Rule Based Combined Tagger for Marathi Text

Kalpana B. Khandale, C. Namrata Mahender

Rule Based Combined Tagger for Marathi text

Kalpana B. Khandale[1] and C. Namrata Mahender[2]

[1,2] Department of Computer Science and I.T, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India
kalpanakhandale1788@gmail.com
cnamrata.csit@bamu.ac.in

Abstract. Part of speech (POS) tagging is most necessary concept in the natural language processing categorized each word in the corpus. This paper focus on the development of Marathi part of speech tagger using the N-gram models. For this we have designed rules for the development of the POS tagger. These rules are framed on the basis of the Marathi grammar. The corpus is of 635 Marathi sentences written in considering variations, for better evaluation of the POS tagger. Total 5715 words has been tagged from in which 1918 words unigram tagger and 106 words are tagged by the bigram tagger. The overall accuracy of POS tagger is 79.34%.

Keywords: Annotated corpus, Unigram, Bigram, POS tagging.

1 Introduction

Around 90 million people speak the Marathi language. Large amount of information available in Marathi in electronic form and this information is not in standard format and the uniformity of the text is the main problem. Many tools available for the normalization of the text, but those tools are not supported to the Marathi up to the mark. Therefore, neither the corpus nor the processing tools available for the Marathi language. The tools available in the other language are not adequate or they are limited to some domain specific. Almost all the application of natural language processing there is need the part of speech tagger for identifying the grammatical tag of each word. Marathi is the verb final language relatively free order language. Word order is a topological property of any language, hence the study of syntax word order is not complete when there is no deliberation or discussion on the word order of any language. The order of Marathi language is SOV (Subject Object Verb). In Marathi many words do not follow the strict order in the sentences. For example,

Table 1. Interchangeable words in sentence.

Sr. No.	Marathi Sentence	English Sentence
1.	मनोज रविला उद्या भेटवस्तू देईल.	Tomorrow Manoj will give a gift to Ravi.
2.	मनोज उद्या भेटवस्तू रविला देईल.	
3.	मनोज उद्या रविला भेटवस्तू देईल.	
4.	रविला उद्या मनोज भेटवस्तू देईल.	
5.	उद्या रविला मनोज भेटवस्तू देईल.	
6.	उद्या रविला मनोज भेटवस्तू देईल.	

The above table 1 depicts the same meaning, but the order of the words is inter-changeable. The meaning of the entire sentence is "Tomorrow Manoj will give a gift to Ravi." It is so simple in English, but in Marathi sometimes when word order has been changed but the meaning of the sentence has not been changed. But on the other hand Marathi has such unit which does not change that order like verb followed by the auxiliary verb, postpositions always follow the noun, adjectives come before the noun, etc.. On the basis of the transformational grammar of the Marathi language, noun phrase looks like the different meanings as described in the following example,

Table 2. Transformational grammar based noun phrase.

Sr. No.	Marathi Words	English Words
1.	मंदिर	A temple
2.	मंदीरासाठी	For temple
3.	मंदीरावर	On the temple
4.	मंदीराजवळ	Near the temple
5.	मंदिरात	In the temple
6.	मंदीरामध्ये	In the temple
7.	मंदीराचे	Of the temple
8.	मंदीरामधील	In the temple
9.	मंदीरासमोरील	In front of the temple
10.	मंदीरामागे	Behind the temple

From the above table the context of the 'मंदीर', if different types of suffixes and post-positions add into the word then the meaning has been change.

From the above example, we can understand the importance of the Marathi POS tag-ger. Because when the suffixes and the postpositions are added into the word the tag-ging of the word will be changed.

1.1 Part of speech tagger

In language, words have a number of different meanings. Contemplate the formal and functional distinction of words and they are grouped into the different word classes called as the part of speech tagger. There are different techniques are available for finding out the word level ambiguity, and it is based on the context of the words in the text under consideration with the respective words with POS. The work of the POS tagging in Marathi is based on the different types of rules. The rules of the POS tagging for Marathi are based on the categories of words belong to. Marathi is one of the Prakrit languages which are developed from Sanskrit. Marathi is the free ordered language and also a verb final language. There are some types of part of speech tagging described as below.

Rule based part of speech tagger
One of the most useful and oldest techniques of word tagging is the rule based POS tagging, in which we can use the handwritten rules for tagging the word. If a word has more than one feasible tag, then this rule is useful for identifying the correct tag of the word. For Marathi is it necessary because the sentences below have the more complexity of identifying the correct tag for the word. For example,

1) तू झाड लाव.
2) तू घर झाड.

In these examples the 'झाड' has two meanings into two different sentences. In the first sentence the word 'झाड' is tagged as noun and in the second sentence, is it tagged as va verb. The advantage of a rule based tagger is that knowledge driven tagger and its built manually.

Stochastic tagger
This tagger is also called as the statistic POS tagger. This is included the frequency and probabilities of words. Here we can use the word frequency and encountered the most frequent words which are present in the training set. On the other hand, the probability of tag also calculated. It requires that training corpus of tagged sentences. For this tagger if the word is not present in the corpus it is not calculating the probability of that word.

Transformation based tagging
It is also called as Brill tagging. This allows linguistic knowledge in the readable form and transforms it into one state to another state by using this transformation based rules. The advantage of this tagging it reduced the complexities of tagging because in this tagger there is entwine of the machine learning and hand-written or human generated rules are there.

2. Literature Survey

Table 3. Previous study on POS tagger.

Author, Year and Language	Methods/ Technique	Dataset	Issues/limitation	Result
AlKhwiter, W., & Al-Twairesh, N. 2021 (Arabic) [1]	-Conditional Random Fields and Bidirectional Long Short-Term Memory (Bi-LSTM) models	-Total 7750 tweets from three different types of tweets -Tweets with a length of less than seven words -Spam tweets -Tweets that were written in dialectal Arabic	-The tagger was not able to tag twitter-specific items correctly. -There was need a good tokenized as tool in NLP for Arabic	-The overall accuracy of 'Mixed' dataset was96.5%. -'MSA' and 'GLF' datasets achieve an accuracy of 95.6% and 95% respectively
Bacon, G. 2020 (Latin) [2]	-unidirectional LSTM -hyper parameter optimization	-the evolution 2020 shared task dataset -containing 14,399 sentences with of 259,645 words	-Not specified	-The overall performance of the system was above 95.3 %.
Yousif, J., & Al-Risi, M. 2019 (Arabic) [3]	-SVM -Evaluation Measures	-Which contains 131 or 77 tags. -Training and testing the SVMmodels starting with 1K, and the maximum size is 156K	-More research in the direction of analyzing and categorizing the Arabic tags and tag sets. -Needed to build a standard Arabic corpus for NLP applications	-Not mention specifically.
Kurniawan, K., &Aji, A. F. 2018(Indonesian) [4]	-bidirectional LSTM (biLSTM). -Scikit-learn library -Majority tag (MAJOR) -Memorization (MEMO) -CRF	-The corpus contains 10K sentences and 250K tokens that are tagged manually.	-Not specified	-The POS tagger demonstrated an accuracy of 69%.
Nita.V. Patil 2018 (Marathi) [5]	-N-gram -HMM used the Viterbi algorithm	-From the domain of news stories they have used 15,000 sentences.	-POS tagging is challenging task for Marathi.	-The overall accuracy of the system is 86.61%.
Ajees A P, Sumam Mary Idicula	-Statical approach -CRF Model	-They have used to 23K words for training	-Not specified	-The overall accuracy of the system is 91.2%.

2018 (Malayalam) [6]	-Maximum entropy Markov models	and 5.7K words for testing.		
-Mittal, S., Sethi, N. S., & Sharma, S. K.2014 (Punjabi) [7]	-Bigram model	-They have collected 2400 sentences from 10,000 words randomly.	-The performance of the n-gram based method was not so accurate because when encountered the unknown words like foreign language words.	-The accuracy was 92.12%.
Joshi, N., Darbari, H., Mathur, I. 2013 (Hindi)[8]	-HMM -Evaluation matrix	-From the domain of the tourism 15,200 sentences with 3, 58,288 words for trained the system.	-They have not concentrate on adding more tags the classification of the text.	-The accuracy of the system is 92.13%.
Singh, J., Joshi, N., &Mathur, I2013 (Marathi) [9]	-Statistical approach -Trigram model -Accuracy	-2000 sentences with 48,635 words have been used for the trained the system.	- The morphological complexity of the Marathi is hard to identify the morphemes.	-The accuracy was 91.63%.
Jyoti Singh ,ItiMathur&Nisheeth Joshi 2013 (Marathi) [10]	-Statistical Approach -Unigram -Bigram -Trigram -HMM methods	-They have used 1000 sentences with 25744 -Words for trained the system.	-The data they have been used was less.	-The overall system result was 93.82%.
Dhanalakshmi, V., Shivapratap, G., &SomanKp, R. S. 2009 (Tamil)[11]	-SVM tools -Non-linear SVM using Linear programming	-25,000 Sentences -10,000 sentences have used for the testing	-Corpus is not available in Tamil - Nouns get inflected for number and cases. Verbs get inflected for tense, person, number, gender suffixes.	-The overall accuracy 95.63%.
Patel, C., &Gali, K.2008 (Gujrati) [12]	-Machine learning approach -Conditional Random Fields (CRF) model	-Out of 600 sentences 10,000 words they have used for training and 5,000 words used for the testing.	-The CRF used for the features and the probabilities of the tag because of the flexible nature of the language.	-The accuracy of the system was 92%.
Asif Ekbal , Sivaji Bandyopadhyay2008 (Bengali) [13]	-Hidden Markov Model (HMM) -Support Vector Machine (SVM)	-They have used the 72,341 words for trained the system	-The data they have used were very less.	-The overall result was 91.23%.
Singh, T. D., &Bandyopadhyay, S. 2008 (Manipuri) [14]	-Machine learning approach -Accuracy	-Out of 3784 sentences10917 unique words have been used for the training purpose	-Disambiguation scheme is necessary for the Noun-Adjective ambiguity.	-The overall accuracy was 69%.

Dalal, A., Nagaraj, K.,Sawant, U., &Shelke, S. 2006 (Hindi) [15]	- Maximum Entropy Markov Model(MEMM)	- They have used NLPAI-ML 2006 contest which have consisting of around 35000 words.	-The features which have language specific that would improve the performance system, particularly in case of chunking	- The total accuracy of 88.4% for POS tagging and 86.45% for chunking
Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., & Chute, C. G. 2005 (English) [16]	- Trigram models of MED data.	The PennTreebank-2 corpus -the GENIA corpus -MED corpus of clinical notes.	-Not specified	-The accuracy of the system was 92% accuracy from 87% in our studies.

3. Proposed System

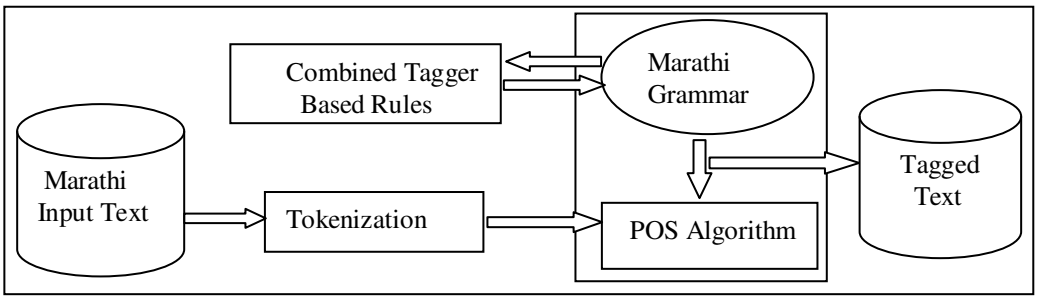

Fig1 Architecture of POS Tagger

3.1 Input Text

As specific corpus of Marathi text or dataset is not available, 907 sentences were manually created, with many variations to sense the difficulties while generating tags using the proposed POS tagger.

3.2 Tokenization

Tokenization is the process of splitting each word from the input text. In Marathi the words have separated by the white space and the punctuation marks. With the help of this we can easily tokenize the sentence into separate tokens.

3.3 Marathi Grammar

The tag to the word are assigned on the basis of the Marathi grammar rules. Marathi is the augmented language. It has eight types of part of speech (Naam (noun), Sarvnam (pronoun), Kriyapad (verb), Visheshan (Adjectives), Shabdyogi Avyay (Postposition), KriyaVisheshanAvyay (Adverb), Ubhayanvayi Avyay (Conjunction), and Kevalprayogi Avyay (Interjection)).

3.4 Combined Tagger Based Rules

We aim to develop the part of speech tagger for Marathi and for that we have used here the combined tagger to train the data.

Combined Tagger

The combination of taggers is one of the main features in natural language processing. The importance of the combined tagger is that when one tagger is unable to tag the word, then it would be passed another one is called as sequential backoff tagging.

In our case we took the default tagger as the backoff tagger because while we train the unigram tagger and in case it has not been tag the word then would be passed to the default tagger and it will tag the word with the default value like NN.

Actually the unigram tagger is the subclass of the n-gram tagger. N-gram also has the two subclasses like bigram and trigram namely. The unigram tagger as the name implies that only the single word of its context for assigning the part of speech tag. It is the context based tagger whose context is a single word. As the name implies the bigram tagger tags the two words, one is previous tag and another is the current tagged word. The result of tagging of the bigram is not on the basis of the context of the word and on the other hand the unigram tagger does not care about the previous word it is tag each current word. So while we combine both the tagger the result is better. Rules framed for the POS tagging, some of them given as below,

Table 4. Sample rule for POS tagging.

Rules For Noun	If suffixes or postpositions precedes the word: Then tag as the NN
Rules for Verb	If NN or JJ or RB in sentence: Then tag as VM Else NN or JJ or VM in sentence: Then tag as VAUX
Rule for Adjective	If preceding word of noun present in the sentence: Then tag as JJ

Algorithm for part of speech tagger:

Step 1: Take Marathi text as input
Step 2: Create a list of tokenized words of sentences of input text.
Step 3: Apply rule based POS tagger on the tokenized text
Step 4: If the tag is not correct repeat the Step 3
Step 5: Stop

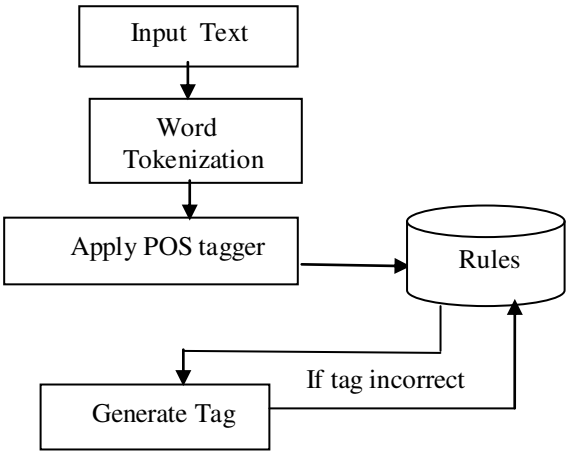

Fig.2. Workflow of the System

Main issues encountered while designing a rule based POS tagger

1. Some issues are encountered while tagging text of Marathi language because it has words like (बोलता-बोलता, दुखः) these words include the hyphen (-) and colon (:) but no other punctuation mark. But hyphen is come across into the sentences it is treated as the individual token. And the 'बोलता' and 'बोलता' also consider two different tokens in the sentence.
2. While doing the segmentation with the full stop at the end like (पु.ल.देशपांडे) sustain the segmentation ambiguity. The full stop denotes the end of the sentence as like English.
3. Another issue, the word in Marathi like 'नवी मुंबई', it is proper nounoun, but during the tokenization it is considered as two different tokens as 'नवी' and 'मुंबई'.
4. Variation in spelling: There are word containing four vowels in Marathi described in the below table,

Table 5. Vowels in Marathi

Vowel in Marathi	Vowel Sign
इ(i)	(ि◌)
ई(ī)	(◌ी)
उ(u)	(◌ु)

ऊ (ū)	(◌ू)

All the vowels do not make a phonetic difference, but those are differing in writing style and spellings like the words such as 'विराज'is grammatically correct, but 'वीराज' is incorrect or 'राणी'is grammatically correct but the 'रानि' is incorrect. These affect the part of speech tagger system because the correct word is not coming across into the training dataset.

5. Issue related to encoding:

Various fonts are used to write the Marathi text. When the character of language is opened on the computer, it is not displayed correctly means the document is unreadable or unusable. When the document created one computer with specific operating system may not display another computer with other configuration of the computer. In these cases, some character is readable but some are not fully displays there may be some hollow circles or some symbolic interpretation. For example, केंद्र is printed as केंद. Thus, all font types have to be traced and converted to one form, which is a very tedious job.

While tagging the sentence there are two verbs are adjacent each-other. First word act as main verb that indicate the incomplete action and second verb that indicate the complete action act as an auxiliary verb. But in some sentences more than one verb occurred. For example,'मोहन राधाला कुल्फी घेऊन देत होता.' In this type of example, the verb like 'घेऊन'and 'देत' are adjacent and one auxiliary verb that is 'होता'.

4. Result

In the following figure 3 and 4, POS tagger is shown with the tagged data on the training data and the testing data. Total 635 tagged sentences are considered for the development of the part of speech tagger. For training 90% (571 tagged data) of the tagged sentences and for testing 10% (126 tagged data) sentences are used from the total sentences. The below figure 3 depict the result of the combined tagger and we mainly concentrate on the unigram and bigram tagger, Total tagged words are 5715 and the size of the unigram tagger is 1918 words are tagged with unigram trained tagger and remaining untagged by the unigram is passed to the bigram tagger and bigram tagger is has successfully tagged 106 words. The table below shows some sentences and our trained tagger results. The overall result of the POS tagger for Marathi is 79.34%. the performance can be enhanced by increasing the training corpus.

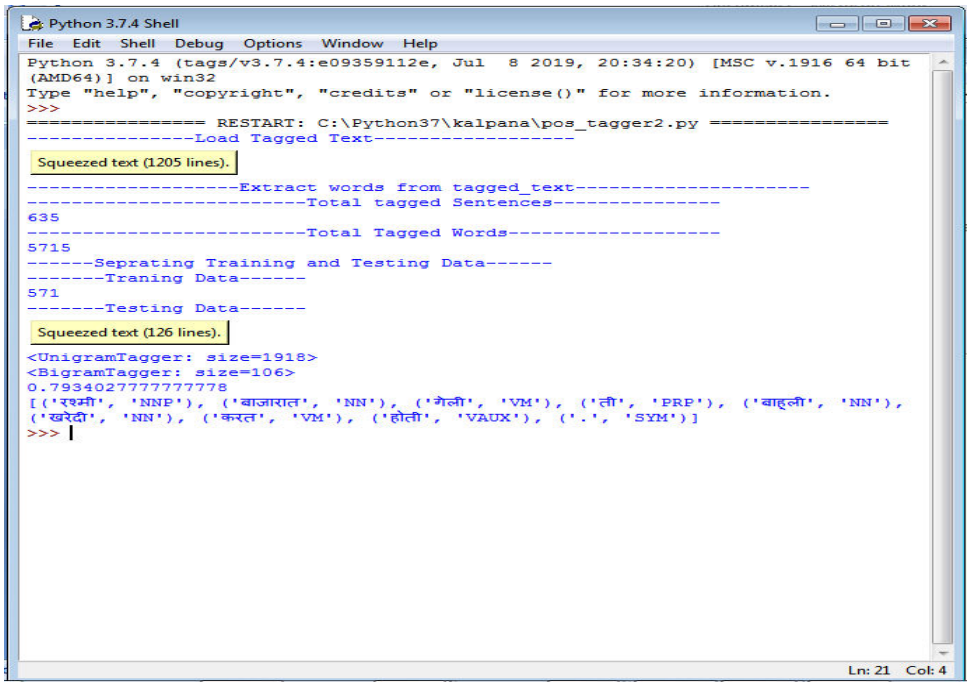

Fig.3. Result of the Marathi POS tagger

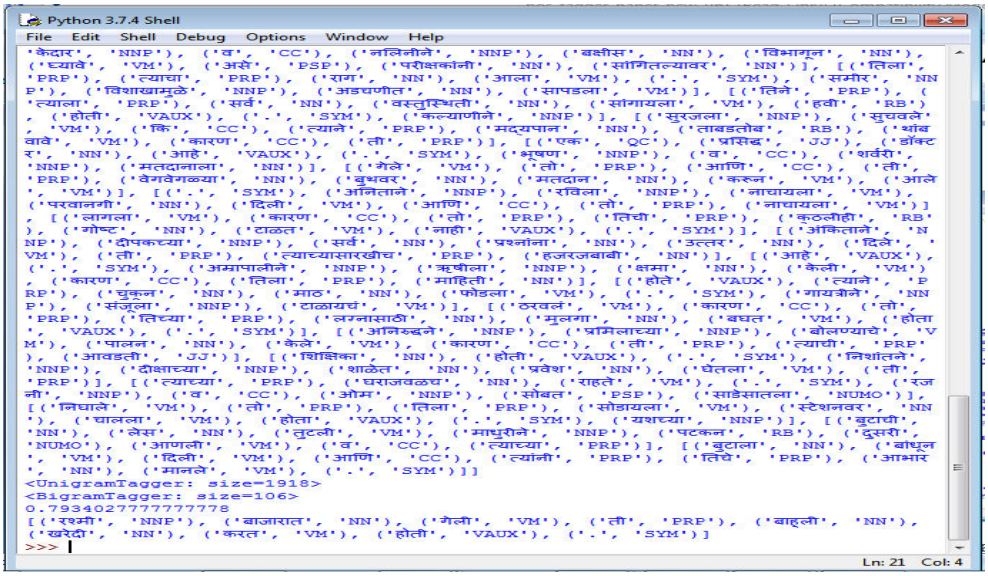

Fig.4. Sample of trained POS tagger

Table 4 Sample dataset with result

Sr. No.	Sentence	POS tagging

1.	राधाने चित्रपट बघितला, ती फार खुश झाली.	(‘राधाने’,’NNP’),(‘चित्रपट’,’NN’),(’बघितला’,’VM’),(‘,’,’SYM’),(‘ती’, ’PRP’),(‘फार’,’INTF’),(‘खुश’,’JJ’),(‘झाली’,’VM’),(‘.’,’SYM’)
2.	रश्मी बाजारात गेली, ती बाहुली खरेदी करत होती.	('रश्मी', 'NNP'), ('बाजारात', 'NN'), ('गेली', 'VM'), ('ती', 'PRP'), ('बाहुली', 'NN'), ('खरेदी', 'NN'), ('करत', 'VM'), ('होती', 'VAUX'), ('.', 'SYM')
3.	सुनील चांगला व्यक्ती आहे, तो सगळ्यांची मदत करतो.	(‘सुनील’,’NNP’),(‘ चांगला’,’JJ’),(‘ व्यक्ती’,’NN’),(‘आहे’,’VAUX’),(‘ ,’,’SYM’) (‘तो’,’PRP’), (‘सगळ्यांची’,’NN’),(‘ मदत’,’NN’),(‘ करतो’,’VM’),(‘.’,’SYM’)

5. Conclusion

Word order is a topological property of any language, hence the study of syntax word order is not complete when there is no deliberation or discussion on the word order of any language. The order of Marathi language is SOV (Subject Object Verb). Contemplate the formal and functional distinction of words and they are grouped into the different word classes called as the part of speech tagger. There are main three types of POS tagger that are rule based POS tagger, stochastic or statistic tagger and the transformation based tagger. Our work is based on the rule based on the combined tagger with the unigram and the bigram tagger. The work of the POS tagging for Marathi is based on the different types of rules. Total 5715 words tagged during the training process and the overall result of the POS tagger is 79.34%. In future we will be adding the more data for training for enhancing the performance of the POS tagger.

References

1. AlKhwiter, W., & Al-Twairesh, N. (2021). Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM. Computer Speech & Language, 65, 101138.
2. Bacon, G. (2020, May). Data-driven Choices in Neural Part-of-Speech Tagging for Latin. In Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages (pp. 111-113).
3. Yousif, J., & Al-Risi, M. (2019). Part of Speech Tagger for Arabic Text Based Support Vector Machines: A Review. ICTACT Journal on Soft Computing: DOI, 10.
4. Kurniawan, K., &Aji, A. F. (2018, November). Toward a standardized and more accurate Indonesian part-of-speech tagging. In 2018 International Conference on Asian Language Processing (IALP) (pp. 303-307). IEEE.
5. Kurniawan, K., &Aji, A. F. (2018, November). Toward a standardized and more accurate Indonesian part-of-speech tagging. In 2018 International Conference on Asian Language Processing (IALP) (pp. 303-307). IEEE.
6. Nita V. Patil (2018) POS Tagging for Marathi Language using Hidden Markov Model International Journal of Computer Sciences and Engineering E-ISSN:2347-2693
7. Ajees A P, Sumam Mary Idicula (2018) A POS Tagger for Malayalam using Conditional Random Fields International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018) Spl.

8. Mittal, S., Sethi, N. S., & Sharma, S. K. (2014). Part of speech tagging of Punjabi language using N gram model. International Journal of Computer Applications, 100(19).
9. Joshi, N., Darbari, H., &Mathur, I. (2013). HMM based POS tagger for Hindi. In Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013).
10. Singh, J., Joshi, N., &Mathur, I. (2013). Part of speech tagging of Marathi text using trigram method. arXiv preprint arXiv:1307.4299. Communications and Informatics (ICACCI), 2013 International Conference on (pp. 1554-1559). IEEE.
11. Dhanalakshmi, V., Shivapratap, G., &SomanKp, R. S. (2009). Tamil POS tagging using linear programming.
12. Patel, C., &Gali, K. (2008). Part-of-speech tagging for Gujarati using conditional random fields. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.
13. Ekbal, A., &Bandyopadhyay, S. (2008). Web-based Bengali news corpus for lexicon development and POS tagging. Polibits, (37), 21-30.
14. Singh, T. D., &Bandyopadhyay, S. (2008). Morphology driven Manipuri POS tagger. In Proceedings of the IJCNLP-08 Workshop on NLP for less privileged languages.
15. Dalal, A., Nagaraj, K., Sawant, U., &Shelke, S. (2006). Hindi part-of-speech tagging and chunking: A maximum entropy approach. Proceedings of the NLPAI Machine Learning Contest, 6.
16. Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., & Chute, C. G. (2005). Domain-specific language models and lexicons for tagging. Journal of biomedical informatics, 38(6), 422-430.