Domain Specific POS tagger for Marathi

[1]Vijay Dhangar, [2]Madhav Kankhar, [3]Kalpana Khandale, [4]C. Namrata Mahender,

[3]Assistant professor, Department of CS and IT, Dr.B.A.M.University, Aurangabad (MS), INDIA.

Abstract— **In the field of natural language processing there are many problems faced by the researcher and one of them is part-of-speech tagging (POS). In the application of natural language processing like question answering system, text summarization, anaphora resolution and many more the POS tagging is most important and the useful for the language understanding. Part-of-speech (POS) tagging is defined as to assign a correct tag to the sentence or word. POS tag include the noun, pronoun, verb, adverb, conjunction and so on.**

This paper present for the design of the part-of-speech tagging for the Marathi. The POS tagger for English is easily available for the tagging of the data but for Marathi the appropriate tagger is not available. This paper focuses on the development of the POS tagger for the Marathi to assign the correct tag to the word. This paper is domain specific it is based on the children's stories. This POS tagger gives the 67% correct tag of words in Marathi.

Keywords—Tokenization; Shallow parser; POS tagging; Marathi

I. INTRODUCTION

POS tagging is defined as to assigning correct tag to the given word. The POS tagging is the difficult task for Marathi. There are vast variety of text data available in Marathi and hence there is need to sort out it. POS tagging is one of the most important thing in the linguistic to understanding the natural language. In English we can easily tagged the word using POS tagger defined for the English but to tagging the word the POS tagger not available for the Marathi. Because of the ambiguity of the Marathi language there is big problem to tag the word properly. Hence we have design the POS tagger with the help of shallow parser (IIT-Hyderabad). But there is also problem with these parser some words are not tagged properly. To the development of the POS tagger we have used the shallow parser to find out the tag of the word but there is also a problem with the Marathi language, some words are not correctly tagged by the shallow parser.

There are various approaches for the POS tagger like rule based tagger, statistical tagger, hybrid tagger etc. The rule based tagger is based on the hand written rule and the statistical tagger is based on the probability and the frequency and the hybrid tagger is based on the combination of both the tagger. For designing the POS tagger we have used the tag set from developed by IIIT Hyderabad [1].

A. *Issues or challenges of the POS tagger:*

There are some issues in POS tagging of Marathi because of the ambiguity of the language there is problem to understanding to the system. We have use the shallow parser to understand the tag of the word but it is not given correct tag for the word.

For ex,

"अकबर/NN आणि/CC बिरबल/NN एकदा/RB फिरायला/VM गेले/NNP होते/NNP"

The above example shows the incorrect tag for the word गेले/NNP and होते/NNP. The correct tag for that word is the "गेले" as VM and the word "होते" is the VAUX.

Such types of issues are faced by the researcher now-a-days. Hence we trying to develop the POS tagger of Marathi language and trying to resolve the ambiguity of the Marathi sentences.

II. LITERATURE SURVEY

The following table shows the development of POS tagger for different languages. Various methods used by the researcher and they have faced some issues or challenges while developing the POS tagger. They have used the dataset and the overall result of the system.

TABLE I. LITERATURE REVIEW OF POS TAGGER

Author, Year and Language	**Methods/ Technique**	**Dataset**	**Issues/limitation**	**Result**
Patel, C., & Gali, K. 2008 **(Gujrati)** [2]	-Machine learning approach - Conditional Random Fields (CRF) model	-They have used 600 sentences in the dataset and 10,000 words for training and 5,000 words used for the testing.	-due to flexible nature of the language the unknown word made mistake while CRF using the features and the probabilities of the tag.	-While using CRF the accuracy of the system was 92%.

Asif Ekbal , Sivaji Bandyopadhyay 2008 **(Bengali)** [3]	-Hidden Morkov Model (HMM) -Support Vector Machine (SVM)	-Corpus of 72,341 tokens trained on the POS tagger.	-When small amount of data has been used to estimate the model parameters the HMMs do not work well.	-The overall result of the Bengali POS tagger using HMM and SVM is 91.23%.
Singh, T. D., & Bandyopadhyay, S. 2008 **(Manipuri)** [4]	-Machine learning approach -Accuracy	-They have used 3784 sentences containing 10917 unique words.	- The noun group words handling are not incorporated - The Noun-Adjective ambiguity disambiguation scheme is required as a separate module.	-The accuracy of the POS tagger is 69%.
Jyoti Singh , Iti Mathur &Nisheeth Joshi 2013 **(Marathi)** [5]	-Statistical Approach -Unigram -Bigram -Trigram -HMM methods	-a test corpus of 1000 sentences with 25744 words.	-They worked on limited corpora they should increase the amount of data for better accuracy of the system.	-The results of POS tagger is 93.82% which is impressive rather than other language systems.
Jyoti,Singh, Nisheeth Joshi,Iti Mathur. 2013 **(Marathi)** [6]	-Statistical approach -Trigram model -Accuracy	-They have developed test corpus of 2000 sentences with 48,635 words	- There are the morphological complexity of the Marathi makes it a little hard.	-The overall accuracy of the system is 91.63%.
Nisheeth Joshi, Hemant Darbari Iti Mathur 2013 **(Hindi)** [7]	-Hidden Morkov Model (HMM) -Precision -Recall -F-measures	-They have used 15,200 sentences with 3,58,288 words from tourism domain to trained the system.	They should improve the tagset and adding more tags for the less ambiguous classification of the text.	-The accuracy of the system is 92.13% on test data.
Nita.V. Patil 2018 **(Marathi)** [8]	-HMM used the Viterbi decoding algorithm -Unigram -Bigram -Trigram	They have used dataset consisting of 15,000 sentences from news stories domain	-POS tagging is challenging task for Marathi.	-The overall accuracy of the system is 86.61%.
Ajees A P, Sumam Mary Idicula 2018 **(Malayalam)** [9]	-Stastical approach -CRF Model -Maximum entropy Markov models	The system is trained on 23K words and tested on 5.7K words	-	The system performs with an accuracy of 91.2% on test data.

III. PROPOSED SYSTEM

The proposed system designed for the development of thr POS tagger for Marathi language. The tagging of the Marathi sentence is the most challenging task for the researcher. Here we have used dataset of 5 children stories and doing pre-processing on it. The flow of the tagger as shown in the Fig. 1

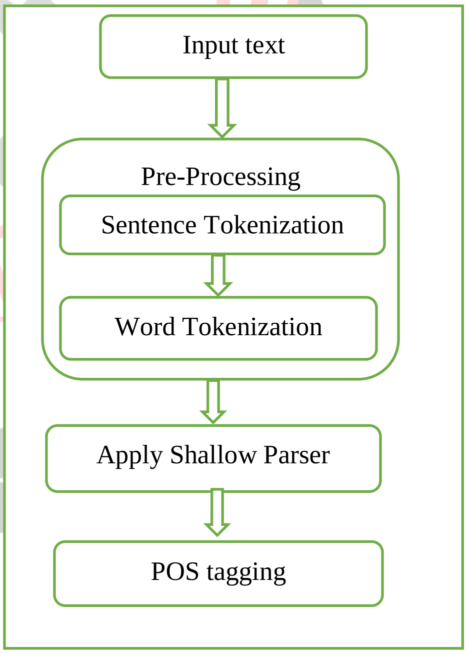

Fig. 1. **Architecture of POS tagger**

A. *Input Text:*

In this step we have to read the text input from the children's story.

B. *Pre-Processing:*

In this pre-processing step the sentence and word tokenization is done.

a) *Sentence Tokenization:*

The input paragraph is splitted or tokenize into the sentences is called as sentence tokenization. For example,

“अकबर आणि बिरबल एकदा फिरायला गेले होते. चालता चालता बादशाह अचानक थांबला.”

This is the sentence tokenization wherever (.) is occur it spit the paragraph into sentence.

b) Word Tokenization:

It is the term which tokenize the sentences into the words is called as word tokenization.

For example,

“'अकबर', 'आणि', 'बिरबल', 'एकदा', 'फिरायला', 'गेले','होते', '.' 'चालता', 'चालता', 'बादशाह', 'अचानक','थांबला', '.'”

The above example shows the word tokenization from sentences.

*C. **Shallow Parser:***

The shallow parser is used to parsing the sentences of the Marathi for understanding language. With the help of it we can find out the tags of the word of given input text. We have used here IIIT-Hyderabad shallow parser for Marathi. But it give incorrect tag for some words.
For example,

“अकबर/NN आणि/CC बिरबल/NN एकदा/RB फिरायला/VM गेले/NNP होते/NNP”

In the above example, the words एकदा/RB, गेले/NNP and होते/NNP tagged as incorrectly. The correct tag of this words एकदा as QC, गेले as VAUX and होते as VM.

*D. **POS Tagger:***

This is the ultimate step of the proposed system which we have design. While using the shallow parser we have got incorrect tag for the some words of dataset. When we apply our tagger on data it gives the better result. For example,
“अकबर/NN आणि/CC बिरबल/NN एकदा/QC फिरायला/VM गेले/VAUX होते/VM”
The above example of our tagger is correctly identify the tags of words.

IV. Conclusion

To assign the correct tag to the word in Marathi is difficult task. Because of the flexibility of the language the tagging became more complex for the researcher. Using the shallow parser we have compare the tags of the words in Marathi.

We have worked on the 5 children’s stories for the POS tagger. While comparing the shallow parser with our tagger it gives the better result than it. The overall result of tagger is 67% to tag the correct words in Marathi.

Acknowledgment

Authors would like to acknowledge and thanks to CSRI DST Major Project sanctioned No.SR/CSRI/71/2015(G), Computational and Psycholinguistic Research Lab Facility supporting to this work and Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India.

References

[1] Bharati, A., Sharma, D.M., Bai, L., Sangal, R., (2006) “AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages”.

[2] 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.

[3] Ekbal, A., & Bandyopadhyay, S. (2008). Web-based Bengali news corpus for lexicon development and POS tagging. Polibits, (37), 21-30.

[4] Singh, T. D., & Bandyopadhyay, S. (2008). Morphology driven Manipuri POS tagger. In Proceedings of the IJCNLP-08 Workshop on NLP for less privileged languages.

[5] 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.

[6] 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).

[7] Bagul, P., Mishra, A., Mahajan, P., Kulkarni, M., & Dhopavkar, G. (2014). Rule Based POS Tagger for Marathi Text. Proc. Int. J. Comput. Sci. Inf. Technol.(IJCSIT), 5(2), 1322-1326.

[8] Nita V. Patil (2018) POS Tagging for Marathi Language using Hidden Markov Model International Journal of Computer Sciences and Engineering E-ISSN:2347-2693

[9] 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.

DOI : 10.18231/2454-9150.2018.0875