IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-1 E-ISSN: 2347-2693

POS Tagging for Marathi Language using Hidden Markov Model

Nita V. Patil

School of Computer Sciences, North Maharashtra University, Jalgaon, India

*Corresponding Author: nvpatil@nmu.ac.in

Available online at: www.ijcseonline.org

Received: 02/Jan/2018, Revised: 10/Jan/2018, Accepted: 23/Jan/2018, Published: 31/Jan/2018

Abstract— Part-of-speech (POS) tagging plays significant role in almost every natural language processing task. This paper addresses a problem of POS tagging for Marathi language. Marathi is free word order, morphologically rich and highly inflectional Indian language. Supervised learning method that uses Hidden Markov Model is implemented to mark Marathi text using POS tags. The dataset required for training the algorithm consists of 12,000 Marathi sentences comprising news from popular Marathi newspaper. The algorithm for POS tagging predicts the tag for current word using the previous word tag pair. The POS tagging system has reported 86.61% accuracy in predicting correct POS to the words.

Keywords— Marathi, HMM, POS, Part of Speech, Tagset, Supervised learning

I. INTRODUCTION

Part of Speech (POS) tagging is a sequence labeling task, which assigns POS tags to words in the natural language text. It is a powerful tool for preprocessing required by many natural language processing applications. Many available POS taggers for English cannot be used to tag Marathi text. This has motivated researchers to implement POS tagging system for Marathi language. POS tagging in Marathi is depicted in following example.

Sentence:

केवळ अडीच मिनिटांची प्रतीक्षा आणि जास्तीत जास्त फक्त शंभर मीटर चालत जावे लागेल अशा पद्धतीने जर शहरांतर्गत प्रवासी वाहतूक यंत्रणा विकसित झाली तर?

POS tagged Sentence:

केवळ/INTF अडीच/QC मिनिटांची/NNP प्रतीक्षा/NN आणि/CC जास्तीत/QF जास्त/QF फक्त/INTF शंभर/NNP मीटर/NNP चालत/VM जावे/VM लागेल/VAUX अशा/CC पद्धतीने/NN जर/CC शहरांतर्गत/NN प्रवासी/NN वाहतूक/NN यंत्रणा/NN विकसित/RB झाली/VAUX तर/CC ?/SYM

POS tagging is a complex task because words may be marked with different tag categories in different context. This results in lexical ambiguity. The sentences shown in table 1 depicts the lexical ambiguity. Such ambiguities can be resolved by observing surrounding words of word tag pairs.

Table 1. Lexical Ambiguity

Example Sentence	Word	Grammatical Category	Meaning
हा कृष्ण धवल चित्रपट आहे.	धवल/ dhaval	Adjective	पांढरा/White
धवल नारायण पुढे या.		Proper noun	नाव/Name
उठा दिवाकर उगवला.	दिवाकर/ divakar	Common noun	सूर्य/Sun
दिवाकर रतिराम पाटील, उटखेडा.		Proper noun	नाव/Name

POS tagging is significant task in many NLP applications such as text parsing, machine translation, name entity recognition, opinion summarization etc. POS tagging adds structure to the document. The structured information assist computers to perform useful NLP tasks. The main approaches for development of POS tagger are linguistic approach where language and domain specific rules are required, or statistical or hybrid approach which could be combination of rule based and statistical approach. The lot of research related to POS tagging for English language is reported, limited research is reported for Indian languages especially for Marathi language. Marking Marathi text with POS tags is difficult task because Marathi is inflectional language where suffixes are added to words to add semantics in context. Patil et al. [1] explored issues and challenges related to text processing for Marathi. In this paper we discuss the development of a POS tagger for Marathi language using Hidden Markov Model (HMM).

The paper is organized into five main sections. The work related to POS tagging for Indian languages is described in Section 2. HMM based method POS tagging for Marathi text

is described in section 3. Section 4 discusses results and evaluation followed by section 5 which concludes the paper.

II. Related Work

POS tagging task attracted researchers in Indian languages context in 1990s. Bharti et al. [2] proposed a POS tagger for Hindi using morphological analysis. Singh et al. [3] have used morphological analysis and decision tree-based classifier for POS tagging. Shrivastava and Bhattacharya [4] proposed an approach that uses stemming to generate POS tags and reported 93.12% accuracy. Manju et al. [5] proposed the tagger using support vector machines for Malayalam. Authors have proposed a tagset for Malayalam, developed an annotated corpus and reported 94% accuracy with their own data and tagset. H.B. Patil et al. [6] reported a Part-of-Speech Tagging system for Marathi Language. The authors have applied morphological analysis to Marathi text, generated tokens, applied stemming to remove all possible affix and disambiguated using rule-based model for Marathi. Bagul, et al. [7] presented a POS Tagging system for Marathi language using rule based approach on tourism domain. The ambiguities in words are addressed using grammatical language rules. Jyoti Singh, et al. [8] presented a POS tagging system for Marathi language using statistical approach based on trigram method. Mishra et al. [9] proposed a POS sagging system for Hindi language with the accuracy 92.13%.

Mahar and Memon [10] proposed a system for Sindhi language based on rule Based approach. The system generates tokens from input. The selected token is compared with words from lexicons developed for POS tagging. If token found, then tag specified in lexicon is assigned to input token, else token is added into lexicon.

III. HMM Based POS Tagging for Marathi

A Hidden Markov Model is the statistical model based on markov chain rule where current state depends on previous state only. The model can be used to predict most probable state sequence for given observations. Supervised learning algorithm based on hmm requires large corpus for training. Thus, POS tagging system for Marathi language (figure 1) is described using dataset preparation, working of the system, and evaluating the performance.

A. Dataset Preparation

Assigning POS tags to Marathi words using HMM needs a POS tagged dataset. We used POS tag set proposed by IIIT Hyderabad [2] shown in table 2. FIRE-2010 (http://www.isical.ac.in/~fire/2010/) corpus is obtained for our research purpose. FIRE-2010 is a corpus developed by IIT Bombay by collecting four years (2004-2007) news stories from Marathi newspaper.

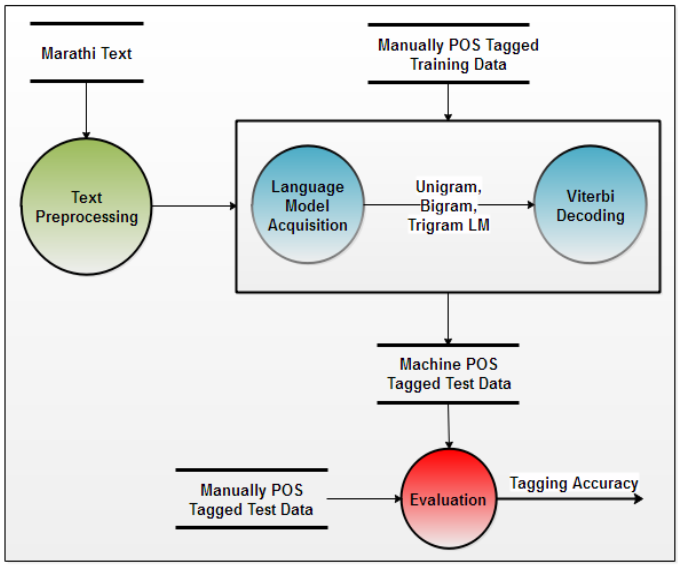

Figure 1. POS Tagging System for Marathi

Table 2. POS Tagset

Tag	Description	Tag	Description
NN	Common nouns	QF	Quantifiers
NST	Spatial and Temporal Expressions	QC	Cardinals
NNP	Proper nouns	CC	Conjuncts
PRP	Pronouns	WQ	Question words
DEM	Demonstrative pronouns	QO	Ordinals
VM	Finite or non-finite verbs	INTF	Intensifiers
VAUX	Auxiliary verbs	INJ	Interjections
JJ	Adjectives	NEG	Negative words
RB	Adverbs	SYM	Symbols
PSP	Postpositions	UNK	Foreign words
RP	Particles		

Algorithm 1 is implemented to develop POS tagged dataset for Marathi language. The dataset consisting of 15,000 sentences from news stories domain were used to train and test HMM based POS tagger. The task of development of POS tagged corpus is completed in six months. The dataset is further divided into two parts using in 80:20 (%) proportion. The large part of corpus i.e. 12,000 sentences (POS-TRAINING) consisting of average 15 number of words in each sentence were used to the train the HMM based POS tagger and small part of the corpus i.e. 3,000 sentences were used to create held out data. The POS tags and tokens are further separated to develop test dataset.

Algorithm 1: Proposed algorithm for Manual POS Tagging

Input:	Text Document (D_T)
Output:	POS Tagged Document (D_{POS})
Definitions:	S={POS tagset}
Begin	

1. Tokenize D_T using Stanford tokenizer. Mark the sentence boundary.
2. Mark the sequence of tokens to preserve context of tokens
3. Sort tokens to group similar tokens
4. Assign POS tag manually
5. Sort D_T to get original sequence of tokens.
6. Manually check for correct POS tagging
7. Return D_{POS} after corrections if any

End

B. Working of POS Tagger for Marathi

HMM uses interconnected states connected by their transition probability. The model consists of a finite set of possible words and tags with the parameters to model the probability of seeing the tag immediately after the bigram/unigram of tag(s) and probability of seeing observation paired with state. Transition probabilities between bigrams or trigrams are recorded in transition probability matrix. The relation between the occurrence of given tag and the set of words in which it occurs is recorded in emission probability matrix. The HMM consists of a finite set of words and possible tags with the maximum likelihood parameters (MLE) q and e. The parameter q for bigram language model is interpreted as the probability of seeing the tag v immediately after the tag u and value of parameter e is interpreted as the probability of seeing observation w paired with state u. The probability $P(w_1, w_2 \dots \dots w_n \mid t_1, t_2 \dots \dots t_n)$ where $w_1, w_2, \dots w_n$ is word sequence and $t_1, \dots \dots t_n$ is a tag sequence using maximum likelihood parameter is,

$$P(w_1, w_2 \dots \dots w_n \mid t_1, t_2 \dots \dots t_n) \cong \prod_{i=1}^{n} q(t_i \mid t_{i-1}) \times e(w_i \mid t_i) \quad (1)$$

and MLE parameters are computed as,

$$q(v|u) = \frac{C(u,v)}{C(u)}$$

$$e(w|s) = \frac{C(s \rightsquigarrow w)}{C(s)}$$

C. Tagging

Viterbi algorithm is used to find the most likely POS tag sequence for the given input word sentence. Most likely sequence of POS tags for a given word sequence is selected using markov chain model which assumes that the current tag depends only on the previous tag in the sequence. The algorithm based on Viterbi algorithm that uses back pointers is used for Viterbi decoding to predict most likely POS tag sequence for given input sequence.

IV. EVALUATION

Accuracy of POS Tagger is computed by comparing words tagged by the system with the hand annotated words and measured the accurate POS tag assignments. Unigram language model for POS tagging reported better performance on small training data. The performance of the system based on unigram language model decreased when the size of training dataset increased, and the satisfactory performance is reported by bigram language model based POS tagger with the training dataset consisting of 12,000 sentences. It is observed that the accuracy reported by trigram language model is poor compared to accuracy compared to bigram language model. The performance of the system using trigram language model can be increased with increase in training dataset size because many trigrams were absent in small training corpus. The accuracy is calculated using the formula,

$$Accuracy\ (\%) = \frac{No.of\ correctly\ POS\ tagged\ tokens}{Total\ No.of\ POS\ tags\ in\ the\ text} \times 100 \quad (2)$$

Performance of POS tagging system for Marathi language is as follows,

No. of correct POS tags assignments by the system = 21,246

Total no. of POS tags in the text = 24,532

Thus, the accuracy reported by the POS tagging system = 86.61%.

V. CONCLUSION

POS Tagging is plays significant role in many NLP tasks. Since Marathi is morphologically rich and free word order language, POS tagging is a challenging task. Hidden Markov Model technique is used to train and test POS tagger for Marathi. Unigram, bigram and trigram language models are used in our experiment. The HMM model uses Viterbi decoding algorithm for predicting most probable tag sequence for given word sequence for Marathi language. The system is trained using 12,000 sentences and tested using 3000 sentences in Marathi. The proposed system based on HMM algorithm has obtained satisfactory accuracy of 86.61% on test data. In future, this POS tagging system will be integrated with named entity recognition system in order to discover the effectiveness of POS tagging for named entity recognition for Marathi language.

REFERENCES

[1] Nita Patil, Ajay S. Patil and B. V. Pawar,*"Issues and Challenges in Marathi Named Entity Recognition* " International Journal on Natural Language Computing (IJNLC) Vol. 5, No.1, pp:15-31(2016) .

[2] Bharati, A., Sharma, D.M., Bai, L., Sangal, R., "*AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages*" (2006). http://ltrc.iiit.ac.in/tr031/posguidelines.pdf
[3] Singh Thoudam Doren and Bandyopadhyay Sivaji, "*Morphology Driven Manipuri POS Tagger*", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98, Hyderabad, India (2008)
[4] Shrivastava, M., Bhattacharyya, P., (2008) "*Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge*". In: International Conference on NLP (ICON08), Macmillan Press, New Delhi.
[5] Manju K., Soumya S., Sumam, M. I., (2009) "*Development of a POS Tagger for Malayalam - An Experience*". In International Conference on Advances in Recent Technologies in Communication and Computing, pp.709-713.
[6] H B Patil, A S Patil and B V Pawar. "*Part-of-Speech Tagger for Marathi Language using Limited Training Corpora*". IJCA Proceedings on National Conference on Recent Advances in Information Technology NCRAIT(4), 2014, pages 33-37.
[7] Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar, "*Rule Based POS Tagger for Marathi Text".* In proceeding of: International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1322-1326.
[8] Jyoti Singh, Nisheeth Joshi, Iti Mathur "*Part Of Speech Tagging Of Marathi Text Using Trigram Method*". International Journal of Advanced Information Technology (IJAIT) Vol. 3, No.2, DOI: 10.5121/ijait2013.3203.
[9] Nidhi Mishra, Amit Mishra, "*Part of Speech Tagging for Hindi Corpus".* In proceeding of International Conference on Communication Systems and Network Technologies, 978-0-7695-44373/11, 2011 IEEE DOI 10.1109/CSNT.2011.118.
[10] Javed Ahmed Mahar, Ghulam Qadir Memon, "*Rule Based Part of Speech Tagging of Sindhi Language*". In proceeding of International Conference on Signal Acquisition and Processing 978-0-7695-3960-7/10,2010 IEEE DOI 10.1109/ICSAP.2010.27.

Authors Profile

Mrs. Nita Patil pursued Bachelor of Computer Science (1999) and Master of Information Technology from North Maharashtra University, Jalgaon (2001). She has recently completed her Ph.D. in Information Technology (2017) and she is working as Assistant Professor in School of Computer Sciences, North Maharashtra University, Jalgaon since 2001. She is a of member of IAENG, since 2013. She has published several research papers in reputed national and international journals and conferences including IEEE. Her main research work focuses on Natural Language Processing. She has 15 years of teaching experience and 10 years of Research Experience.

