Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

Improve accuracy of Parts of Speech tagger for Gujarati language

Sirajuddin Y. Hala [1], Sagar H. Virani[2]

[1]Department of Computer Engineering, VVP Engineering College, Rajkot
[2]Ass.Prof. Computer Engineering Department, VVP Engineering College, Rajkot

Abstract — *Parts-of-Speech (POS) tagging is a technique for annotation of lexical categories to every word in input sentence. POS tagging is widely used for linguistic analysis of input text. It is very essential task and preprocessing step for all the natural language processing activities. A POS tagger takes a sentence from input data and assigns a unique parts of speech tag to each lexical item of the sentence. Parts-of-Speech (POS) tagger has been developed for many languages as a part of Natural Language Processing. Till now many POS Taggers are available for Indian Languages like Hindi, Bengali, Tamil, Telugu, Malayalam, and Panjabi etc. Gujarati is a resource poor Indo-Aryan Language on which very less languages processing work is done. Very fewer tools are available for language processing on Gujarati. The main concern of developing POS tagger for any Language is to improve accuracy of tagging and remove ambiguity in sentences due to language structure. This work focuses on developing Parts of Speech Tagger for Gujarati Language.*

Keywords – *Parts of speech tagging, Natural Language Processing, Gujarati.*

I. INTRODUCTION

In corpus linguistics, Parts-of-speech tagging (POS tagging), also called grammatical tagging, is the process of marking up a word in a text (corpus) as corresponding to their particular parts of speech, based on both its definition, as well as its context - i.e. Relationship with adjacent and related words in a phrase, sentence or paragraph.[1]

Once performed manually, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

Parts of speech tagger can be developed by various methods. It depends on the type of language for which tagger is being developed, Language structure and method used for tagging. Parts of speech tagging is useful in Information Retrieval (IR), Word sense Disambiguation, text to speech, Morphological Analysis etc. As it is pre-processing step for all these natural language processing activities, it is necessary to build a POS tagger with higher accuracy.

II. METHODS FOR PARTS OF SPEECH TAGGING

There are three methods for parts of speech tagging. Rule based, Stochastic (or Statistical) and Hybrid.

- Rule Based :-

Rule based approach for Parts of Speech tagging uses linguistic rules for deciding which POS tag to be assigned to the input word. It is oldest approach for parts of speech tagging. Rule based approach requires in depth knowledge of language. Natural languages are of complex nature. To apply language rules, also called context rules, in POS tagging, it necessitates vast linguistic knowledge.

- Stochastic Approach :

In this approach statistical or probabilistic model uses corpus. Hidden Markov Model is one of such techniques. Stochastic Models predict probabilities of POS tags from tagged corpus, in order to calculate the most likely tags of an input sequence. Stochastic models are of two types: Supervised and Unsupervised. Supervised methods for stochastic tagging will use tagged corpus only. Therefore it requires large amount of tagged data to achieve higher accuracy. On the other hand, unsupervised methods do not require any tagged corpus. Instead, these methods calculate probabilities required by statistical tagger by using automatic word groupings.

- Hybrid approach

Hybrid techniques for POS tagging are combination of both rule based and stochastic approach. It depends on working model implemented for particular language. In this approach, first rule based approach is used and than stochastic techniques can be applied. And vice versa is also possible

III. EXISTING WORKS

There are many works done in Natural Language Processing in Indian languages. Here in this table, comparison of various part of speech tagger for Indian languages is given along with size of used corpus and its accuracy.

No.	Language	Approach	Accuracy	Corpus Size
1 .	**Assamese**	HMM	87 %	10,000 words
2 .	**Bengali**	Hybrid, HMM	95%	50,000 words
3 .	**Gujarati**	Stochastic approach - CRF	92%	10,000 words
4 .	**Hindi**	HMM	93.12%	66900 words
		MEMM	94.38%	15562 words
5.	**Punjabi**	Bigram Model	92.16%	10,000 words
6.	**Telugu**	SVM based	95%	25,000 words
7.	**Tamil**	Hybrid, Morpheme Composition	95.92%	10,000 words

Table-1 Comparison of POS Taggers in Indian Languages [2][3][4][5][6][7][8]

IV. PROPOSED SYSTEM

4.1 Proposed System Architecture:

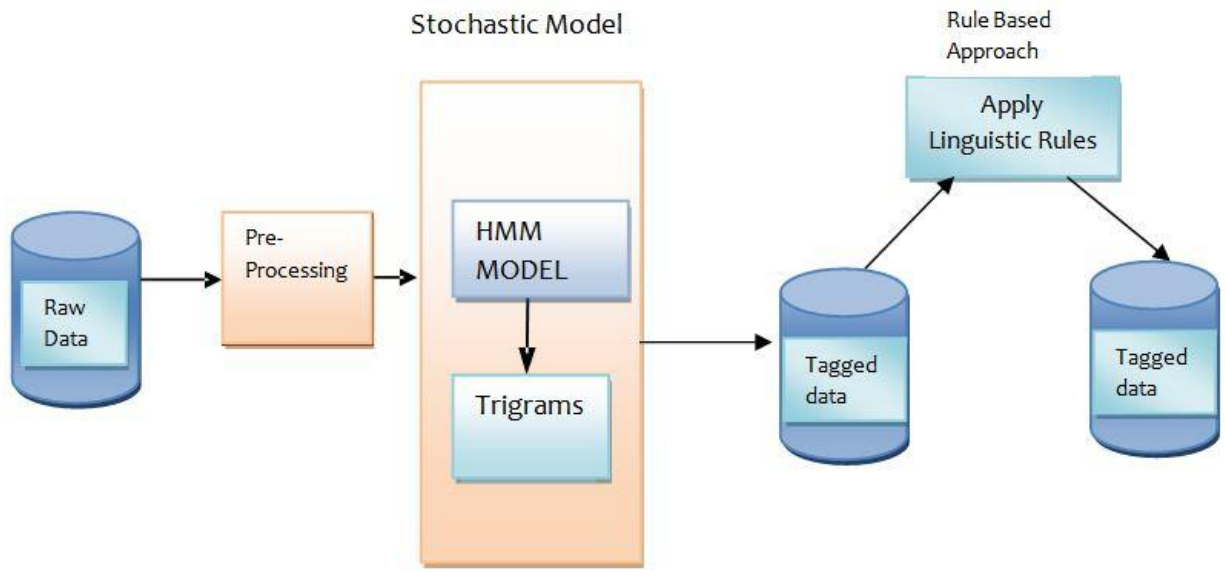

Fig-1 System architecture for Proposed POS tagger

The proposed system works in three phases.
1.Preprocessing
2.Apply stochastic method- HMM Model
3. Apply Linguistic Rules on tagged data

Phase-1 Pre-processing

In this phase pre-processing is done over the raw data. Resources necessary for tagging the text is generated i.e. Cleaning of data and tokenization of the data are done in this phase.

Tokenization:

```
>>> import nltk
>>>text="પાણીની ઊંડાઈ વધારે હતી."
>>>nltk.word_tokenize(text)
['પાણીની',' ઊંડાઈ,' વધારે','હતી', '.']
```

Phase-2 Apply Hidden Markov Model

In second phase Hidden Markov Model is applied on tokens. It is stochastic technique which computes probabilities of tags from corpus and predicts the best sequence.

HMMs are simple 3-tuple models described as

$\lambda(\Pi, A, B)$, where,

Π = Initial Probabilities B = Emission Probabilities

A = Transition Probabilities

For a given input sequence $W=(w_1,w_2,\ldots,w_n)$ we wish to determine a tag sequence $T=(t_1,t_2,\ldots,t_n)$ such that P(W, T) is maximized,

$$P(W, T) = \Pi^{N}_{i} [P(w_i|t_{1,i},w_{1,i-1})P(t_i|t_{1,i-1},w_{1,i-1})]$$

Hidden Markov Model are class of probabilistic models that assume that we can predict the probability of some future model without looking too far into the past. N-Gram approach, used in the proposed system, which looks into previous *n-1* words. A Trigram called as second order Markov model looks into two previous tags and computes probability of current word.[2][10]

Phase-3 Applying Linguistic Rule

Rule 1 : If token is having postfix like –યુક્ત,-વિહીન,-તર,-શીલ, -સભર then the token will be adjective.

e.g. લાગણીશીલ,કેસરયુક્ત ,અભ્યાસેતર, ગતિશીલ, ઉષ્માસભર

lAganiiviihiin,kEsaryUkta, abh-yAsEttar, gatiishIl, ushmAsabhar

all above words will be tagged as adjectives.

Rule 2: If in a sentence token is unknown and previous noun is pronoun followed by noun then it has to be a verb

e.g. તેઓનું સ્કૂટર ચાલી રહયુ છે.

Pronoun noun *verb* *verb*

Similarly other rules of Gujarati language can be applied in step of rules application on tokens. Some ambiguity can be solved by applying language rules to the tokens.

V. IMPLEMENTATION METHODOLOGY

5.1 Implementation Tools

This proposed work for developing part of speech tagger will use **Python** programming language. Python is included with an Open-source language processing library called **NLTK (Natural Language Toolkit).**

NLTK is having inbuilt tokenizer, chunker, pre-tagged corpus, various language processing tools. The hidden Markov Model trained on training data will be developed using NLTK library.

5.2 Tagset and Corpus

The tagset will be used for POS tagging of input data, is Bureau of Indian Standards (BIS) standard tagset (Annexure-I) developed for Gujarati Language. It contains 11 main Parts of Speech and further its subtypes for Parts of Speech tagging. Tagged corpus for this work manually developed of 10,000 sentences will be used.

V. CONCLUSION AND FUTURE EXPANSION

This work will focus on improving accuracy of Gujarati POS tagger using proposed hybrid approach. As in our approach both stochastic and rule based models used for tagging. It will resolve ambiguity and incorrect assigned tags to a word by applying Linguistic rules of Gujarati. On completion of the system implementation – precision, recall and f-measure will be calculated and compared with existing system. Based on that improvement in the proposed system will be done.

REFERENCES

[1] *"Part of Speech Tagging"*- http://en.wikipedia.org/wiki/Part-of-speech_tagging

[2] ***"Part of Speech Tagging using HMM"***. Available at: http://nlp.stanford.edu/courses/cs224n/2010/reports/parawira.pdf

[3] N.Saharia, D. Das,U.Sharma,J.Kalita **"Part of Speech Tagger for Assamese Text".** In proceedings of the ACL-IJCNLP 2009 conference, Singapore

[4] A.Dalal,K.Nagraj,U.Sawant,S.Shelke,P.Bhattacharya."***Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi",***In ICON 2007: International Conference on Natural Language Processing2007

[5] Chirag Patel, Karthik Gali *"**Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields**"*. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 117–122,2008

[6] N. Chandra, S.Kumawat,V.Srivastava "***Various tagsets for indian languages and their performance in part of speech tagging***", In Proceedings of 5th IRF International Conference, Chennai, 23rd March. 2014

[7] G.S Binulal, P. A. Goud, K.P.Soman***"A SVM based approach to Telugu Parts Of Speech Tagging using SVMTool"*** International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[8] Sumeer Mittal et al. "***Part of Speech Tagging of Punjabi Language using N Gram Model."*** International Journal of Computer Applications (0975 – 8887) Volume 100– No.19, August 2014

[9] "Unified Parts of speech standards for indian languages", http://www.tdildc.in/tdildcMain/articles/780732Draft%20POS%20Tag%20standard.pdf

[10] Jurafsky, D and Martin H. James, (2000), Speech and Language Processing, Prentice Hall.

Annexure-I

BIS TAGSET FOR GUJARATI LANGUAGE[9]

Sl. No	Category			Label	Annotation Convention**	Examples	Remarks
	Top level	Subtype (level 1)	Subtype (level 2)				
1	Noun			N	N		
1.2		Proper		NNP	N__NNP	મોહન,રવિ 'Mohan', 'Ravi'	
1.3		Nloc		NST	N__NST	ઉપર,અહીં નીચે	
2	Pronoun			PR	PR		
2.1		Personal		PRP	PR__PRP	હું,તું, 'me', 'you',	
2.2		Reflexive		PRF	PR__PRF	પોતે,જાતે સ્વયં 'herself/himself'	
2.3		Relative		PRL	PR__PRL	જે,તે,જ્યાં	

2.4		Reciprocal		PRC	PR__PRC	અરસ-પરસ,પરસ્પર 'mutually','e	
2.5		Wh-word		PRQ	PR__PRQ	કોણ, ક્યારે, ક્યાં 'who', 'when', 'where'	
2.6		Indefinite				કોઈક, 'someone',	
3	Demonstrative			DM	DM		
3.1		Deictic		DMD	DM__DMD	આ 'this'	
3.2		Relative		DMR	DM__DMR	જે,જેને 'which/who', 'whom'	
3.3		Wh-word		DMQ	DM__DMQ	કોણ ,શું, 'who',what	
3.4		Indefinite				કોઈ, 'someone',	
4	Verb			V	V		
4.1		Main		VM	V__VM	ખાવું,રમવું Eat,play	
4.2		Auxiliary		VAUX	V__VAUX	છે,હતું,કર્યું	
5	Adjective			JJ		મહત્વ	
6	Adverb			RB		મનોમન	
7	Postposition			PSP		સાથે With	
8	Conjunction			CC	CC		
8.1		Co-ordinator		CCD	CC__CCD	અને,અથવા	
8.2		Subordinator		CCS	CC__CCS	તેથી,જેવું કારણકે	
9	Particles			RP	RP		

9.1		Default		RPD	RP__RPD	પણ 'but',	
9.2		Interjection		INJ	RP__INJ	હે! અરે!	
9.3		Intensifier		INTF	RP__INTF	બહુ,ઘણુ	
9.4		Negation		NEG	RP__NEG	નહિ,નાં 'no'	
10	Quantifiers			QT	QT		
10.1		General		QTF	QT__QTF	થોડું,ઘણુ,	
10.2		Cardinals		QTC	QT__QTC	એક,બે,ત્રણ	
10.3		Ordinals		QTO	QT__QTO	પહેલું,બીજુ 'first',second	
11	Residuals			RD	RD		
11.1		Foreign word		RDF	RD__RDF	tv, perasitemol	
11.2		Symbol		SYM	RD__SYM	$, *,&	
11.3		Punctuation		PUNC	RD__PUNC	, : ; {} ()	
11.4		Unknown		UNK	RD__UNK		
11.5		Echowords		ECH	RD__ECH	કામ-બામ પાણી-બાણી	

