Part-of-Speech Tagger for Marathi Language using Limited Training Corpora

H.B. Patil
School of Computer Sciences
North Maharashtra University,
Jalgaon.

A.S. Patil
School of Computer Sciences
North Maharashtra University
Jalgaon.

B.V. Pawar
School of Computer Sciences
North Maharashtra University
Jalgaon.

ABSTRACT
Part-of-speech tagging in Marathi language is a very complex task as Marathi is highly inflectional in nature & free word order language. In this paper we have demonstrated a rule-based Part-of-Speech tagger for Marathi Language. The hand–constructed rules that are learned from corpus and some manual addition after studying the grammar of Marathi language are added and that are used for developing the tagger. Disambiguation is done by analyzing the linguistic feature of the word, its preceding word, its following word, etc. After testing the system with three data sets we got encouraging results. The accuracy of our system is of an average 78.82% after testing it on three different data sets.

General Terms
Natural Language Processing.

Keywords
POS tagger, Morphological analysis, rule-based.

1. INTRODUCTION
Part-Of-Speech (POS) tagging is an important process used as a building block for various NLP tasks like Machine translation, Natural language text processing and summarization, User interfaces, Multilingual and cross language information retrieval, Speech recognition, Artificial intelligence, Parsing , Expert system and so on. POS tagging is the process of choosing the correct grammatical tag for a word based on the context or morphological properties. Automated POS has been extensively used for more than one decade. POS taggers are designed with the aim of analyzing text of sample language using corpora to determine the syntactic categories of the words or phrases used in the text. POS tagger is a program that accepts an unprepared raw text as input and to each word adds a tag specifying its grammatical properties. It performs a mapping from sequence of words to a sequence of lexical categories. POS tagging consists of 3 stages: Tokenization, Morphological Analysis and Disambiguation.

2. LITERATURE SURVEY
The work on Part-of-Speech (POS) tagging has begun in the early 1960s [30]. The POS tagger can be implemented by using either a supervised technique or an unsupervised technique [12, 27]. Under these two categories different approaches have been used for the implementation of POS taggers such as : Rule-based [8,13,19,30], Stochastic or probabilistic [4,15, 10], Neural networks [17] and Hybrid [33].The earliest taggers S. Kelin & R. Simmons (1963) and Barba B. Greene & Gerald M . Rubin (1971) has large sets of hand constructed rules for assigning tags on the basis of word character patterns and on the basis of tags assigned to preceding or following word, but they had only small lexical, primarily for exceptions to the rules [30]. The rule-based approach for developing POS tagging was continued, most notably by Karlsson (1990) [16], Voltilainen and colleagues (1995) [8, 10], Tapanainen and Chanod (1994) [19], and Brill (1992) [13] Kh Raju Singha et.al (2012)[23].The statistical approach is also used for POS tagging. Various approaches in stochastic tagging are –Hidden Markov Model (HMM) taggers [4], Transformation –based Taggers. e.g. -Brill's Tagger (1995) , Decision Tree learning for Taggers. e.g. -Helmut Schmid Tree Tagger (1994) , Maximum Entropy Taggers [2], Neural Networks [17], Memory Based Learning [33]. The alternative approaches for development of POS tagging systems includes Neural Networks and Hybrid taggers. Nakamura (1990) trained a 4-layer feed-forward network with upto three preceding Part-of-Speech tag as input to predict the word category of next word [17]. Federici and Pirrelli (1993) & Helmut Schmid (1994) developed a POS tagger which is based on a Multilayer Perceptron network. Adwait Ratnaparkhi (1996) came up with a POS tagger based on Maximum Entropy model [2]. Weischeded (1993), Merialdo (1994) are based on hidden markov model. K.T. Lua (1996) used genetic algorithms for POS tagging of Chinese sentence [26]. Jelink (1994) Magerman (1995) uses Statistical Decision Tree [2]. Hybrid taggers are also developed such as CLAWS. Graside and Smith (1997) used both statistical and rule-based approaches [31]. In this scenario, POS tagging for highly inflectional languages presents an interesting study. Morphologically rich languages are typically free-word ordered, which causes fixed-context systems to be hardly adequate for statistical approaches (Samuelsson and Voutilainen 1997) [31]. Morphology based POS tagging of some languages like Turkish (Oflazer and Kuruoz 1994), Arabic (Guiassa 2006), Czech (Hajic 2001), Modern Greek (Orphanos 1999), Hungarian (Megyesi 1999) and Hindi (Singh, Gupta, Shrivastava and Bhattacharyya 2006) has been tried out using hand- crafted rules and statistical learning [31]. A lot of work for a language like English, related to POS tagging has been carried out. Brill (1992) developed a POS tagger for English using rule-based approach, which is one of the most successful tagger [13].

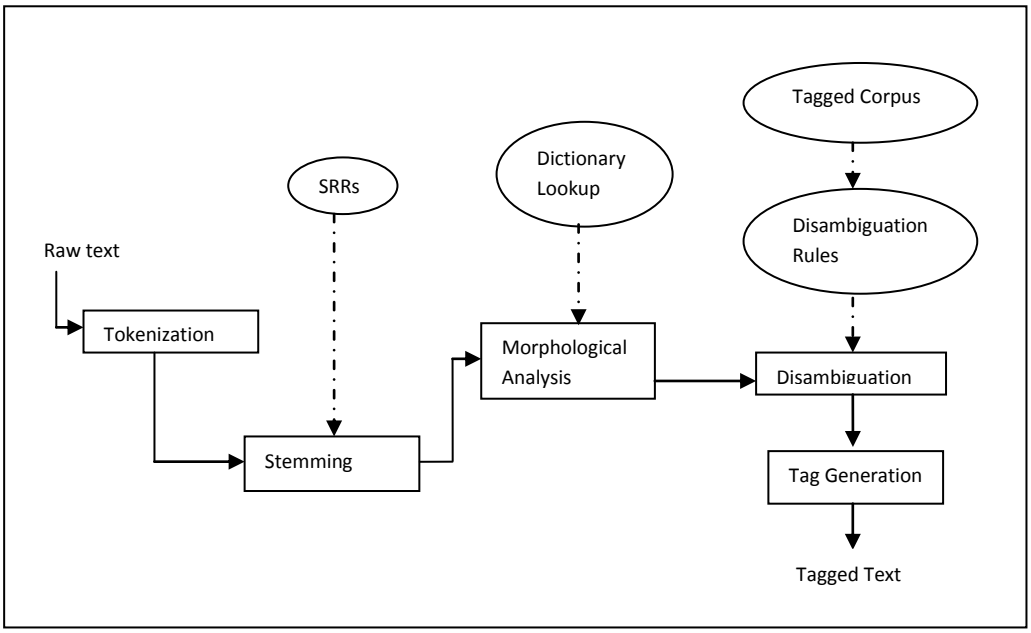

Fig 1 : Architecture of our POS tagging system.

Dinesh kumar et. al. in 2010 did a survey for part-of-speech taggers for Morphologically rich Indian languages such as (Hindi, Punjabi, Malayalam, Bengali and Telgu)[11]. Jyoti Singh et. al. in 2013 used a trigram method for part of speech tagging of Marathi text. So by taking this literature as a basis we have decided to develop a rule-based POS tagger for Marathi language.

3. EXPERIMENTAL SETUP

The process of POS tagging consists of three stages: Tokenization, Morphological analysis and Disambiguation By considering these three stages of POS tagging we have developed our own architecture for Marathi POS tagger as given in fig 1.

3.1 Tokenization

Tokenization is the process of separating tokens from raw text. Marathi is a segmented language where word boundaries are fixed. Words are separated by white spaces or punctuation marks. In segmented languages like Marathi since word boundaries are clear tokenization becomes easy. So by using this we can easily find out the tokens from the sentence.

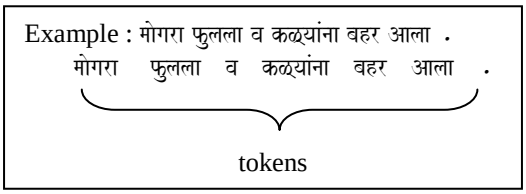

Fig 2: Tokens

3.1.1 Problems of Tokenization for Marathi language

The word in Marathi may include hyphen (-) and colon (:) like (Ex- c*halata – chalata* , *swatacha*) but no other punctuation mark. The major problems in tokenization for Marathi language are listed below:

Segmentation

Segments with the period at the end (Ex- *va. pu. kale*.) suffer from segmentation ambiguity. The period can denote an abbreviation or the end of the sentence, or both as like English. If a hyphenated segment such as (Ex- *don – tin*) is encountered then hyphens should be treated as independent tokens and the words *don* and *tin* are also considered as two different tokens.

Round up

If a word consisting of a sequence of segments such as a proper noun (Ex- *navi delhi*) then this proper noun is considered as two different tokens like *navi* and *delhi*.

3.2 Stemming

Stemming process removes all possible affixes and thus reduces the word to its stem. Stemmer stems the term by pattern matching. The stemming process is carried out as given in fig.3

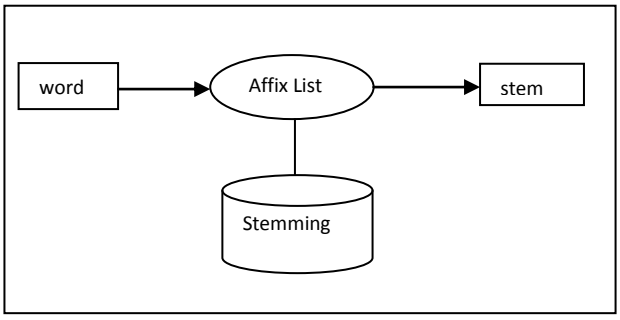

Fig 3 : Stemming process

Example: By considering our above example after stemming the stems of the tokens as given in table 1.

Table 1. Stemming example

Word	मोगरा	फुलला	व	कळ्यांना	वहर	आला	.
Stem	मोगरा	फुल	व	कळ्या	वहर	आ	.

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. The SRRs are used to convert the stem word into the root-word. We have developed 25 SRRs. The sample rule for masculine nouns is as below:

Table 2. SRR example

Rule	If the stem word ends with आ replace आ by अ
Example	कागदा becomes कागद
Exception	मामा , काका

Example: For the word *Kalya* the SRR applies is If the stem word ends with '*ya*' replace '*ya*' by *e / i* is used and then the word *kali* is identified as a root-word. The root-words that are identified are then given to morphological analyzer.

3.3 Morphological Analysis

Morphological analysis is the process of formation and alteration of words. Morphological analysis gives the information about the words like possible POS tags, gender, etc. The morphological analysis is carried out by dictionary lookup and morpheme analysis rules.

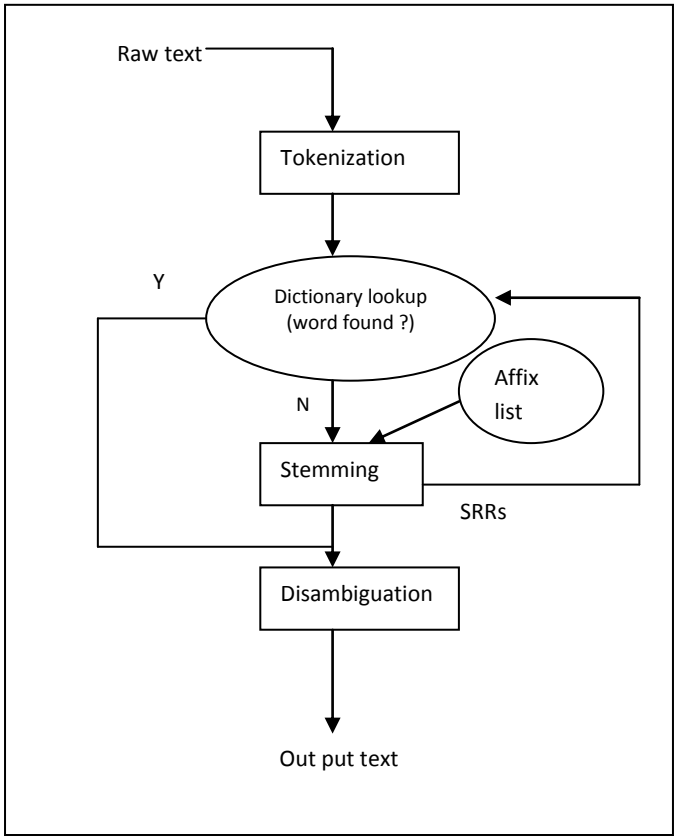

Fig 4: Process of morphological analysis

Table 3. Morphological analysis example

Root Word	मोगरा	फुल	व	कळी	वहर	आला	.
Possible POS Tags	N	N/V	C	N/V	N/V	V	P

3.4 Disambiguation

A word can have more than one grammatical categories based on the context where it is used. So disambiguation is necessary to resolve the ambiguity. Disambiguation selects the most possible sequence of lexemes by the use of rule-base model or Hidden Markov Model. Based on the corpus we have identified 11 disambiguation rules that are used to remove the ambiguity. The sample disambiguation rules that are developed are as follows:

Table 4. Disambiguation rule

Rule	If suffix is (पण / पणा / त्व / ता / य / ई / आई / वा / की / गिरी) If root-word's tag = = A Then tag = N
Example	लहानपण , जडत्व , नवलाई

According to the disambiguation rules the tags that are assigned as follows:

Table 5. Example of disambiguation

Root word	मोगरा	फुल	व	कळी	वहर	आला	.
Possible Pos tag	N	V	C	N	N	V	P

3.5 Tag Generation

This is final phase of the POS tagger. Tag generator generates the appropriate tag based on tokenization, morphological analysis and disambiguation process. The example are as given below:

Table 6. Tag generation example

Root word	मोगरा	फुल	व	कळी	वहर	आला	.
Possible Pos tag	N	V	C	N	N	V	P

4. RESULTS & DISCUSSION

We have developed our own corpus consisting of 576 unique words. Our tag set consists of 9 tags for main POS categories of Marathi language only. In order to evaluate the POS tagger for Marathi language we have used three different Test Data Sets (TDS). The first one (TDS1) and third (TDS3) are a sets constructed from the corpus sentences and the second one (TDS2) is constructed from the random sequence of the words from the corpus. The accuracy in percentage of the tagger is calculated using the formula given below :

$$Accuracy = \frac{Correctly\ tagged\ words}{Number\ of\ words\ in\ evaluation\ set} * 100$$

The summarized results of the evaluation are as given in following tables:

Table 7. Tagging accuracy

Test Data Sets	# Sentences	# Words	correct tagged words	incorrect tagged words	% age Accuracy
TDS1	30	109	81	28	74.31
TDS2	10	63	51	12	80.95
TDS3	23	121	101	20	83.34

The accuracy of the system in the form of recall, precision & f-measure is as given in the following table:

Table 8. Tagging results

Test Data Set	Precision	Recall	F-Measure
TDS1	0.7788	0.9412	0.8225
TDS2	0.8644	0.9272	0.8946
TDS3	0.8416	0.9901	0.9098

We can say that the system is working with a quite good accuracy. The errors occur because Marathi is very ambiguous language. The errors are also due to the small size of corpus. If the size of the corpus is increased then more rules can be discovered which will help to reduce the error rate. Most of the errors occur during the disambiguation module. The ambiguity and error rate can be reduced by studying Marathi Grammar structure and Marathi linguists in more details. Based on table 8 we can conclude that the rule based technique is well suitable for morphologically rich language like Marathi.

5. CONCLUSION

In this work we have reported the POS tagger for Marathi language using the rule-based technique. After developing the system and testing it with three data sets we came to the conclusion that our system is working with a quite good accuracy at an average of 78.82% which is acceptable. Although the corpus size is relatively small but our tagger still cope up with other taggers, and if the size of corpus is increased then more rules can be discovered and thus the error rate can be reduced which will ultimately increase the accuracy of the tagger.

6. FUTIRE WORK

In stemming procedure the suffixes are removed and then the word is searched in dictionary. These suffixes belong to cases and preposition category. So for the word like ‘ *vikas* ’ where the word itself consist of cases at their end, the problem arises due to the rule of stemming. From the word ‘ *vikas* ’ the last ‘*s*’ is removed first and then the characteristics mark ‘ a ’ thus we get the word ‘ *vik* ’ as stem. Similarly the same type of wrong result will be generated for the words like ‘*kunchala*’, ‘*kes*’, ‘*darshana*’, ‘*nate*’ that ends with cases. In Marathi language almost all verb in present tense ends with ‘*Ne*’ like ‘*KhaNe*’, ‘*GaNe*’, etc., but some noun like *vataNe* , *futaNe*, etc. also ends with ‘*Ne*’.The statements like ‘ *tu Jhaad laav*.’ , and ‘*tu ghar Jhaad*.’ , ‘ *Jhaad* ’ is appearing in both the statements but with different tag at former statement the correct tag is assigned but in later statement the wrong tag is given to the word. Handling all the issues is an interesting task.

7. ACKNOWLEDGMENTS

The authors are thankful to the University Grants Commission, New Delhi for supporting this research under the Special Assistance Programme (SAP) at the level of DRS-I (No: F.3-52/2011(SAP-II).

REFERENCES

[1] “A Part of Speech Tagger for Indian Languages”. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

[2] A. Ratnaparkhi “A maximum entropy model for Part-of-Speech tagging”, 1st Conference on Empirical Methods in Natural Language Processing (*EMNLP-1996*). PP133-142

[3] A. Bharati, V. Chaitanya and R. Sangal, “Computational Linguistics in India: An Overview”, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics 2000, VOL 38; PART 1, PP 595-596

[4] A. Azimizadeh, M. M. Arab, S. R. Quchani “ Parsian part of speech tagger based on Hidden Markov Model”, JADT 2008 : 9es Journées internationales d’Analyse statistique des Données Textuelles , 2008, PP 121-128.

[5] A. Ramanathan, D. D. Rao, “A Lightweight Stemmer for Hindi”, In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2003.,Workshop on Computational Linguistics for South Asian Languages (Budapest, April 2003).

[6] A. Dalal, K. Nagaraj, U. Sawant, S.Shelke and P. Bhattacharyya [2007]: “Building a Future Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi”, ICON 2007, Hyderabad, India.

[7] Arulmozhi. P, Sobha. L. Kumara Shanmugam. B, [2004]: “Parts of Speech Tagger for Tamil”, Symposium on Indian Morphology, Phonology & language Engineering IIT Kadagpur India March 19-21 2004 PP 55-57.

[8] A. Voutilainen, “A Syntax-based part-of-speech analyser”, Conference of the European Chapter of the Association for Computational Linguistics, 1995, EACL – 95 PP 157-164.

[9] B.N. Patnaik ,“Computational linguistics for Indian Languages”, Symposium on Indian Morphology Phonology and Language Engineering 2004, PP 3-4.

[10] C. Samuelsson and A. Voutilainen “Comparing a Linguistic and a Stochastic Tagger”, Proceedings of the 35th Annual meeting of the ACL and 8th Conference of the European chapter of the ACL 1997 PP 246-253.

[11] Dinesh Kumar & Gurpreet Singh Josan “Part of Speech Tagger for Morphologically rich Indian languages : A Survey”, International Journal of Computer Applications Vol. 6, No. 5, September 2010.

[12] E. Brill, “Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging”, Proceedings of the Third Workshop on Very Large Corpora, 1995.

[13] E. Brill, “A Simple Rule Based Part of Speech Tagger”, In Proceeding of the Third Conference on Applied

Natural Language Processing 1992 Toronto, Italy, PP 152-155.

[14] F. M. Hasan, N. UzZaman and M. Khan "Comparison of Different POS Tagging Techniques (n-gram, HMM and Brill's Tagger) for Bangla", Proceedings of the International Conference on Systems, Computing Sciences and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 06), December 4 - 14, 2006.

[15] F. M. Hasan, N. UzZaman, and M. Khan, " Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages.", Proceedings of the Conference on Language and Technology (CLT07), Pakistan, August 7 - 11, 2007

[16] F. Karlsson "Constraint grammar as a framework for parsing running text ", In COLING-1990, PP 163-173.

[17] H. Schmid, "Part-of-Speech Tagging with Neural Networks", In Proceeding of the International Conference on Computational Linguistics 1994, Kyoto, Japan, PP 172-176.

[18] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", In International Conference on New Methods in Language Processing 1994.

[19] J. Chanod and P. Tapanainen, "Tagging French-comparing a statistical and a constraint-based method", In EACL- 1995 PP 149-157.

[20] Jyoti Singh, Nisheeth Joshi Iti Mathur, "Part of Speech Tagging of Marathi text using trigram method", International Journal of Advanced Information Technology, Vol. 3, No. 2, April 2013.

[21] K. Bali, S. Baskaran, T. Bhattacharya, P. Bhattacharyya, M. Choudhury, G. Nath Jha, and et. al. , "A Common Part-of-Speech Tagset Framework for Indian Languages", Lexical Resources Engineering Conference (LREC08), Marrakech, Morocco, May 26-June 1, 2008.

[22] Kh. Raju Singha, Bipul Syam Purkayastha & kh. Dhiren Singha "Part of Speech Tagging in Manipuri with Hidden Markov Model", International Journal of Computer Science Issues, Vol. 9, No. 2, November 2012 PP: 146-149.

[23] Kh. Raju Singha, Bipul Syam Purkayastha & kh. Dhiren Singha "Part of Speech Tagging in Manipuri : A rule-based approach", International Journal of Computer Applications, Vol. 15, No. 14, August 2012.

[24] K. W. Church, "Current practice in Part of Speech Tagging and Suggestion for the Future", In Simmons (ed.) 1992 Sbornik Praci : In honor of Henry Kucera Michigan Slavic studies.

[25] K. Gupta, M. Shrivastava, S. Singh and P. Bhattacharyya, " Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi", In Proceedings of the COLING/ACL on Main conference poster sessions , Sydney, Australia 2006,. PP: 779 – 786.

[26] K. T. Lua, "Part of Speech Tagging of Chinese Sentences Using Genetic Algorithm", Proceedings of ICCC96, National University of Singapore, 1996.

[27] L. V. Guilder, "Automated Part Of Speech Tagging A Brief Overview", Handout for LING 361 Georgetown University Fall 1995.

[28] N. Agrawal, M. Shrivastava, S.Singh, B. Mohapatra, P. Bhattacharya, "Morphology Based Natural Language Processing tools for Indian Languages.".Workshop on Morphology 2005 .PP 71-75.Online link: http://www.cse.iitk.ac.in/users/iriss05/m_shrivastava.pdf

[29] R.M. Carrasco and A. Gelbukh, "Evaluation of TnT Tagger for Spanish.", Computer Science, 2003. ENC 2003. Proceedings of the Fourth Mexican International Conference on, ISBN:0-7695-1915-6 . PP: 18- 25 .

[30] S. Abney " Part-of-Speech Tagging and Partial Parsing", Corpus-Based Methods in Language and Speech Processing 1996,.

[31] S. Singh, K. Gupta, M. Shrivastava, P. Bhattacharyya "Morphological Richness Offsets Resource Demand-Experiences in Construction a Pos Tagger for Hindi.", Proceedings of the COLING/ACL- 2006, on Main conference poster sessions. PP: 779 – 786. Sydney, Australia

[32] U. Sawant, S.Shelke, K. Nagaraj, and A. Dalal, "Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach.", Proceeding of the NLPAI Machine Learning, 2006.

[33] Y. Tlili-Guiassa, L. M. Tayeb "Tagging by Combining Rules-Based and Memory-based Learning", Information Technology Journal 5 (4), PP 679-684. 2006. ISSN:1812-5638.