ORIGINAL RESEARCH PAPER

Volume-7 | Issue-4 | April-2018 | PRINT ISSN No 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

PARTS OF SPEECH TAGGING FOR HINDI LANGUAGES USING HMM

Engineering

Rajesh Kumar M.Tech Scholar, Arya Institute of Engineering & Technology, Jaipur

Sayar Singh Shekhawat Associate Professor, Arya Institute of Engineering & Technology, Jaipur

ABSTRACT

This paper describes the Part of Speech (POS) tagging for Indian Languages "HINDI". Part of Speech tagging is the one of the most basic problems of Natural language processing NLP. Part of speech (POS) tagging is the process of assigning the part of speech tag or other lexical class marker to each and every word in a sentence. A lot of POS tagging work has been done by the researchers for various languages using different approaches HMM (Hidden Marcov Model), SVM(Support Vector Machine), ME (Maximum Entropy). HMM approaches concerned for POS tagging of sentences written in Hindi languages are discussed in this paper. This paper also discussed a hybrid based approach, for tagging hindi language.

KEYWORDS

Part of Speech tagging(POS), Natural Language Processing(NLP) , Indian Languages HINDI

INTRODUCTION

Parts of Speech (POS) Tagging is an initial stage of information extraction, summarization, retrieval, machine translation, speech conversion. POS-tagging is the process of assigning the part of speech tags to the natural language text based on both its definition and its context. Identifying the POS-tags in a given text is an important aspect of any Natural Language Application.

Parts of speech (POS) tagging is one of the most well studied problems in the field of Natural Language Processing (NLP). Natural language processing is the skill of a computer program to understand human language as it is spoken. It is a component of computer science, linguistics and artificial intelligence. To build NLP application is a difficult because human speech is not always specific.

If tag is found then we assign the tag to Hindi word with the help of Hindi wordnet dictionary.

Hindi POS tagging using Hindi WordNet dictionary and HMM.

Input Hindi sentence	POS tagged output
पुस्तक मेज पर पड़ी है।	पुस्तक <Noun> मेज <Noun> पर <postprepos> पड़ी है <Verb> |<Punc>

Summarized Statement: in above table all the words are tagged with the help of Hindi WordNet dictionary and Hidden Markow Model (HMM).

LITERATURE REVIEW

A detailed review of 18 research papers on POS tagging system, published within the period of year 2006 to the year 2018 is presented in this section. The review process based on five stage analysis as discussed in the previous section was adopted.

- Get the Feel
- Get Big Picture
- Get the Details
- Evaluate the Details
- Synthesize the Details

S. No.	Issue	Types of Papers	Number of Papers
1.	POS Tagging System	Different Conferenced in Journals	18

The Different approaches have been used for part-of speech tagging and different researchers have developed POS taggers for various languages Foreign Languages like English, Arabic and other European languages have more POS taggers than Indian languages. Indian Languages for which POS taggers have been developed are Panjabi, Hindi, Bengali, and Tamil.

[Avinesh.PVS, Karthik G, 2006] "Conditional Random Fields and Transformation Based Learning" In this paper we describe Part Of Speech (POS) tagging and Chunking using Condi-tional Random Fields (CRFs) and Transfor-mation Based Learning (TBL) for Telugu, Hindi and Bengali. The overall results obtained for POS tagging is 77.37% and for chunking it is 79.17% (Telugu). [1]

[Er.Davinder Kaur1, Er.Ubeeka Jain,2017] "Automatic Rule Detection and POS Tagging of Punjabi Text" In this paper study the problem and generate "automatic rule detection and POS Tagging of Punjabi text" by using rule based approach. supervised learning approach provides better results according to this paper. [2]

[Pratibha Singh, Aditya Tripathi, 2017] "Hindi Language text search : a literature review" The literature review focuses on the major problems of hindi text searching over the web and the review availability of number of techniques. supervised learning approach provides better results. [3]

[M.C. Padma, R.J. Prathibha, 2016] "Morpheme based parts of speech tagger for kannada Language" This paper presents a Morpheme based parts of speech tagger for kannada language. This proposed work uses hierarchical tag set for assigning tags. The system is tested on some kannada words taken from emille corpus. Experimental result shows that the performance of The proposed system is above 90%. The Board of indian standards, dravidian and hierarchical Tag set is used to assign parts of speech tags. It is Shown that the performance of morpheme based pos Tagger is better even without using manually Pre-tagged training data set and statistical or machine Learning algorithms. [4]

[Aniket Dalal, Kumar Nagaraj, Uma Sawant, Sandeep Shelke, 2006] "Hindi Part-of-Speech Tagging and Chunking : A Maximum Entropy Approach" System has good performance with an average accuracy of 88.4% for POS tagging and 86.45% for chunking, with best accuracies being 89.35% and 87.39% for POS tagging and chunking, respectively. [5]

[Archit Yajnik,2017] "Part of Speech Tagging Using Statistical Approach for Nepali Text" This article presents POS tagging for Nepali text using Hidden Markov Model and Viterbi algorithm. From the Nepali text, annotated corpus training and testing data set are randomly separated. Viterbi algorithm is found to be computationally faster and accurate as compared to HMM. The accuracy of 95.43% is achieved using Viterbi algorithm. [6]

[Arnav Sharma, Raveesh Motlani, 2016] "POS Tagging For Code-Mixed Indian Social Media Text : Systems from IIIT-H for ICON NLP Tools Contest" POS tagger using only the given dataset, namely a constrained system, which gave an accuracy of 75.04% after being evaluated on unseen test dataset. The text had sentences composed of a mixture of words in English and one of the three Indian languages, Hindi, Bengali and Tamil. [7]

[Manish Shrivastava, Pushpak Bhattacharyya,2008] "Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge" In this paper, we present a simple HMM based POS tagger, which employs a naive(longest suffix matching) stemmer as a pre-processor to achieve reasonably good accuracy of 93.12%. This method does not require any linguistic resource apart from a list of possible suffixes for the language. [8]

PROBLEM DEFINITION

The problem statement of the dissertation work is **"**Part of speech tagging of Hindi language using Hybrid approach".

Objective:

- To design the knowledge base of Hindi Corpus by collecting data from social media.
- To perform data pre-processing on the created Hindi Corpus.
- To design and implement the POS tagging on Hindi corpus using Hybrid Approach(Rule Based +Statistical) to tag the token
- To carry out the performance analysis of proposed system.

POS TAGGING TECNIQUES

Part of Speech Tagger is an important tool that is used to develop information extraction and language translator. The problem of tagging in natural language processing is to find a way to tag every word in a text as a particular part of speech. Part of Speech tagger is an important application of natural language processing. It is an important part of morphological analyzer. Part of speech tagging is the process of assigning a part of speech like noun, verb, preposition, pronoun, adverb, adjective or other lexical class marker to each word in a sentence.

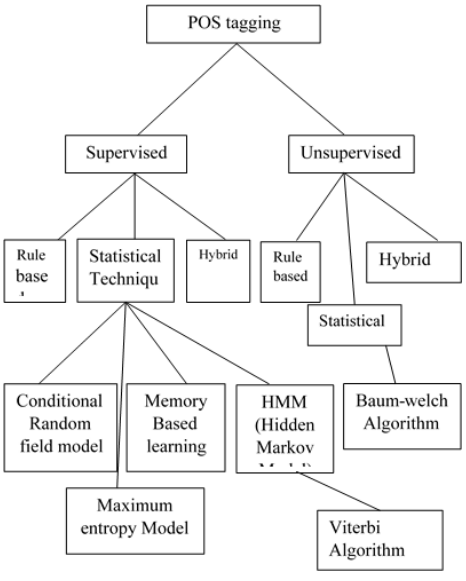


Figure 1.1 POS tagging Techniques

Supervised POS taggers

The supervised POS tagging models require on pre-tagged corpora which are used for training to learn information about the word-tag frequencies, rule and tagset, sets etc. The performance of the models generally increases with the increase in size of these corpora.

Unsupervised POS tagger

The unsupervised POS tagging models do not require pre-tagged corpora. Instead, they use those methods through which automatically tags are assigned to words. Advanced computational methods like the Baum-Welch algorithm to automatically include tag sets, transformation rules etc. Again supervised and unsupervised techniques are fallen into three subcategories:

a) Rule based
b) Stochastic or Statistical based POS tagger
c) Hybrid

Rule Based Approach / Transformation Based

The rule based POS tagging approach that uses a set of hand written rules. Rule base taggers depend on word list or lexicon or dictionary to assign appropriate tag to each word. The tagger divided into two stages. First, it search words in dictionary and second, it assigns a tag by removing disambiguity of words using linguistic features of word . On the basis of level rule divided as lexical rules act in a word level, each sentence splits into small words called lexeme or token And, the context sensitive rules act in a sentence level, to check the grammar for the sentence . The transformation based approach is similar to the rule based approach in the sense that it depends on a set of rules for tagging. The transformation based approaches use a pre-defined set of handcrafted rules as well as automatically induced rules that are generated during training. The main drawback of rule based system is that it fails when the text is not present in lexicon. Therefore the rule based system cannot predict the appropriate tags.

Statistical Approach / Stochastic Tagger:

Stochastic tagger as a simple generalization of the stochastic taggers generally resolves the ambiguity by computing the probability of a given word (or the tag).The common machine learning models used for POS tag are:

Maximum Entropy Markov Model

MaxEnt stands for Maximum Entropy Markov Model (MEMM). It is a conditional probabilistic sequence model. It can represent multiple features of a word and can also handle long term dependency. It is based on the principle of maximum entropy which states that the least biased model which considers all known facts is the one which maximizes entropy.

Conditional Random Field Model

CRF stands for Conditional Random Field. It is a type of discriminative probabilistic model. It has all the advantages of MEMMs without the label bias problem. CRFs are undirected graphical models (also known as random field) which are used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

Hybrid Models

Hybrid models are basically combination of rules based and statistical models. In Hybrid system, approach uses the combination of both rule-based and ML technique and makes new methods using strongest points from each method. It is making use of essential feature from ML approaches and uses the rules to make it more efficient.

Hidden Markov Model (HMM)

HMM stands for Hidden Markov Model. HMM is a generative model. The model assigns the joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets. It is advantageous as its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. It uses only positive data, so they can be easily scaled. It has few disadvantages. In order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Hence it makes various assumptions about data like Markovian assumption i.e. current label depends only on the previous label. Also it is not practical to represent multiple overlapping features and long term dependencies. Number of parameter to be evaluated is huge. So it needs a large data set for training.

Sample of ambiguity

INPUT	Tagged output
पर्यावरण वह होता है जो प्राकृतिक रूप से, हमारे चारों **तरफ** होता है |	पर्यावरण <Noun> वह <pron> होता है <verb> जो <pron> प्राकृतिक <Adj> रूप <Noun> से, <postprepos> हमारे <pron> चारों <Noun> **तरफ** <Adj> **तरफ** <Noun> होता है <verb> |<Punc>

Total No of Words	Untagged	Tagged words	
		Total POS Tag	Total Correct word Tag
735	26	639	709

Summarized Statement: in above table "तरफ" has two tag as adjective and noun .this situation is called Amibiguity. So for this purpose we use HMM.

Table: Input parameters

Number Of Dataset	Number of Sentences	Number of words
10	100	735

For evaluation of the experimental results, Standard IR (Information Retrieval) performance measures accuracy, precision, recall and f-measure are used. The values of the above performance measures are calculated, for existing and proposed systems compared

Table: Experiment results of proposed system

EVALUATION METRICS

The evaluation metrics for the data set is precision, recall and F-Measure. These are defined as following:

$$\text{Precision} = \frac{\text{Total Number of correct word tag by POS after HMM}}{\text{total number of Word}}$$

$$\text{Recall} = \frac{\text{Number of correct tag by POS}}{\text{total number of correct word tag by POS after HMM}}$$

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

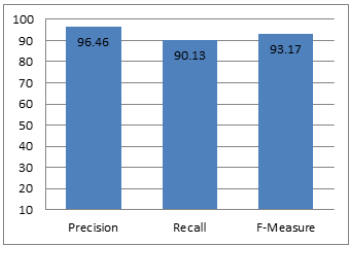
The results of our proposed system are given in following table.

Table: Experimental results in terms of precision ,Recall , F-measure.

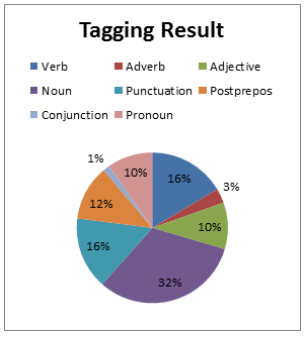
EXPERIMENTAL

We collected total 100 Hindi sentences from different websites.

The graph plotted for the obtained result is shown in Figure 1.



The graph plotted for the obtained result For Different Tag. is shown in Figure 2.



Performance Parameters	Precision
Precision	96.46
Recall	90.13
F-Measure	93.17

CONCLUSION

A part-of-speech tagger is a system that uses context to assign parts of speech to words The performance analysis has been carried out for Precision, Recall and F1-Measure. we obtained 93.17 % precision ,96.46 % Recall and 90.13 % F-measure. we also calculated the performance parameter for each tag. Finally Future work has also been discussed as; to deal with complex sentences .in future we also do Hindi to English translation.

REFERENCES

1. Avinesh.PVS, Karthik G " Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning", 2006.
2. Er. Davinder Kaur, Er.Ubeeka Jain "Automatic Rule Detection and POS Tagging of Punjabi Text", 2017 International Journal Of Engineering And Computer Science(IJECS).
3. Pratibha Singh and Aditya Tripathi "Hindi Language Text search: a Literature Reviw", 2017 Annals of Library and Information Studies.
4. M. C. Padma, R. J. Prathibha "morpheme based parts of speech tagger for kannada Language" 2016 international journal of management and applied science.
5. Aniket Dalal, Kumar Nagaraj, Uma Sawant, Sandeep Shelke, "Hindi Part-of-Speech Tagging and Chunking : A Maximum Entropy Approach",2005 IJCNLP.
6. Archit Yajnik "Part of Speech Tagging Using Statistical Approach for Nepali Text",2017 International Journal of Cognitive and Language Sciences.
7. Arnav Sharma, Raveesh Motlani "POS Tagging For Code-Mixed Indian Social Media Text : Systems from IIIT-H for ICON NLP Tools Contest",2016.
8. Neetu Aggarwal, Amandeep kaur Randhawa "A Survey on Parts of Speech Tagging for Indian Languages",2015 International Conference on Advancements in Engineering and Technology (ICAET)
9. Manish Shrivastava, Pushpak Bhattacharyya "Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge", 2008 http://ltrc.iiit.ac.in/proceedings/ICON
10. Nisheeth Joshi, Hemant Darbari and Iti Mathur, "HMM based POS tagger for hindi", 2013 Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA.
11. Rijuka Pathak, Somesh Dewangan "Natural Language Chhattisgarhi: A Literature Survey", 2014 International Journal of Engineering Trends and Technology (IJETT)
12. Antony P J, Dr. Soman K P "Parts Of Speech Tagging for Indian Languages: A Literature Survey", 2011 International Journal of Computer Applications.
13. Ramandeep Kaur, Lakhvir Singh Garcha, Dr. Mohita Garag, Satinderpal Singh" Parts of Speech Tagging for Indian Languages Review and Scope for Punjabi Language ",2017 International Journal of Advanced Research in Computer Science and Software Engineering.
14. Shachi Mall, Umesh Chandra Jaiswal "Survey: Machine Translation for Indian Language", 2018 International Journal of Applied Engineering Research
15. Shubhangi Rathod, Sharvari Govilkar "Survey of various POS tagging techniques for Indian regional languages", 2015 International Journal of Computer Science and Information Technologies
16. Atul Kr. Ojha Pitambar Behera, Srishti Singh and Girish Nath Jha "Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case ofHindi, Odia and Bhojpuri".