POS Tagging Approaches: A Comparison

Article *in* International Journal of Computer Applications · May 2015

DOI: 10.5120/20752-3148

CITATIONS
62

READS
6,002

2 authors:

Deepika Kumawat
Indian Institute of Technology Kanpur
1 PUBLICATION **62** CITATIONS
SEE PROFILE

Vinesh Jain
Malaviya National Institute of Technology Jaipur
5 PUBLICATIONS **166** CITATIONS
SEE PROFILE

POS Tagging Approaches: A Comparison

Deepika Kumawat
Department of Computer Science
Govt. Engineering College
Ajmer, Rajasthan

Vinesh Jain
Department of Computer Science
Govt. Engineering College
Ajmer, Rajasthan

ABSTRACT
Part of speech (POS) cataloguing is the process of allocating the part of speech tag or other philological class sign to each and every word in a sentence. In many Natural Language Processing presentations such as word intellect disambiguation, information recovery, information handling, analyzing, interrogating, and machine interpretation, POS tagging is reflected as the one of the basic obligatory tool. Categorizing the uncertainties in language philological items is the puzzling objective in the procedure of emerging an effectual and correct POS Tagger. Works survey displays that, for Indian lingoes, POS taggers were established only in Hindi, Punjabi, Bengali and Dravidian languages. Some POS taggers were also established generic to the Hindi, Telugu and Bengali tongues. All scheduled POS taggers were grounded on diverse Tag-set, established by diverse organization and individuals. This paper speaks the various developments in POS-taggers and POS-tag-set for Indian language, which is very essential computational verbal tool needed for many natural language processing (NLP) presentation [15].

Keywords
Tag-set, Ambiguity, Trigram, HMM, NPL, Tokenized, Indian Languages.

1. INTRODUCTION
Part of speech tagging is very significant pre-processing task for Natural language processing activities [1]. A Part of speech (POS) tagger has been developed in order to check off the words and punctuation in a textual matter having suitable POS labels of Hindi text. POS tagging makes up a primal task for processing a natural language. It is built up using linguistic theory rule, random pattern and sometimes a combining both [1]. My work shows the evolution of an easy and effective automatic tagger in support of inflectional and derivational morphologically rich language Hindi. Indian languages are morphologically rich with less linguistically peculiar patterns and rules and heavy annotated corpora and thus the development of POS tagger is a difficult task [6]. POS tagging is a phenomenon of allotting the words in a textual matter as matching to a picky component of speech. In general, POS tagging is as well denoted to as grammatical tagging of textual matter as representing to a specific component of speech because of both its definition and context.

A part-of-speech is a grammatical category, commonly including verbs, nouns, adjectives, adverbs, determiner, and so on.

1.1 Tagging
The process of assigning a part-of-speech or lexical class marker to each word in a collection. There are many potential distinctions we can draw leading to potentially large tag sets. To do POS tagging, we need to choose a standard set of tags to work with. We could pick very coarse tag sets as N, V, Adj, Adv.

Words	**Tag**
Sohan	N
Put	V
The	DET
Boy	N
On	P

1.2 Problems
The major problems in the process of POS tagging are: Ambiguous words and unknown words [2] . The first and foremost problem is with those words whose more than one tag can exist. This problem can be solved by emphasizing on context rather than single words. These can an easy task for humans but not so for the automatic word taggers. In the process of tagging we can sometimes get such words that have different tag categories when they are used in different context. Thus it is a very tedious job. This phenomenon is known as lexical ambiguity. But while occupying the same part of speech many words can have multiple meanings. Ambiguous words are the major problem in the part of speech tagging. Many words can have tags which are more than one [3]. Some words can have different meaning in different context but they have same POS. In order to solve such problem single word is considered rather than the context.

भारत/NN **सोने/JJ** की/CC चिड़िया/NN कहलाता/VM था/VAUX

अक्षय/NNP **सोने/VM** चला/VAUX गया/VAUX

2. CLASSIFICATION OF POS TAGGER
A Part-Of-Speech Tagger (POS Tagger) is defined as a part of software which assigns parts of speech to every word of a language that it reads. The approaches of POS tagging can be divided into three categories; rule based tagging, statistical tagging and hybrid tagging [1]. A set of hand written rules are applied along with it the contextual information is used to assign POS tags to words in the rule based POS. The disadvantage of this system is that it doesn't work when the text is not known. The problem being that it cannot predict the appropriate text. Thus in order to achieve higher efficiency and accuracy in this system, exhaustive set of hand coded rules should be used. Frequency and probability are included in the statistical approach. The basic statistical approach works on the basis of the most frequently used tag for a specific word in the annotated training data and also this information is used to tag that word in the unannotated text. But the disadvantage of this system is that some sequences of tags can come up for sentences that are not correct according to the grammar rules of a certain language. Another approach is also there that is known as the hybrid approach. It may even perform better than statistical or rule based approaches. First of all the probabilistic features of the statistical method are

used and then the set of hand coded language specific rules are applied in the hybrid approach. There are different types of statistical tagging approaches discussed in this paper that are- Unigram, Bigram and Trigram. Along with this the studies done on the basis of comparisons and evaluation are also shown.

POS tagging works on different approaches. The different models of POS tagging are shown in the following figure.

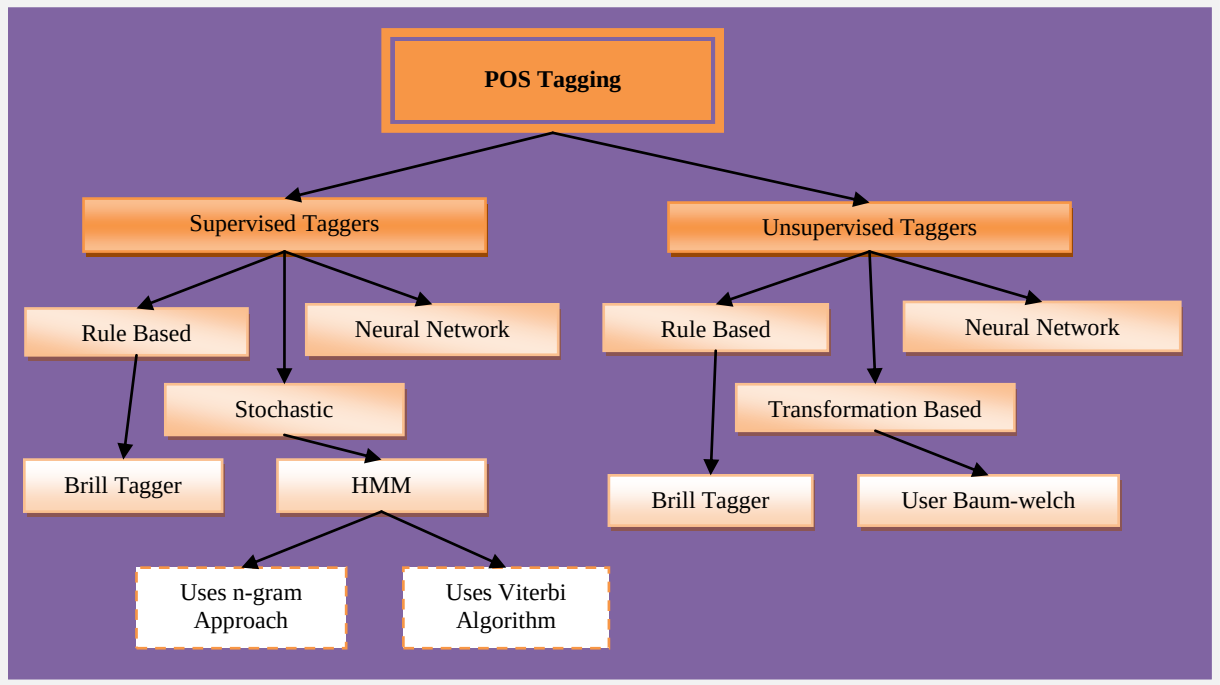

Fig 1: POS Classification

2.1 Supervised POS Tagging

Frequency or probability is the fundamentals used the Statistical taggers to tag the text. With the simplest Statistical tagger the problem of ambiguity of words based on the probability that word occurs with a particular tag can be resolved. The most common areas in which these tags are frequently used are the training set and are the one assigned to an ambiguous instance of that word in the testing data. Pre-tagged models are required by the supervised POS tagging models as they are used to learn information about the tag-set, word-tag frequencies, rule sets etc for training [13]. Increase in the size of corpora generally increases the performance of the models.

This approach is termed as the n-gram approach, which refers to the fact that the tag which is the best for a given word is determined by the probability which occurs with the n-1 previous tags. The drawback of this method is that it can of course retrieve a correct tag for a given word but along with this it can also sometimes retrieve invalid sequences of tags. The stochastic model is based on various models such as Hidden Markov Model (HMM), Maximum Likelihood Estimation, Decision Trees, N-grams, Maximum Entropy, Support Vector Machines and Conditional Random Fields [9].

2.1.1 Rule Based Approaches

The oldest part-of-speech tagging system was the one which used rule based approach. A set of hand written rules were applied and also contextual information was used in order to assign POS tags to words in the rule based POS tagging. These rules are generally known as context frame rules. Two-stage architecture was applied in the earliest algorithms for automatically assigning part-of-speech [10]. Firstly in the initial stage a dictionary is used in order to assign each and every word a list of potential parts of speech. After this in the second stage used large lists of hand-written disambiguation rules are used with the purpose to lessen down this list to just a single part-of-speech for each word.

Supervised training is required usually in the rule based tagging models that is pre-annotated corpora. The main disadvantages of the rule based systems are the necessity of a linguistic background and manually constructing the rules.

2.1.2 Stochastic

The frequency, probability or statistics are included in the stochastic approach. But the disadvantage of this approach can be that sometimes those sequence of tags can come which are not correct as per the grammar rules of a language. An approach which is known as the n-gram approach which calculates the probability of a given sequence of tags can be used as an alternative to the word frequency approach. The best tag can be determined by it for a word by finding out the probability that it occurs with the n previous tags, where the value of n is set to 1, 2 or 3 for practical purposes. These models are termed as Unigram, Bigram and Trigram [1]. Viterbi algorithm, which is a search algorithm that avoids the polynomial expansion of a breadth first search by trimming the search tree at each level using the best m Maximum Likelihood Estimates (MLE).

2.2 Unsupervised POS Tagging

The unsupervised POS tagging models is not like supervised models as they do not require pre-tagged corpora. Rather than this, they use advanced computational methods such as the

Baum-Welch algorithm so as to automatically induce tag sets, transformation rules etc.

There are basically two classes in which most of the tagging algorithms fall: rule-based taggers and stochastic taggers. The supervised approaches cannot be practically done easily to make them work in applicative settings but they reach the best performance in many NLP tasks [7]. Not only this, the supervised systems should be trained on a large amount of annotations which are manually provided.

2.2.1 Transformation Based Learning (TBL)

Brill described a system which learns a set of correction rules which helps to avoid linguistic rules that are manual. A set of rules is obtained by instantiating every rule template which has data from the corpus, with the help of predetermined rule template. This is done after the initialization process. The words that are tagged incorrectly are applied with each rule temporarily and hence the rule which reduces the maximum number of errors is identified and considered to be the best. Now this rule is added to the leaned rules and on the new corpus formed this process iterates by taking the recently added rule, because with the help of remaining rules, the reduction of error rate less than a predetermined threshold cannot be possible[5].

Both the transformation based approach and the rule based approach are similar as they depend on a set of rules for tagging. Initially, the tags to words are assigned based on a stochastic method. For example- for a particular word, the tag which has the higher frequency is assigned. Then to get the final result, the set of rules are applied to the initially tagged data.

3. IMPLEMENTATION OF STATISTICAL TAGGERS

3.1 Experimental Setup

3.1.1 Corpus Creation

Collection of text for corpus creation is a tedious job in Marathi language but because of availability of Books, News and Other informative Documents on web it become little bit easy but still Marathi document on web are limited rather than English.

In similar way for Part of Speech tagging we do not had tagged data in Marathi as compared to English, so to develop the annotated data we manually tagged 20,000 sentences for the part of speech tagging work.

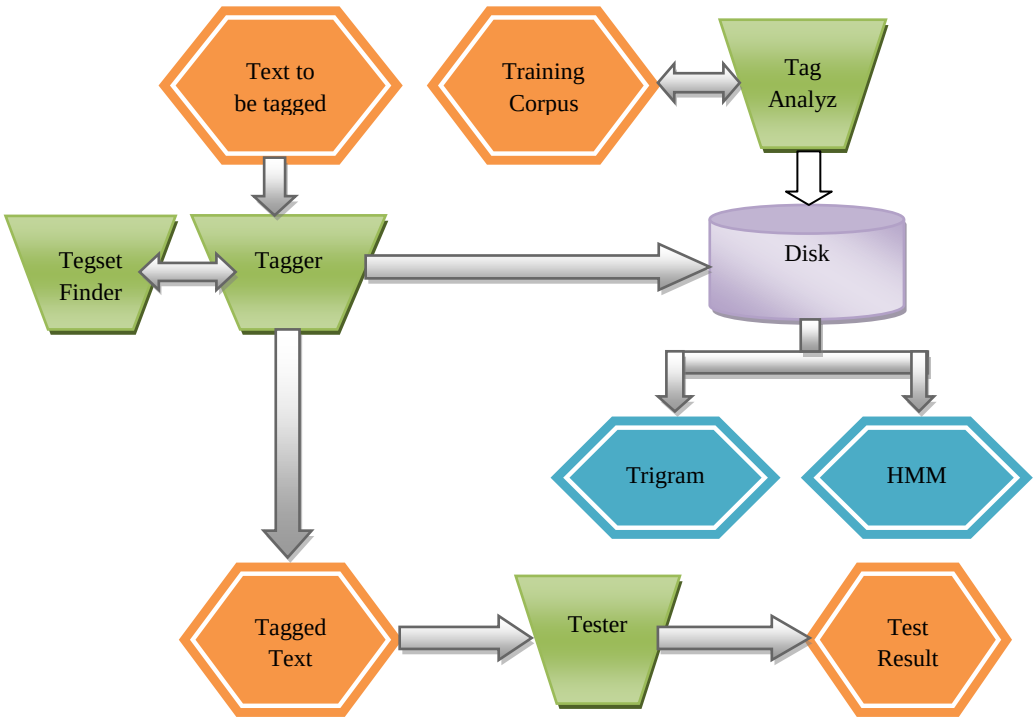

Fig 2: Working of POS Tagger

3.1.2 Tag-set finder

Tag-set finder module contains information about words observed in the corpus. In tag-set finder each word is assigned a set of tags. The tag-set finder supports fetching word information by providing information required to determine word feature.

3.1.3 Tag Analyzer

Tag Analyzer firstly split the corpus into sentences and then split the sentences into words. After that store those words into lexicon table which lies in Disk. 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.

3.1.4 N-Gram

3.1.4.1 Trigram

For describing Trigram Model for POS tagger, our main aim is to perform POS Tagging to determine the most likely tag for a word, given the previous two tags. So if t1, t2 …tn are tag sequence and w_1, $w_2 \ldots w_n$ are corresponding word sequence then the following equation explains this fact-

$P(t_i/w_i) = P(w_i/t_i).\ P(t_i/t_{i-2}, t_{i-1})$..... (1)

Where t_i denotes tag sequence and w_i denote word sequence. $P(w_i/t_i)$ is the probability of current word given current tag.

Here, $P(t_i|t_{i-2}t_{i-1})$is the probability of a current tag given the previous two tags.

This provides the transition between the tags and helps capture the context of the sentence. These probabilities are computed by following equation.

$P(t_i/t_{i-2}, t_{i-1}) = f(t_{i-2}, t_{i-1}, t_i)/f(t_{i-2}, t_{i-1})$…… (2)

Each tag transition probability is computed by calculating the frequency count of two tags which come together in the corpus divided by the frequency count of the previous two tags coming in the corpus.

3.1.4.2 HMM Tagger

The Idea behind Hidden Markov Model tagger is that "pick the most likely tag for the word" approach. After collecting statistical data of the tagged corpus from Tag analyzer, the tagger is activated on the test set which is already tokenized by the tokenizer [8]. The tagger employs a sentence based aapproach rather than a word based approach. That is, first all the possible tags for the words and the word sequences in the sentence are determined, and then the combination of the tags with the highest probability for the whole sentence is selected. A HMM is Statistical Model which can be used to generate tag sequences. Basic idea of HMM is to calculate or determine the most likely tag sequences. For this purpose we have to calculate Transition probability. Transition probability shows the probability of traveling between two tags i.e. forward tag and backward tags.
The Transition probability is generally estimated based on previous tags and future tags with the sequence provided as an input. The following equation explains this idea-

$P(t_i/w_i) = P(t_i/t_{i-1}).\ P(t_{i+1}/t_i).\ P(w_i/t_i)$……. (3)

$P(t_i/t_{i-1})$ is the probability of current tag given previous tag.

$P(t_{i+1}/t_i)$ is the probability of future tag given current tag.

$P(w_i/t_i)$ Probability of word given current tag

It is calculated as-

$P(w_i/t_i) = freq(t_i, w_i)/ freq(t_i)$……. (4)

This is done because we know that it is more likely for some tags to precede the other tags.

In HMM we consider the context of tags with respect to the current tag. Powerful feature of HMM is context description which can decides the tag for a word by looking at the tag of the previous word and the tag of the future word.

3.1.5 Tester

Tester performs testing based on 3 different domain test corpus. On the basis of that tester produces the result and give tagged data.

3.2 Tag set for Part Of Speech Tagging:

The significance of large annotated corpora in the present day NLP is widely known. It proves to be a basic building block for constructing statistical models for automatic processing of natural languages [14]. Depending on some general principle of tag-set design strategy, a number of POS tag-sets have been developed by different organizations. For developing tagger we were first required to annotate a corpus based on a tag-set. We used IL POS tag-set[14] proposed by Bharti et. Al. Table 2 shows brief description of the tags used. A detailed explanation can be sought from their paper. They have around 20 relations (semantic tags) and 15 node level tags or syntactic tags. Subsequently, a common tag-set has been designed for POS tagging and chunking of a large group of the Indian languages. The tag-set consist of 26 lexical tags. The tag-set was designed based on the lexical category of a word.

Sr. No.	Grammatical Word (Tag used for)	Tag	Example
1.	Common Noun	NN	कुर्सी, मेज़, लड़का, अजमेर, कलम
2.	Noun Denotating Spatial and Temporal Expressions	NST	पहले, बाद में, ऊपर, नीचे, आगे, पीछे
3.	Proper Nouns (name of person)	NNP	अक्षय, शुभम, हिमांशु, दीक्षा
4.	Pronoun	PRP	मैं, तुम, वह, हम, उसका, वो, तुम्हारा
5.	Demonstrative	DEM	वो, उस, यह, वह
6.	Verb Main (finite or non-finite)	VM	पड़ता, लिखता, खाता, सोता, खाते, सोते
7.	Verb Auxiliary (any verb, present besides main verb shall be marked as auxiliary verb)	VAUX	है, हुए
8.	Adjective (modifier of noun)	JJ	नयी, आधुनिक, सुनहरी, शानदार

9.	Adverb (modifier of verb)	RB	देर, जल्दी, धीरे,
10.	Postposition	PSP	ने, को, से, में
11.	Particles	RP	तो, ही, भी
12.	Quantifiers	QF	थोडा, बहुत, ज्यादा, कम
13.	Cardinals	QC	एक,दो ,तीन,चार
14.	Conjunctions (coordinating and subordinating)	CC	और,की,परन्तु,लेकिन
15.	Question Words	WQ	क्यों,कौन,क्यों,कब
16.	Ordinals	QO	पहला,दूसरा,तीसरा,चौथा
17.	Intensifier	INTF	बहुत,कम
18.	Interjection	INJ	अरे,हाय
19.	Negative	NEG	नहीं,ना
20.	Symbol	SYM	? , : ; !
21.	Compounds	XC	केंद्र/XC सरकार/NN रंग/XC बिरंगे/JJ
22.	Reduplications	RDP	बार/RB-बार/RDP गली/NN-गली/RDP
23.	Echo Words	ECH	प्यार-व्यार, चाय-वाय

4. PRACTICAL WORK

We apply Trigram and HMM methods on Hindi text. In order to measure the performance of our systems, we developed a test corpus of 3000 sentences. 1000 sentences belongs to tourism, 1000 sentences belongs to health and 1000 sentences belongs to general domain and finally report results of all POS taggers in terms of accuracy.

4.1 For Trigram

The accuracy was calculated using the formula:

Accuracy (%) = (No. of correctly tagged token/ Total no. of POS tags in the text)*100

4.1.1 For tourism sentences: Test scores of our system are as follows:

No. of Correct POS tags assigned by the system = 16958
No. of POS tag in the text = 18160
Thus the accuracy of the system is 93.38%.

संतो/**NN** ने/**PSP** किया/**VM** पुष्कर/**JJ** सरोवर/**NN** में/**PSP** शाही/**NN** स्नान|/**NN**

कार्तिक/**JJ** मास/**NN** के/**PSP** पंचतीर्थ/**NN** स्नान/**NN** के/**PSP** चलते/**VM** बुधवार/**NN** को/**VM** ब्रह्म/**NN** चतुर्दशी/**NN** के/**PSP** उपलक्ष्य/**NN** में/**PSP** दूरदराज़/**NN** से/**PSP** आये/**VM** संत-महात्माओं/**NN** ने/**PSP** सरोवर/**NN** में/**PSP** शाही/**JJ** स्नान/**NN** किया|/**NN**

4.1.2 For Health sentences:

No. of Correct POS tags assigned by the system = 18360
No. of POS tag in the text = 17059
Thus the accuracy of the system is 92.93%.

जेएलएन/NN के\PSP रेजिडेंट्स\NN आज\NN से\PSP हड़ताल\NN पर|\NN

कोटा/NN मेडिकल\XC कोलेज\NN में\PSP पिछले\JJ दिनों\NN रेजिडेंट\NN डॉक्टर्स\NN के\PSP साथ\NST हुई\VM मारपीट\NN मामले\NN में\PSP आरोपियों\NN के\PSP खिलाफ\PSP:? कार्यवाही\NN ना\NEG होने\VAUX पर\PSP प्रदेशभर\NN में\PSP रेजिडेंट\NN डॉक्टर्स\NN में\PSP रोष\NN पनपने\VM लगा\VM है|\NN

4.1.3 For General sentences:

No. of Correct POS tags assigned by the system = 16906
No. of POS tag in the text = 18247
Thus the accuracy of the system is 92.65%.

ट्रक/NN चालक\NN की\PSP दिलेरी\NN ने\PSP बचाई\NN २०\NN बस\RP यात्रियों\NN की\PSP जिंदगी|\NN

गलत/JJ दिशा\NN में\PSP जा\VAUX रही\VAUX बस\RP को\PSP बचाने\PSP के\PSP लिए\XC ट्रक\NN को\PSP खाई\NN में\PSP गिरा\VM दिया|\NN

गुरुवार/NN सुबह\NN डायटा\NN बांध\NN के\PSP पास\NST यात्रियों\NN से\PSP भरी\VM एक\QC निजी\JJ बस\RP चालक\NN ने\PSP कार\NN को\PSP ओवरटेक\NN किया|\NN

Average accuracy of Trigram model is- 92.98%.

4.2 For HMM

The accuracy was calculated using the formula:
Accuracy (%) = (No. of correctly tagged token/ Total no. of POS tags in the text)*100

Test scores of our system are as follows:

4.2.1 For tourism sentences:

No. of Correct POS tags assigned by the system = 17301
No. of POS tag in the text = 18160
Thus the accuracy of the system is 95.27%.

संतो/NN **ने**/PSP **किया**/VM **पुष्कर**/JJ **सरोवर**/NN **में**/PSP **शाही**/NN **स्नान**|/NN

कार्तिक/JJ **मास**/NN **के**/PSP **पंचतीर्थ**/NN **स्नान**/NN **के**/PSP **चलते**/VM **बुधवार**/NN **को**/VM **ब्रह्म**/NN **चतुर्दशी**/NN **के**/PSP **उपलक्ष्य**/NN **में**/PSP **दूरदराज़**/NN **से**/PSP **आये**/VM **संत-महात्माओं**/NN **ने**/PSP **सरोवर**/NN **में**/PSP **शाही**/JJ **स्नान**/NN **किया**|/NN

In above sentence HMM assigns correct tag.

4.2.2 For Health sentences:

No. of Correct POS tags assigned by the system = 18360
No. of POS tag in the text =17744
Thus the accuracy of the system is 96.64%.

जेएलएन/NN के\PSP रेजिडेंट्स\NN आज\NN से\PSP हड़ताल\NN पर|\NN

कोटा/NN मेडिकल\XC कोलेज\NN में\PSP पिछले\JJ दिनों\NN रेजिडेंट\NN डॉक्टर्स\NN के\PSP साथ\NST हुई\VM मारपीट\NN मामले\NN में\PSP आरोपियों\NN के\PSP खिलाफ\PSP कार्यवाही\NN ना\NEG होने\VAUX पर\PSP प्रदेशभर\NN में\PSP रेजिडेंट\NN डॉक्टर्स\NN में\PSP रोष\NN पनपने\VM लगा\VM है|\NN

4.2.3 For General sentences:

No. of Correct POS tags assigned by the system = 17240
No. of POS tag in the text = 18252
Thus the accuracy of the system is 94.46%.

ट्रक/NN चालक\NN की\PSP दिलेरी\NN ने\PSP बचाई\NN २०\NN बस\RP यात्रियों\NN की\PSP जिंदगी|\NN

गलत/JJ दिशा\NN में\PSP जा\VAUX रही\VAUX बस\RP को\PSP बचाने\PSP के\PSP लिए\XC ट्रक\NN को\PSP खाई\NN में\PSP गिरा\VM दिया|\NN

Average accuracy of HMM model is- 95.45%

4.3 Results

The results obtained from our taggers are summarized in below, each column corresponding to one of the above methods output respectively.

Table 1. Average results of all the taggers

Trigram	HMM
92.98%	95.45%

Studying the resulting tagged corpora we concluded that Most of the errors could be categorized as follows:

a. Errors in the case of the word are the highest. Those are partially due to the fact that some of the tags do not reflect the case of the word, and hence it is hard for the learner to conclude the reason of the next word being given its tag, examples of that are proper nouns, common noun and pronouns.
b. Unknown proper nouns (of people and places) cannot be guessed. Only few rules may lead to realizing a proper noun. Having a large corpus would reduce this problem by inserting many names in the lexicon.
c. Distinction between adverb and compounds is not easily guessed by some methods.

Taking in consideration the large and rich tagset we worked with, and the unavailability of a standard truth corpus, we think the results obtained here are very promising, and can be enhanced by many actions like: enlarging the training corpus, and enhancing the lexical analysis program. **We are presently working in this direction.**

5. CONCLUSION

Natural Language is the medium for communication which is incorporated by every human being. One of the most important activities in processing natural languages is Part of Speech tagging. In POS Tagging we assign a Part of Speech tag to each word in a sentence and literature. POS tagging is one of the simplest, most constant and statistical model for many NLP application. POS Tagging is an initial stage of linguistics, text analysis like information retrieval, machine translator, text to speech synthesis, information extraction etc. Since many of the companies like Google and Microsoft are concentrating on Natural language processing applications. Currently many tools are available to do this task of part of speech tagging. The POS tagger described here is very simple and efficient for automatic tagging.

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 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. The Hybrid approaches use a pre-defined set of handcrafted rules as well as automatically induced rules that are generated during training.

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.

6. ACKNOWLEDGMENTS

I want to give my sincere thanks to my husband Mr. Akshay Kumawat, beloved parents, my In-Laws, and teachers. And also to who participated in the study. I would like give a very special thanks to Mr. Sameer Meherishi for his support.

7. REFERENCES

[1] Jyoti Singh, Nisheeth Joshi, Iti Mathure, "Development of Marathi Part of Speech Tagger Using Statistical Approch"

[2] Dhanalakshmi V, Anand Kumar1, Shivapratap G, Soman KP and Rajendran S, "Tamil POS Tagging using Linear Programming", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[3] Gurleen Kaur Sidhu, Navjot Kaur, " Role of Machine Translation and Word Sense Disambiguation in Natural Language Processing", IOSR Journal of Computer Engineering (IOSR-JCE), May. - Jun. 2013.

[4] Asif Ekbal and Shivaji Bandyopdhyay, (2008) "Web-based Bengali News Corpus for Lexicon Development and POS Tagging", In Proceeding of Language Resource and Evaluation.

[5] Siva Reddy, Serge Sharoff, (2011) "Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources". In Proceeding of IJCNLP workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies.

[6] Dinesh Kumar and Gurpreet Singh Josan, (2010) "Part of Speech Tagger for Morphologically rich Indian Language: A survey". International Journal of Computer Application. Vol. 6(5).

[7] Singh Thoudam Doren and Bandyopadhyay Sivaji, (2008) "Morphology Driven Manipuri POS Tagger", Proceeding of Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98, Hyderabad, India.

[8] Nisheeth Joshi, Hemant Darbari, Iti Mathure, (2013) "HMM based Pos Tagger for Hindi". In Processing of 2013 International Conference on Artificial Intelligence and Soft Computing.

[9] Hasan Fahim Muhammad, Zaman Naushad Uz and Mumit Khan, (2007) "Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages", In proceeding of Center for Research on Bangla Language Processing.

[10] Aniket Dalal, Nagraj Kumar, Uma Sawant, Sandeep Shelke and Pushpak Bhattacharyya, (2007) "Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi". In Proceeding of International Conference on Natural Language Processing (ICON).

[11] Chirag Patel, Karthik Gali, (2008) "Part of Speech Tagging for Gujarati Using Conditional Random Feilds", In Proceeding of IJCNLP-08 Workshop on NLP for Less Privileged Language, pp 117-122.

[12] Mandeep Singh, Gurpreet Lehal, and shiv Sharma, (2009) "Part-of-Speech Tagging for Grammar Checking of Panjabi" in Proceeding of The Linguistics Journal Volume 4 Issue.

[13] Manju K, Soumya S, Idicul S. M., (2009) "A Development of A POS Tagger for Malayalam – An Experience" In Proceeding of International Conference on Advance in Recent Technologies in Communication and Computing.

[14] Akshar Bharti, Dipti Misra Sharma, Lakshmi bai, Rajeev Sangal. AnnCorra: Annotating Corpora Guidelines for POS and Chunk with Annotation For Indian Languages , Language Technologies Research Centre IIT, Hyderabad.

[15] Antony P J, Amrita, Dr. K P Soman, "Parts Of Speech Tagging for Indian Languages: A Literature Survey", IJCA (0975-8887) Volume 34-no. 8, November 2011.