Marathi Parts-of-Speech Tagger Using Supervised Learning

Jyoti Singh, Nisheeth Joshi and Iti Mathur

Abstract In this paper, we present a parts-of-speech tagger for inflectional and derivational morphologically rich language Marathi. Marathi is spoken by the native people of Maharashtra. The general approach used for the development of tagger is statistical-based hidden Markov model (HMM). We establish a methodology of parts-of-speech (POS) tagging for Marathi using HMM. The main concept of HMM is to calculate probabilities to determine which is the best sequence of tags that correspond to observation sequence of words. In this paper, we show the development of the tagger. Moreover, we have also shown the evaluation done.

Keywords HMM · POS tagger · F-measure · Recall · Precision

1 Introduction

Parts-of-speech (POS) tagging is a process of assigning the words in a text as corresponding to particular parts of speech. POS tagging is also called word category disambiguation. A simplified version of POS tagging is the identification of words as nouns, verbs, adjectives, etc. POS tagging can be regarded as a simplified form of morphological analysis, where it only deals with assigning an appropriate POS tag to the word, whereas morphological analysis deals with finding the internal structure of the word.

J. Singh (✉) · N. Joshi · I. Mathur
Department of Computer Science, Apaji Institute, Banasthali University, Rajasthan, India
e-mail: jyoti.singh132@gmail.com

N. Joshi
e-mail: nisheeth.joshi@rediffmaial.com

I. Mathur
e-mail: mathur_iti@rediffmail.com

D. P. Mohapatra and S. Patnaik (eds.), *Intelligent Computing, Networking, and Informatics*, Advances in Intelligent Systems and Computing 243,
DOI: 10.1007/978-81-322-1665-0_24,

Indian languages are morphologically rich, and they have more than one morpheme of a word and due to this, tagging of Indian languages is difficult.

POS tagging is used in various applications like parsing where word and their tags are transformed into chunks, which can be combined to generate the complete parse of a text.

A POS tagger is a piece of software that reads text in some language and assigns parts of speech to each word. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction, and their subcategories. There are various approaches of POS tagging, which can be divided into two categories: rule-based tagging and statistical tagging.

The rule-based POS tagging models apply a set of handwritten rules and use contextual information to assign POS tags to words. These rules are often known as context frame rules. The earliest algorithms for automatically assigning POS were based on two-stage architecture. The first stage used a dictionary to assign each word a list of potential parts of speech. The second stage used large lists of handwritten disambiguation rules to bring down this list to a single POS for each word. The necessity of a linguistic background and manually constructing the rules are the main drawbacks of the rule-based systems. A stochastic approach includes frequency and probability or statistics. The simplest stochastic approach finds out the most frequently used tag for a specific word in the annotated training data and uses this information to tag that word in the unannotated text. The problem with this approach is that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules of a language.

2 Motivation

Automatic text tagging is an important concept in natural language processing. Till now no work has been done on POS tagging in Marathi. Therefore, there is a necessity to develop an automatic POS tagger for Marathi.

3 Problems of Parts-of-Speech Tagging

Ambiguous words are the main problems in parts-of-speech tagging. There may be many words, which can have more than one tag. Many words have multiple meanings. Sometimes it happens that a word has same POS but have multiple meaning.

For example,

नयन/**NNP** नी/PP नयन/**NN** मधे/CC झाकून/VM बघतिला/VAUX

The same word ‘नयन’ is given a different label in a sentence. In the first case, it is termed as a proper noun as it is referring to a name (person). In the second case,

it is termed as a common noun as it is referring to body part (eyes). POS tagging tries to correctly identify a POS of a word by looking at the context (surrounding word) in the sentence.

4 Previous Work on Indian Language POS Tagging

A lot of work has been done in POS tagging. There have been several researches carried out in this area. Singh et al. [2] proposed a Manipuri POS tagger using conditional random fields (CRF) and support vector machine (SVM). In this paper, they described a tagger for using CRF and SVM. Their evaluation result demonstrated the accuracies of 72.04 and 74.38 % in the CRF and SVM, respectively. Ekbal and Sivaji [3] proposed Web-based Bengali News Corpus for Lexicon Development and POS tagging, the POS tagger using hidden Markov model (HMM) and SVM. The POS taggers have been developed for Bengali and provide the accuracies of 85.56 % and 91.23 % for HMM and SVM, respectively. Dhanalakshmi et al. [4] present Tamil POS tagging using linear programming. In this paper, they propose the POS tagger for Tamil using machine learning techniques. They found that SVM-based machine learning tool affords the most encouraging result for Tamil POS tagger (95.64 %).

Kumar et al. [5] presented Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi. A statistical part-of-speech tagger for a morphologically rich language Hindi. This Tagger employs the maximum entropy Markov model with a rich set of features capturing the lexical and morphological characteristics of the language. The system achieved the best accuracy of 94.89 % and an average accuracy of 94.38 %. Singh et al. [6], in 2008, proposed POS tagging for Grammar Checking of Punjabi. In this paper, they discuss the issues concerning the development of a POS tagset and a POS tagger for use as a part of the project on developing an automated grammar checking system for the Punjabi Language.

In 2009, Manju et al. [7] proposed Development of a POS Tagger for Malayalam which was a Hidden Markov Model (HMM) based part of speech tagger for Malayalam language. The performance of the developed POS tagger is about 90 %, and almost 80 % of the sequences generated automatically for the test case were found correct. Joshi et al. [8] proposed POS tagging for Hindi. They have used IL POS tagset for the development of this tagger. They disambiguated correct word–tag combinations using the contextual information available in the text. They have achieved the accuracy of 92 %.

Patel et al. [9] proposed POS tagging for Gujarati using CRF. This paper describes a machine learning algorithm for Gujarati POS tagging. The machine learning part is performed using a CRF model. The algorithm has achieved an accuracy of 92 % for Gujarati texts where the training corpus is of 10,000 words and the test corpus is of 5,000 words. Reddy and Sharoff [10] proposed Cross Language POS taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. They use TnT (Brants 2000) [11], a popular implementation of the second-order Markov model for POS tagging.

5 POS Tagset

A number of POS tagsets are developed by different organization or person based on general principle of tagset design strategy. For POS annotation of texts in Marathi, we have use tagset developed by IIIT Hyderabad (Bharti et al. 2006) [1]. Table 1 shows brief description of IL POS tagset.

Table 1 Description of Marathi POS tagset

S.No	Tag	Description (Tag Used for)	Example
1.	NN	Common Nouns	मुलगा, साखर,मंडळी, सैन्य, चांगुलपणा
2.	NST	Noun Denoting Spatial and Temporal Expressions	मागे पुढे,वर,खाली
3.	NNP	Proper Nouns (name of person)	मोहन, राम, सुरेश
4.	PRP	Pronoun	मी,आम्ही,तुम्ही
5.	DEM	Demonstrative	तो, ती, ते, हा, ही, हे
6.	VM	Verb Main (Finite or Non-Finite)	बसणे,दिसणे, लिहिणे
7.	VAUX	Verb Auxiliary (Any verb, present besides main verb shall be marked as auxiliary verb)	नाही,नको,करणे ,हवे, नये
8.	JJ	Adjective (Modifier of Noun)	उत्साही,श्रेष्ठ,बळवान
9.	RB	Adverb (Modifier of Verb)	आता,काल,कधी,नेहमी
10.	PSP	Postposition	आणि,वर,कडे
11.	RP	Particles	भी, तो, ही
12.	QF	Quantifiers	बहुत, थोडा, कम
13.	QC	Cardinals	एक, दोन, तीन
14.	CC	Conjuncts (Coordinating and Subordinating)	आणि,केव्हा,तेव्हां, जर, तर
15.	WQ	Question Words	काय कधी कुठे कोण
16.	QO	Ordinals	पहिला दुसरा तिसरा
17.	INTF	Intensifier	खूप फार पुष्कळ
18.	INJ	Interjection	आहा,छान,अगो, हाय
19.	NEG	Negative	नाही,नको
20.	SYM	Symbol	? , ; : !
21.	XC	Compounds	काळे मांजर-काळमांजर, तेल पाणी- तेलवणी
22.	RDP	Reduplications	जवळ-जवळ
23.	UNK	Foreign Words	English, ગુજરાતી,

6 POS Tagger Procedure

The tagging process follows the following procedure (Fig. 1):

Tagset finder module contains information about words observed in the corpus. In tagset finder, each word is assigned a set of tags. The tagset finder supports fetching word information by providing information required to determine word feature. Statistics analyzer firstly splits the corpus into sentences and then splits the sentences into words. After that, it stores those words into lexicon table, which lies in statistics database. Tagger tags the words in a sentence with their corresponding tags. After the completion of tagging of words, the tester module provides us the test result.

7 Our Approach

In this paper, we are describing HMM for Marathi POS tagger. Our main aim is to perform POS tagging to determine the most likely tag sequences that generate the words of sentences where t_i denotes the tag sequence and w_i denotes the word sequence. Then, the following equation explains this fact

$$P(t_i|w_i) = P(t_i|t_{i-1}) \cdot P(t_{i+1}|t_i) \cdot P(w_i|t_i) \tag{1}$$

Here, $P(t_i|t_{i-1}) =$ the probability of a current tag given the previous tag
$P(t_{i+1}|t_i) =$ the probability of the future tag given the current tag

This provides the transition between the tags. These probabilities are computed by the following equation:

$$P(t_i|t_{i-1}) = \frac{\text{freq}(t_{i-1}, t_i)}{\text{freq}(t_{i-1})} \tag{2}$$

Each tag transition probability is computed by calculating the frequency count of two tags that come together in the corpus divided by the frequency count of the previous tag coming independently in the corpus. We used two special tags <S> and </S> to denote the starting of the sentence and the ending of the sentence, respectively, which was added to all the sentences of the training corpus.

8 Evaluation

For testing the performance of our system, we developed a test corpus of 1,000 sentences (25,744 words). We finally report results of all POS taggers in terms of recall, precision, and F-measure because they are considered to be standard performance indicators of a system.

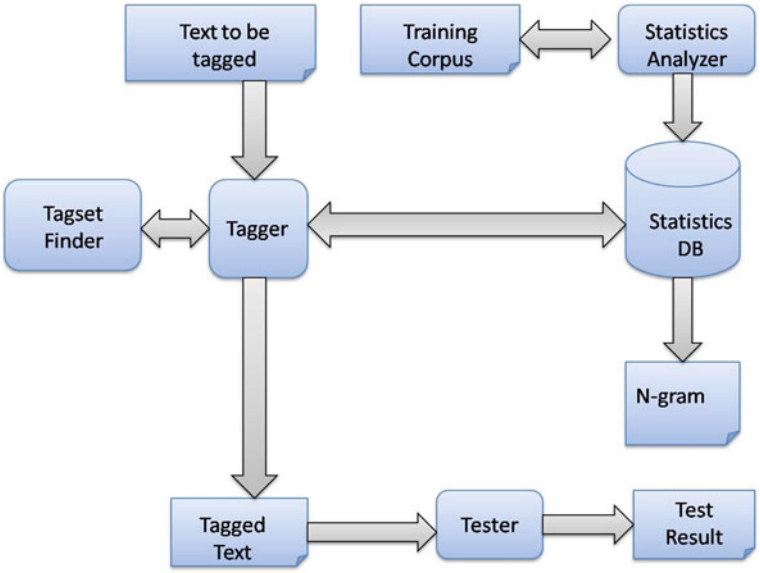

Fig. 1 Working diagram of POS tagger

Precision: The precision rate is the number of times that algorithm correctly identifies an event over the total number of times it actually identifies.

$$\text{Precision}(P) = \frac{\text{No. of correct POS tags assigned by the system}}{\text{No. of POS tags assigned by the system}}$$

Recall: The recall rate is the number of times that algorithm correctly identified an event over the total number of times that it actually occurred.

$$\text{Recall}(R) = \frac{\text{No. of correct POS tags assigned by the system}}{\text{No. of POS tags in the text}}$$

F-measure: F-measure is the weighted harmonic mean of precision and recall.

$$\text{F-measure} = \frac{2 \times P \times R}{P + R}$$

Test scores of our system are as follows:

Number of correct POS tags assigned by the system = 24,156
Number of POS tags assigned by the system = 25,744
Number of POS tags in the text = 25,744.

Thus, the accuracy of the system is 93.82 %.

9 Conclusion

In this paper, we have described a POS tagger for Marathi. The POS tagger described here is very simple and efficient for automatic tagging, but the morphological complexity of the Marathi makes it little hard. The performance of the current system is good, and the results achieved by this method are excellent. We believe that future enhancements of this work would be to improve the tagging accuracy by increasing the size of tagged corpus.

References

1. Bharti, A., Sangal, R., Sharma, D.M., Bai, L.: Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages. *LTRC-TR31* (2006)
2. Singh, T.D., Bandyopadhyay, S.: Morphology driven Manipuri POS tagger. In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 91–98. Hyderabad, India (2008)
3. Ekbal, A., Bandyopadhyay, S.: Web-based Bengali news corpus for lexicon development and POS tagging. In: Proceeding of Language Resource and Evaluation (2008)
4. Dhanalakshmi, V., Anandkumar, M., Rajendran, S., Soman, K.P.: Tamil POS tagging using linear programming. In: proceeding of International Journal of Recent Trends in Engineering, vol. 1, No. 2 (2009)
5. Dalal, A., Nagaraj, K., Swant, U., Shelke, S., Bhattacharyya, P.: Building feature rich pos tagger for morphologically rich languages: Experience in Hindi. In: Proceedings of International Conference on Natural Language Processing (ICON) at IIIT, Hyderabad (2007)
6. Gill, M.S., Lehal, G.S., Joshi, S.S.: Part-of-Speech tagging for grammar checking of Punjabi. Linguis. J. **4**(1), 6–21 (2009)
7. Manju, K., Soumya, S., Idicula, S.M.: Development of a POS tagger for Malayalam-an experience. In: Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, IEEE (2009)
8. Joshi, N., Darbari, H., Mathur, I.: HMM based POS tagger for Hindi. In: Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013) (2013)
9. Patel, C., Gali, K.: Part-Of-Speech tagging for Gujarati using conditional random fields. In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 117–122. Hyderabad, India (2008)
10. Reddy, S., Sharoff, S.: Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In: Proceedings of IJCNLP Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies. Thailand (2011)
11. Brants T.: Tnt: a statistical part-ofspeech tagger. In: Proceedings of the sixth conference on Applied natural language processing, ANLC '00, pp. 224–231, Association for Computational Linguistics, Stroudsburg, PA, USA (2000)