quad \infty \text{ otherwise}$$

$$D_3(X, Y) = \frac{n-m+1}{m} \times \sum_{i=m}^{n} \frac{1}{2^{i-m}} \quad \text{if } m > 0, \quad \infty \text{ otherwise}$$

$$D_4(X, Y) = \frac{n-m+1}{n+1} \times \sum_{i=m}^{n} \frac{1}{2^{i-m}}$$

We use the longest subsequence match to stem the word.

In this study, we have developed a hybrid tagger that is a multi-pass stemmer with a look-ahead and look-back of one word. This look-ahead and look-back for the current word causes the removal of ambiguity in the POS tag for that word, in most of the cases. The longest-sequence match is computed using the string distance measures defined by Majumder et al.[17]. It introduces the concept longest sub-sequence match and penalizes mismatches. Disambiguation rules, in the form of Marathi grammar rules (refer Table 5), are applied when the stemmer looks back or ahead. Marathi language specific rules have been encoded that are used to disambiguate the POS tags assigned to a word. This rule-base has also been used due to the unavailability of exhaustive dictionary and corpus of Marathi.

III. The Working of the Tagger

The tagger designed and being used in this study is multi-pass. In the first pass, the tagger removes suffix (inflections) associated with the word using the suffix list provided to it. A word in Marathi is represented as a collection of UNICODE

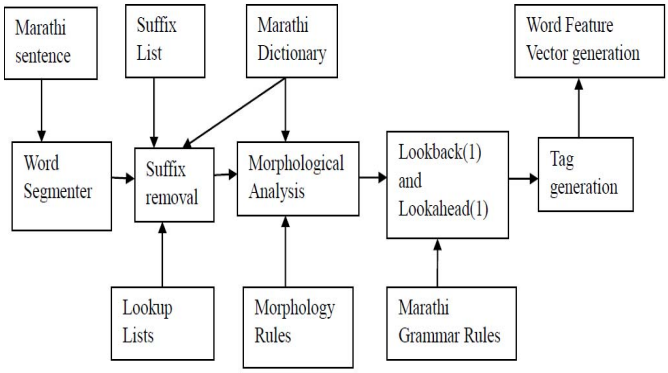

Figure 1: The tagger Architecture

characters. The tagger code, written in Java, processes Marathi sentences that are input in Devnagri script. In addition, we have made substantial use of Marathi grammar - grammar rules and information about suffixes. Marathi Dictionary is implemented in mySQL (using the Devnagri character representation - UTF- 8 format).

The word is scanned from right to left, to obtain the suffix. This suffix is removed from the word. When the suffix is removed from a word we get the सामान्य रूप (samanya roop/stem) for the word. For example for the word शिक्षकाने the tagger identifies the suffix as ने and शिक्षका as the stem. The root of the word शिक्षका is obtained using the dictionary lookup and using longest subsequence match. The suffix ने gives the feature of the word namely the case (विभक्ती) as third person (तृतीया), gender(लिंग) as Masculine (पुल्लिंग) and number(वचन) as singular (एकवचन) with root word शिक्षक. Consider the word भावाला, the tagger identifies the suffix as ला and the stem as भावा. The stem भावा matched the word भाव (meaning हावभाव- expression or दर-rate), since we are using the longest subsequence match and a dictionary lookup. The stemmer then, tags the word भावाला as Masculine, Second Person, Singular. Though a part of the feature vector for

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

भावाला is correct, the root word भाव is incorrect, it needs to be भाऊ.

Table 5: Some Marathi Grammar rules

Rule No.	Rule Description / Pattern
1	S→ NP^+ + VP (Noun Phrase- NP and Verb Phrase VP). A general rule, not used by the tagger.
2	S → NNPC + NNP + [NNC/NNP/VNN]* + [VFM/ VRB] (NNPC will have no inflections, last proper noun will be inflected in case of compounded proper nouns). Example: **विवेकानंदांनी**/NNP **देशातील**/NNC **अज्ञान**/NNC **घालवण्याला**/VNN **ईश्वर**/NNC **सेवा**/NNC **मानले**/VFM **होते**/VAUX
3	S → NNC^+ + [VFM / VRB] (**मुलगा रडतो / मुलगा पुस्तक वाचतो**)
4	S → PRP + [VFM /VRB] (**तू जा**)
5	S → [NNC / NNP / DEM]* JJ + NNC+ + [VFM / VRB] Example: **बबनचा** /NNC **पहिलाच**/JJ **बंगला**/NNC **आहे**/VRB, **सुंदर**/JJ **मुले**/NNC **सर्वांना आवडतात**/VRB, **तो**/DEM **जुना**/JJ **वाडा**/NNC **आहे**/VRB
6	S → PRP + NNC + [VFM / VRB] (**मी शाळेतून आले**)
7	S → PRP + RB +NNC + [VFM / VRB] (**तू आता**/RB **घरी जा**)
8	S → [NNC / NNP] + RB + [VFM/VRB] (**मुलगा जोरात रडतो**)
9	S → RB + NNC^+ + VRB (**काल**/RB **शाळेला सुट्टी होती**)
10	NP → [NNPC+ NNP/NNP] (**सर्वपल्ली राधाकृष्णन**) / NNP NNC^+ (**सीमा बटाटे सोलत**/VNN **होती**/VFM) / NNP UH (**पुणे येथे**) NNP → DEM + NNP^+ + NNC (**त्याने विठ्ठलाच्या पायी**)
11	NNC-> NNC^+ / PRP NNC+ (**त्याचे पुस्तक**) / DEM JJ NNC (**त्या सुंदर फुलाचे**) / PRP JJ NNC^+ (**ती चपळ मुलगी**)//
12	VP -> RB + VFM / RB + VRB / VNN VFM

The feature vectors, which are built for each word in the sentence, are modified as the tagger progress to the next pass. From the second pass onwards, the tagger uses a lookahead and lookback of one word along with disambiguation rules, to tag a word, until each word is uniquely tagged. The algorithm of the tagger is explained in figure 2. An example of the intermediate working of the tagger is shown immediately after the algorithm in table 7.

IV. Handling Pronoun-Adjective Ambiguity by the Hybrid Tagger.

Pronouns such मी, आम्ही, तू, तुम्ही, तो, ती, ते, जो, जी, जे, हा, ही, हे, कोण, काय ideally occur instead of nouns. However, when a pronoun precedes noun, the pronoun acts as an adjective. Table 6 lists the adjectival inflections of pronouns.

The noun following the pronoun will decide (inflect) the gender. For example, तो पक्षी, हा मनुष्य, माझा सदरा, तुझा पेन, तिच्या साड्या, माझे पुस्तक, आमचा दूरध्वनी क्रमांक, कोणता गाव. This list is input as a lookup list for the tagger to use.

Table 6: Adjectival Pronouns in Marathi [23]

Pronoun (सर्वनाम)	Adjective formed from Pronouns (सर्वनामिक विशेषण)	Number (वचन)
मी	माझा, माझे	Singular
आम्ही	आमचा	Plural
तू	तुझा	Singular
तुम्ही	तुमचा	Plural
तो	त्याचा, तसा, तसला, तितका, तेवढा, तमका	Singular
ती	तिचा	Singular
हा	असा, असला, इतका, एवढा, अमका	Singular
जो	जसा, जसला, जितका, जेवढा	Singular
कोण	कोणता, केवढा	Singular
काय	कसा, कसला	Singular

Input: Marathi sentence as UNICODE characters (Si).
Algorithm:
Pass I:
For every word w_i Є S_i
Initialize the feature vector for each word
Lookup Stop words and initialize the feature vector of each stop word. POS tag for stop word = [NA] (meaning POS not applicable) and postagged[i] = [1, F] where i is the position of the stop word in the sentence. The [1, F] indicates that the tagger should not change the tag in the future passes (F indicates that the tag is final and must not be changed).
Lookup the adjectives inflected from pronouns, set feature vector.
begin
Lookup suffixlist for w_i
Remove suffix and populate the feature vector with the root word for w_i
Apply morphological rules
Use dictionary lookup for the root word to extract meaning, gender and number
If word w_i is not found in the dictionary, tag = NNP (Proper noun)
Add this to feature vector for w_i
end
for all w_i where postagged[i] != [1,F]
Initialize a postagged[i] = {0, U}
// postagged[i] = [0, U] indicates word w_i is untagged
// postagged[i] = [1, U] indicates the POS tag for word w_i may change in the succeeding passes
// postagged[i] = [1, F] indicates the POS tag for word w_i is final and will not change in the succeeding passes

For Pass II onwards (till all postagged[i] = [1, F]) do:
For every word w_i Є S_i and till there are changes to t_i after every pass postagged[i] == [0, U] or [1, U])
begin

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

Initialize w_i-1 = previous word occurring before w^i, NULL if w_i is the first word
Initialize w_i+1 = next word occurring after w_i NULL if w_i is the last word.
Let t_i be the current tag for w_i
Apply predefined grammar rules to the tagged sentence using w_i-1 and w_i+1
Recompute tinew for w_i
If tag changes to tinew (t_i != t_{inew}) then
postagged[i] =[1, U] // tag changes are happening
lookup new meaning in the dictionary, change corresponding features in the feature vector, word meaning in feature vector of w_i
else
postagged[i] = [1, F]
end
Store the feature vector of each word w_i belonging to of S_i to backend store.

Figure 2: The hybrid tagger Algorithm

The tagger works well for grammatically correct sentences with no ambiguity in the meaning of the words. Minimum three passes are required by the tagger to tag sentences. Simple sentences such are tagged by the tagger accurately; refer to the concise output of the tagger shown in table 7 for some of the Marathi sentences.

Table 7: Output of the tagger for some simple Marathi sentences.

Marathi Sentence	Concise Tagger output
जगनने कोमलला पुस्तक दिले	जगनने/NNP कोमलला/NNP पुस्तक/NNC दिले/VRB(Past Tense, पूर्ण भूतकाळ).
गीता चांगले खात जाईल	गीता/NNP चांगले/RB खात/VRB जाईल/VAUX(Future tense, रीति भविष्यकाळ)
चोराने दागिने चोरले	चोराने/NNC दागिने/NNC चोरले/VRB(Past tense, पूर्ण भूतकाळ)
नंदू गाडी चालवीत असे	नंदू /NNP गाडी/NNC चालवीत/VRB असे/VAUX (Past tense, रीति भूतकाळ)
पुण्यात सवाईचा कार्यक्रम नेहमी रंगतो.	पुण्यात/NNP सवाईचा/NNP कार्यक्रम/NNC नेहमी/RB रंगतो/VRB (Present tense, रीति वर्तमान)

V. AMBIGUITIES OBSERVED IN THE PROCESS OF TAGGING BY THE TAGGER

Consider the following Marathi sentence.

Input sentence 1: मधुला चांदीचे पदक मिळाले.

Output of the hybrid POS tagger:

Pass I: (Gender indicated as 1: Masculine, 2:Feminine, 3: Neuter, 0: Unknown)

The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector
Word : **मधुला** POS : [NNC/JJ] Person : Second Gender : 1 Vachan :Singular Stem : [**मधु**] Root : [**मधु**] Suffix : [**ला**] Meaning: [**मध /गोड**] Postion : 1 Prayog : null Prefix : null	Word : **चांदीचे** POS : [NNC] Person : Sixth Gender : 2 Vachan : Singular Stem : [**चांदी**] Root : [**चांदी**] Suffix : [**चे**] Meaning: [**शुध्द रूपें**] Postion : 2 Prayog : null Prefix : null	Word : **पदक** POS : [NNC] Person : First Gender : 3 Vachan : Singular Stem : [**पदक**] Root : [**पदक**] Suffix : [null] Meaning:[**एक कंठभूषण**] Postion : 3 Prayog : null Prefix : null	Word : **मिळाले** POS : [VRB] Tense : simple Past Aspect : null Mood : null Stem : [**मिळ**] Root : [**मिळणे**] Meaning:[**प्राप्त होणे**] Suffix : [null] Postion : 4

In Pass II; after applying grammar rule (Start of a sentence should be a noun or pronoun) to word मधुला and using the tag of the next word, the tag changes to NNP.

The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector
Word : **मधुला** POS : [NNP] Person : Second Gender : 1 Vachan :Singular Stem : [**मधु**] Root : [**मधु**] Suffix : [**ला**] Meaning: [NULL] Postion : 1 Prayog : null Prefix : null	Word : **चांदीचे** POS : [NNC] Person : Sixth Gender : 2 Vachan : Singular Stem : [**चांदी**] Root : [**चांदी**] Suffix : [**चे**] Meaning: [**शुध्द रूपें**] Postion : 2 Prayog : null Prefix : null	Word : **पदक** POS : [NNC] Person : First Gender : 3 Vachan : Singular Stem : [**पदक**] Root : [**पदक**] Suffix : [null] Meaning:[**एक कंठभूषण**] Postion : 3 Prayog : null Prefix : null	Word : **मिळाले** POS : [VRB] Tense : simple Past Aspect : null Gender : masculine Mood : null Stem : [] Root : [**मिळणे**] Meaning:[**प्राप्त होणे**] Suffix : [null] Postion : 4

Pass III does not change any of the existing tags and hence the algorithm converges.

Similarly for input sentence अटल बिहारी वाजपेयी भारताचे तेरावे पंतप्रधान होते, the verb inflection on the gender and number of the proper noun is observed in the second pass, after applying Marathi grammar rules.

The first word अटल is not found in the dictionary and the longest subsequence match method used by the tagger matches the word अटल with the word अटळ and the tagger tags the word अटल as JJ (Adjective). In the second pass, the tagger sets w_i+1= बिहारी, w_i = अटल and w_{i-1}=NULL. After applying Marathi grammar rules, the tagger sets the tag of both w_i and w_{i+1} as NNPC. A succession of 4 or more NNP words leads to the tagger to tag the succession (except the last

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

word) with the tag NNPC. Output of the hybrid POS tagger is as follows:

Input Sentence 2 : sentence अटल बिहारी वाजपेयी भारताचे तेरावे पंतप्रधान होते

Pass I:

The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector
Word : **अटल** POS : [JJ] Person : First Gender : 1 Vachan :null Stem : [**अटळ**] Root : [**अटळ**] Suffix : null Meaning: [**न टळणारा**] Postion : 1 Prayog : null Prefix :	Word : **बिहारी** POS : [NNPC] Person : Third Gender : 1 Vachan : Plural Stem : [**बिहार**] Root : [**बिहार**] Suffix : [**ई**] Meaning: null Postion : 2 Prayog : null Prefix : null	Word : **वाजपेयी** POS : [NNPC] Person : Third Gender : 1 Vachan : null Stem : [**वाजपेय**] Root : [**वाजपेय**] Suffix : [**ई**] Meaning: null Postion : 3 Prayog : null Prefix : null	Word : **भारताचे** POS : [NNPC] Person : Sixth Gender : 1 Vachan : Singular Stem : [**भारता**] Root : [**भारत**] Suffix : [**चे**] Meaning: [**हिंदुस्थान**] Postion : 4 Prayog: null Prefix : null	Word : **तेरावे** POS : [NNP] Person : First Gender : 1 Vachan : Singular Stem : [**तेरावे**] Root : [**तेरावे**] Suffix : null Meaning: [**तेरा-आकडा**] Postion : 5 Prayog : null Prefix : null	Word : **पंतप्रधान** POS : [NNC] Person : null Gender : 1 Vachan : Singular Stem : [**पंतप्रधान**] Root : [**पंतप्रधान**] Suffix : null Meaning: [**मुख्य, पेशवा**] Postion : 6 Prayog : null Prefix :	Word : **होते** POS : [VRB] Tense : simple Past Aspect : null Gender : masculine Mood : Passive Stem : [**हो**] Root : [**होणे**] Meaning:[**बनणे, असणे**] Suffix : null Postion : 7

Pass II:

The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector
Word : **अटल** POS : [NNPC] Person : null Gender : 1 Vachan : Singular Stem : [**अटल**] Root : [**अटल**] Suffix : null Meaning: null Postion : 1 Prayog : null Prefix : null	Word : **बिहारी** POS : [NNPC] Person : null Gender : 1 Vachan : Singular Stem : [**बिहार**] Root : [**बिहार**] Suffix : [**ई**] Meaning: null Postion : 2 Prayog : null Prefix : null	Word : **वाजपेयी** POS : [NNPC] Person : null Gender : 1 Vachan : Singular Stem : [**वाजपेय**] Root : [**वाजपेय**] Suffix :[**ई**] Meaning: null Postion : 3 Prayog : null Prefix : null	Word : **भारताचे** POS : [NNPC] Person : null Gender : 1 Vachan : Singular Stem : [**भारता**] Root : [**भारत**] Suffix : [**चे**] Meaning: [**हिंदुस्थान**] Postion : 4 Prayog : null Prefix : null	Word : **तेरावे** POS : [NNP] Person : null Gender : 1 Vachan : Singular Stem : [**तेरावे**] Root : [**तेरावे**] Suffix : null Meaning: [**तेरा-आकडा**] Postion : 5 Prayog : null Prefix : null	Word : **पंतप्रधान** POS : [NNC] Person : null Gender : 1 Vachan : Singular Stem : [**पंतप्रधान**] Root : [**पंतप्रधान**] Suffix : null Meaning: [**मुख्य, पेशवा**] Postion : 6 Prayog : null Prefix : null	Word : **होते** POS : [VRB] Tense : simple Past Aspect : null Gender : masculine Mood : Passive Stem : [**हो**] Root : [**होणे**] Meaning:[**बनणे, असणे**] Suffix : [null] Postion : 7

Pass III leads to no change in the POS tags of any words of the sentence and hence the tagger stops, giving the above result.

When a pronoun acts as a demonstrative word, ambiguity arises in the process of tagging. Consider the following Marathi sentence.

Input sentence 3: माझा चेंडू त्या लोखंडी खाटेखाली गेला.

In the first pass, the tagger has tagged the word 'त्या' as a pronoun instead of DEM (demonstrative). In the second and third pass the tagger does not change the tag for the word त्या. Output of the hybrid POS tagger is as follows:

The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector	The Feature Vector
Word:**माझा** POS : [PRP] Person : First Gender : 0 Vachan :Singular Stem : [**माझा**] Root : [**माझा**] Suffix : [null] Meaning: [**माझा**] Postion : 1 Prayog : null Prefix : null	Word : **चेंडू** POS : [NNC] Person : First Gender : 1 Vachan : Singular Stem : [**चेंडू**] Root : [**चेंडू**] Suffix : [null] Meaning: [**खेळण्याचा गोळा**] Postion : 2 Prayog : null Prefix : null	Word : **त्या** POS : [PRP] Person : First Gender : 3 Vachan : Plural Stem : [**त्या**] Root : [**ती**] Suffix : [null] Meaning:[**त्या**] Postion : 3 Prayog : null Prefix : null	Word: **लोखंडी** POS : [JJ] Gender : 3 Vachan: Singular Stem : [**लोखंडी**] Root : [**लोखंड**] Suffix : [**ई**] Meaning:[**एक कठिण धातू**] Postion : 4 Prayog : null Prefix : null	Word: **खाटेखाली** Gender : 3 Vachan: Singular Stem : [**लोखंडी**] Root : [**लोखंड**] Suffix : [**ई**] Meaning:[**एक कठिण धातू**] Postion : 5 Prayog : null Prefix : null	Word : **गेला** POS : [VRB] Tense : Simple present Aspect : null Gender : 1 Mood : null Stem : [**गेला**] Root : [**गेला**] Meaning: [**जाणे**] Postion : 6

VI. Observations of the Marathi Tagger

A list of 137 Marathi sentences has been used to test the tagger algorithm under study. The smallest sentence has 3 words (for sentence like चोराने दागिने चोरले) and the longest sentence has 13 words (for the sentence like गुरुवारी सकाळी दादा जरी बाहेर जाण्याच्या तयारीत होते, तरी आई बाहेर आली नाही).

Table 8: Types of sentences input to the tagger

Types of Sentences	Number of sentences
Sentences in Present, Past and future Tenses	73
Sentences with Compound Proper Nouns	7
Sentences having conjunctions	16
Sentences having exclamations	10
Sentences having nouns and Pronouns acting as adjectives	17
Sentences with prepositions	8
Sentences having verb acting as nouns	4
Grammatically incorrect sentences/Sentences with unknown words	2
Total Sentences	**137**

Each of these sentences is analysed by Marathi language expert. Every word of each sentence is tagged and its feature vector elicited by the language expert. This is used as a standard to develop Marathi grammar rules that are used by the hybrid tagger. The sentences used for testing are of various

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

types, table 8 details the types of sentences analysed. The dictionary contained 1947 words.

Following are the observations of the hybrid tagger.

1) It was observed that for simple sentences where there is no ambiguity in the meaning of any word, the tagger works correctly.

2) A succession of proper nouns (NNPC), as shown in sentence 2 in section 4.9, is tagged incorrectly. The boundary of NNPC and NNP words is not clearly delimited.

3) Words that are nouns (NNP or NNC) can appear before other nouns to qualify them, such preceding nouns are actually adjectives. For example words such as कापड दुकान (cloth store), फळ भाजी (fruit-vegetable),पुस्तक विक्रेता (book seller), सातारी पेढा (pedhas from Satara), बनारसी शालू and many such more words. The tagger under study, segments such words into two words. Pass I of the tagger tags the first word (that is, कापड, फळ, पुस्तक) as NNC (Common noun) or NNP (Proper Noun) for words not appearing in the dictionary (example: सातारी, बनारसी). The further passes of the hybrid tagger does not change the tags of such words to adjectives.

4) Demonstrative (DEM) words such as त्या are tagged incorrectly as PRP (pronouns) instead of DEM as shown in sentence 3.

5) The word आणि and व can be used as conjunctions (उभयान्वयी अव्यय) in sentences. For example, in the sentence 'विजा चमकू लागल्या आणि पावसाला सुरवात झाली'. The tagger works correctly tagging आणि as a conjunction. However, in the noun phrase 'कृषी आणि फलोत्पादन विभाग', the tagger incorrectly tags the word आणि as a conjunction in the first pass. Similar case is seen for 'व'. In the second pass however, the tagger checks the previous (-1) and next word (+1), applies the rule 'if previous word tag is VRB/VFM/AUX and next word tag is NNP/NNC/PRP then the current word (at position 0) tag is CC (conjuncts) else NNP/NNC/PRP (that is, noun or pronoun).

6) It was observed that in some sentences an adjective can appear as a noun. For example, in the sentence 'चांगल्या कामाचा परिणाम चांगलाच असतो' (which is a phrase in the native language (बोलीभाषा)), the word 'चांगल्या' is treated as a noun instead of an adjective. The tagger tags the word चांगल्या as noun. To handle idioms and phrases, the tagger needs a different approach. One of them would require a dictionary of phrases and idioms to be translated as they are. Currently, such idioms and phrases are not handled by this hybrid tagger.

7) In some Marathi sentences an adjective may appear as a verb even when inflected. For example, some root verbs (धातू) when inflected act as adjectives, For example, वाहती नदी (धातू-वाह), हसरी मुले (धातू-हस), पिकलेला आंबा (धातू-पिक) and so on. The tagger tags such words as verb in the first pass. However, in the second pass when it looks back and ahead one word, it observes that the root verb is not the last word. It also derives that no auxiliary verb is present in the look ahead. It uses the rule noun preceded by adjective and tags the words वाहती, हसरी, पिकलेला as adjective.

8) A few sentences in Marathi were given as input to the tagger in which a noun has been used as a verb (with/without inflections). Consider the following sentences:

i) माझे महाराष्ट्र बँकेत खाते होते.

The word 'खाते' in (i) appears as a VERB (root word - खा + morphology rule to add 'णे' makes the word 'खाणे' which is habitual, plural) instead of NOUN (singular, direct).

ii) १७ तासांच्या कठोर परिश्रमानंतर, मला झोप आवश्यक होती.

The word 'झोप' in (ii) is correctly tagged as VERB. However, in the following sentence (iii) the same word (झोप) appears as verbal NOUN instead of VERB. But the tagger tags the verbal noun as verb.

iii) शरीराला कमीतकमी ६ तास झोप आवश्यक आहे.

To conclude, verbs may appear as verbal nouns in their infinitival form and may function as nouns. Nevertheless, a verbal noun retains many of its verbal properties. In the sentence 'पोहण्याचे शरीराला अनेक फायदे आहेत' the word 'पोहणे', appearing as 'singular /oblique' case, is tagged as a VERB instead of NOUN.

9) Main Verb or Auxiliary Verb ambiguity.

In the sentence बघतच बसला होता (kept on seeing) बघतच is the main verb (progressive). This is because when the tagger sees a 'त' in the word 'बघत' it indicates the locative (सप्तमी) person. It expects a location/place, that is, a noun, when the 'त ' is removed from 'बघत'. The tagger gets the root 'बघ' which is not a location/place, resulting

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

in ambiguity. The POS tagger is unable to resolve this ambiguity since the contextual information is missing. In another sentence like 'हौदातच बसला होता', the root word is 'हौद', indicating a location, which is a noun, the tagger works properly.

10) Main Verb or Noun ambiguity.

The tagger differentiates a verb from a verbal noun using the rule that there can be only one main verb in a sentence. If there are more than one root verbs (धातू) in a sentence, then the tagger uses the lookback and lookahead, to find which is the proper main verb. For example, in the sentence, मुलांना खाऊ आवडतो, the tagger get two verbs with roots (धातू) as खा and आवड. In the second pass the tagger tags the word आवडतो as 'VFM' (Main Verb Finite). In the third pass, with the lookahead(1), it tags the word खाऊ as Noun, since it finds no auxiliary verb.

Another example of noun-verb ambiguity can be observed in the phrase करून टाकले होते (had done), कर can appear as a noun (कर - tax, as in जकात कर वसूल केला) or as root of a verb कर (to do as in तो नियमीत व्यायाम करीत असे). In order to resolve this POS ambiguity, the system requires information that the auxiliary verb (असे) follows the main verb. This information excludes the possible tag Noun and leaves Main Verb as the correct one. For example, in the sentence पावसामुळे क्रिकेटची मैच ३८ षटकांची करण्यात आली होती (Due to rain, the cricket match was reduced to 38 overs), the verb group is identified as (करण्यात आली होती). (कर) is marked as a verb and आली होती as auxillary verbs. Where as in the sentence जकात कर दिला गेला होता, it appears as a noun.

11) Adverb or adjective ambiguity.

An adjective can function as an adverb. For example consider the sentence "गोगलगाय फार हळू चालते". The word "फार" is an adverb since it is directly related to the word "चालणे". It appears in the dictionary as an adjective. The tagger tags the word as adjective (JJ) in the first pass. But in the second pass it sees the next word tagged as adverb (RB) and previous word as NNC. It uses the rule S→ NNC + RB+ + [VFM /VRB], and changes the tag of "फार" to RB. This is a transitive dependency since 'फार' is related to 'हळू' and 'हळू' is related to 'चालणे'. Hence, 'फार' is related to 'चालणे'.

12) In case a noun or pronoun is inflected with the षष्ठी विभक्ती प्रत्यय (चा, ची, चे), the gender and number information of the base noun and the षष्ठी विभक्ती प्रत्यय are different. For example, in the word गाडीचा (of the vehicle), गाडी has the feature singular, feminine and the suffix: चा (षष्ठी विभक्ती प्रत्यय) has the feature masculine, singular. The tagger has to retain the feature vectors for the base noun and inflection (षष्ठी विभक्ती प्रत्यय). The hybrid tagger fails in this case. It maintains only one the feature vector, that of the inflection for the noun.

13) This study currently does not handle affixes and more than one suffixes as in the word करणाऱ्याने (कर + णारा + ने). Words imported from another language or from the colloquial language (बोलीभाषा) are not handled by the tagger. Similarly, Proverbs and idioms are not handled by well the tagger.

VII. ANALYSIS OF RESULTS OF THE MARATHI TAGGER

A total of 137 Marathi sentences were tested using the hybrid tagger. Table 9 summarizes the parameters and results of the tagger. The tagger's accuracy was measured to be 84.63% for the sentences tested. The tagger worked well for inflectional morphology, however handling the complexities in handling derivational morphology remains to be developed in the form of separate suffix replacement rules. The rule base needs to be updated to handle derivational morphology. This will remain a task for the future.

The accuracy of the hybrid tagger is computed using the following measure.

$$\text{Accuracy} = \frac{\text{Number of correctly stemmed words}}{\text{Total Number of words}} \text{ X } 100$$

Table 9: Results of the hybrid tagger

Description	Number
Total Number of Sentences	137
Number of Unique words	1763
Number of words in the Dictionary	1947
Number of words tagged correctly	1492
Number of words with Multiple / incorrect tags	248
Number of words not tagged (UNK)	3

The graph shown in figure 3 shows the results of the hybrid POS tagger.

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

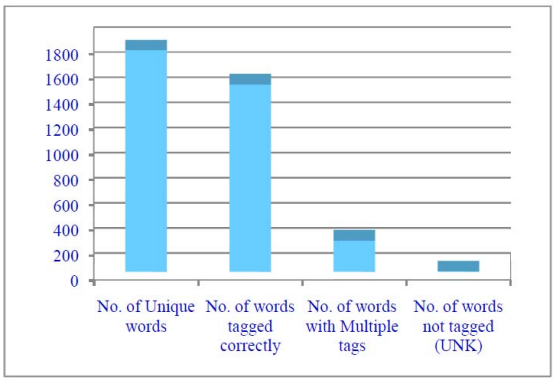

Figure 3: Output analysis of the hybrid POS tagger

VIII. CONCLUSIONS

The Marathi hybrid POS tagger uses a vast rule-base and thus relies on string manipulation. It also uses a number of lookup files and requires a comprehensive Marathi dictionary. These lookup files and dictionary needs to be updated and maintained from time-to-time.

A word will be stemmed incorrectly by the tagger, if either it is spelled incorrectly in terms of Unicode characters. The tagger will not be able to tag words on which the rule base cannot be applied. This typically happens if the tagger comes across a word from another language or from the colloquial language (बोलीभाषा) or proverbs and idioms. The UNK tag is used specifically for this purpose and there will be no features associated with the unknown word.

We have evaluated the performance of the tagger based on accuracy. The set of fully tagged 137 sentences with feature vector associated with each word of every sentence is analysed by a linguistic expert. These analysed sentences are used as a benchmark for comparing the results of the hybrid tagger. The accuracy of the hybrid tagger is around 84%. The analyses of words give a feature list including root and suffix for every word of a sentence, which is mostly sufficient for NLP applications.

The rule-base provides a comprehensive set of rules for each category of sentences. The hybrid tagger makes use of a Marathi dictionary and a lookup list of suffixes. This tagging method is economical in terms of space and time complexity for the Marathi data set that is constructed. Once the suffixes are identified, removing the suffixes and applying proper morpheme sequencing rules can obtain the stem. In order to identify the correct stem and POS tags, the Marathi dictionary needs to be robust and exhaustive as possible.

IX. FUTURE WORK

We plan to extend this study to POS tag derived words in Marathi language that can be used in Marathi sentences. Prefix or affix removal can also be considered as future work with respect to POS tagging.

REFERENCES

[1] Allen James, Natural Language Understanding, Pearson Education, ISBN 978-81-317-0895-8, Second Ed., 2008.

[2] Apte V. G, *मराठी शब्दरत्नाकर (Marathi Shabdaratnakar)*, Aadarsha Vidhyarthi Prakashana, ISBN-13: 978-9382259893 Reprinted 2012.

[3] A. Ramnathan and D. Rao, "A Lightweight Stemmer for Hindi", *Proc. Workshop on Computational Linguistics for South Asian Languages, 10th Conference of the European Chapter of Association of Computational Linguistics*, pp. 42-48, 2003.

[4] Bharti A., V. Chaitanya, and R. Sangal, *Natural Language Processing: A Paninian Perspective*, Prentice Hall, New Delhi, 1995.

[5] Bellaris H. S. K., Askhedkar L. Y, A Grammar of the Marathi Language, Obtained in digital format digitized by the Internet Archive in 2007 with funding from Microsoft Corporation, 1868.

[6] Christopher D. Manning, Hinrich Schütze, *Foundations of Statistical Natural Language Processing,* MIT Press, ISBN 0262133601, 2000.

[7] Dandapat, S. , "Part-of-Speech Tagging and Chunking with Maximum Entropy Model", *Workshop on Shallow Parsing for South Asian Languages*, 2007.

[8] Dandapat, S. and Sarkar, S., "Part-of-Speech Tagging for Bengali with Hidden Markov Model", *NLPAI ML workshop on Part of speech tagging and Chunking for Indian language,* 2006.

[9] Handbook of Natural Language Processing, Second edition (2010) Edited by Nitin Indurkhya, Fred J. Damerau, Chapman & Hall/CRC Press, ISBN-13: 978-1-4200-8593-8 (Ebook-PDF)

[10] Harold Somers, *Computers and translation: a translator's guide,* John Benjamins Publishing Company, Amsterdam, ISBN: 90-272-1640-1, 2003

[11] H.B. Patil, A.S. Patil and 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,* pp. 33 – 37, 2014.

[12] Hooper, R. and Paice, C., "The Lancaster Stemming Algorithm", Available at: http://www.comp.lancs.ac.uk/computing/research/stemming/ (Accessed: 11/7/2012)

[13] J. B. Lovins, "Development of Stemming Algorithm", *Mechanical Translation and Computational Linguistics*, Vol. 11, No. 1, pp. 22-31, 1968.

[14] Mehar Vijay and Phani Gedde, "Improving statistical POS tagging using linguistic features for Hindi and Telugu", Proc. *The International Conference on Natural Language Processing (ICON)*, 2008.

[15] Navalkar Ganpatrao, The Student's Marathi Grammar, Asian Educational Services,2001.

[16] Phadke Arun, *मराठी लेखन - कोश (Marathi Lekhana - Kosha)*, Ankur Publication, Thane, 2008.

[17] P. Majumder, M. Mitra, S. K. Parui, G. Kole, P. Mitra and K. Datta, "YASS: Yet another suffix stripper", *ACM Transactions on Information Systems*, Vol. 25, No. 4, pp. 18-38, 2007.

[18] P. Arulmozhi, L. Sobha and K. Shanmugam, "Parts of Speech Tagger for Tamil", *Proc. The Symposium on Indian Morphology, Phonology & Language Engineering,* Indian Institute of Technology, Kharagpur, pp. 55-57, 2004.

[19] P. J. Antony and K.P. Soman, "Kernel based part of speech tagger for Kannada", *Proc. Machine Learning and Cybernetics (ICMLC)*, Vol. 4, pp. 2139 –2144, 2010.

[20] Paice, "Another Stemmer", *ACM SIGIR Forum*, Vol. 24, No. 3, pp 56-61, 1990.

[21] Phadke Arun, *मराठी लेखन - कोश (Marathi Lekhana - Kosha)*, Ankur Publication, Thane, 2008.

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

[22] Ratnaparakhi. Adwait, “A maximum entropy model for part-of-speech tagging”, *Proc. The Conference on Empirical Methods in Natural Language Processing*, 1996.

[23] Om Vikas, "Language Technology Development in India", *Ministry of Information Technology*, New Delhi, India.

[24] Walimbe M. R., *सुगम मराठी व्याकरण लेखन (Sugam Marathi Vykarana Lekhan)*, Nitin Prakashana, Pune, ISBN 81-86169-80-6, 2012.

[25] https://en.wikipedia.org/wiki/Part-of-speech_tagging

978-1-5386-2051-9/18/$31.00 ©2018 IEEE