ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
Volume 4, Issue 1, February 2014

Hybrid approach for Part of Speech Tagger for Hindi language

Kanak Mohnot, Neha Bansal, Shashi Pal Singh, Ajai Kumar

***Abstract*— 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. In this paper, we present a Hybrid Based Part of Speech Tagger for Hindi. Our System is evaluated over a corpus of 80,000 words with 7 different standard part of speech tags for Hindi. Accuracy is the prime factor in evaluating any POS tagger so the accuracy of proposed tagger is also discussed in this paper.**

***Index Terms*— POS, Tagging, Rules, Hybrid, HMM.**

I. INTRODUCTION

Natural Language Processing is a rapidly growing technology at present and also it is a very important resource for fetching information from collections of huge amount of data with the help of imposing some queries and keywords. But there is problem of fetching information what the user exactly wants because it contains more than one document related to a particular thing, person or incident etc. For instance, when we search for some data in the repository with the help of some query, we may get a lot of un-important or irrelevant data instead of getting the exact data or information. So, in order to fetch the exact information from large collection of documents what the user exactly wants there is great need of some methods or mechanisms. This leads to the Information Extraction Research. Information Extraction (IE) is a method which helps in extracting the required or exact data. It is the process of fetching the required information from large collection of documents what the user want.[1]

Morphology is the field of the linguistics that studies the internal structure of the words. Morphological Analysis and generation are essential steps in any NLP Application. Morphological analysis means taking a word as input and identifying their stems and affixes. Morphological Analysis is essential for Hindi it has a rich system of inflectional morphology as like other languages.[2] Morphological Analyzer and generator is a tool for analyzing the given word and generator for generating word given the stem and its features (like affixes).

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.[3][8] There are a number of approaches to implement part of speech tagger, i.e. Rule Based approach, Statistical approach and Hybrid approach.

A. Rule Based Approach

It uses linguistic grammar-based techniques to find tags. It needs rich and expressive rules and gives good results. It requires great knowledge of grammar and other language related rules. Good experience is needed to come up with good rules and heuristics. It is not easily portable and has high acquisition cost. It is very specific to the target data. [12]

B. Statistical Methods

The common machine learning models used for POS tag are:

1) ***Hidden Markov Model****:* - 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. [7][3]

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. [8]

2) ***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. Each source state has an exponential model that takes the observation feature as input and output a distribution over possible next state. Output labels are associated with states. [10]

The large dependency problem of HMM is resolved by this model. Also, it has higher recall and precision as compared to HMM. The disadvantage of this approach is the label bias problem. The probabilities of transition from a particular state must sum to one. MEMM favors those states through which less number of transitions occurs. [9]

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
Volume 4, Issue 1, February 2014

3) ***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 know as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes. [11]

C. 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. [6]

II. System Description

This system is developed using hybrid based approach and 7 different standard part of speech tags. The system mainly works in two steps-firstly the input words are found in the database; if it is present then it is tagged. Secondly if it is not present then various rules or HMM model is applied.

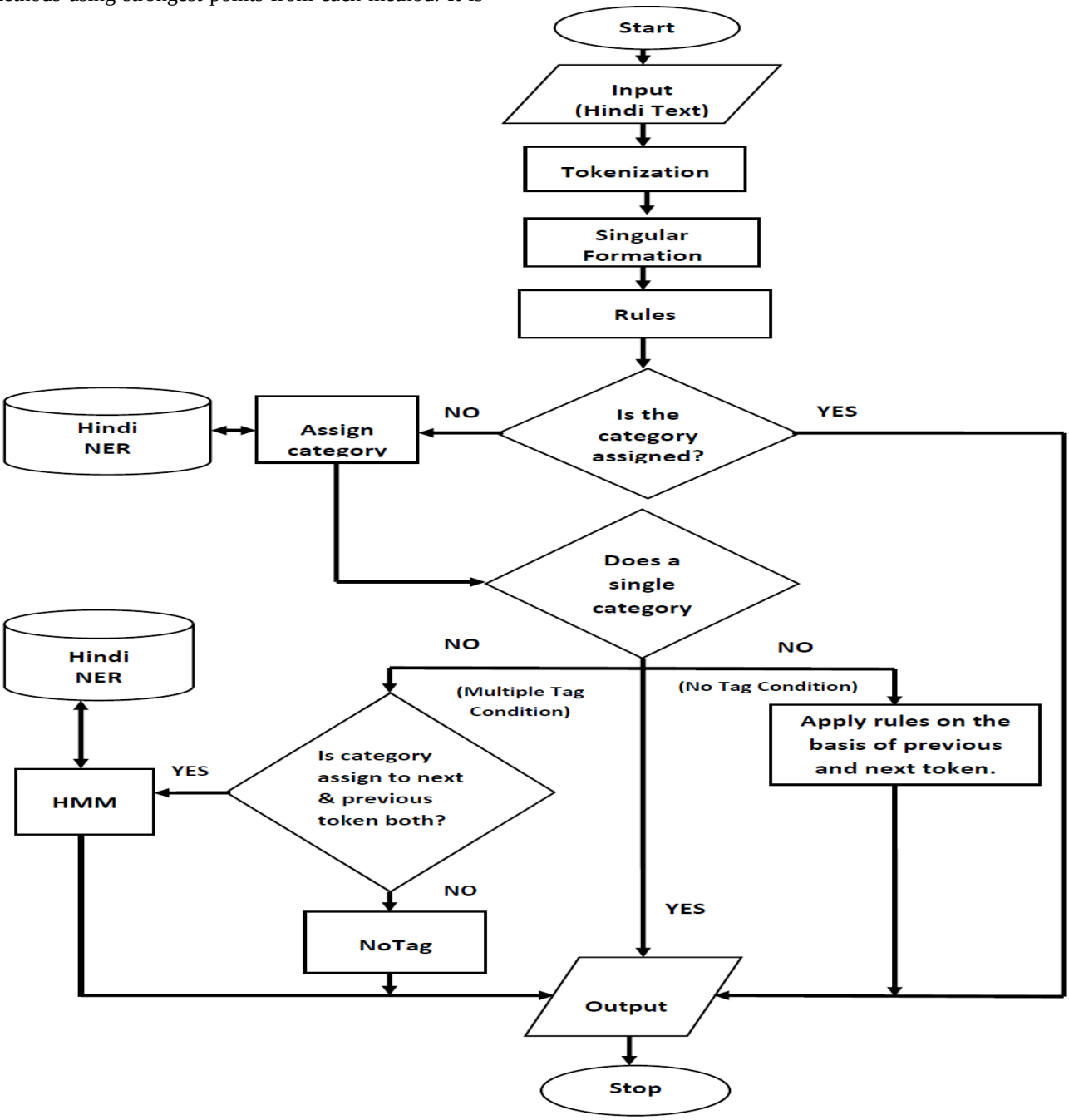

Figure 1:- POS Flow Chart

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
Volume 4, Issue 1, February 2014

A. Algorithm

Steps:

1. Enter the Hindi input text.
2. Tokenize the input text i.e. break the input sentence (according to phrases and single tokens) into independent and meaningful words.
3. If the token is not singular i.e. plural, then convert the token into singular form.
4. Find the POS Category of tokens by applying the POS Rules.
    A. If the category is found then go to step 8.
    B. Otherwise go to step 5.
5. Assign the POS category of token with the help of Hindi NER Database (using Exception Table and Corpus Table).
    A. If the single POS Category of token is found then go to step 8.
    B. If the no POS category is found then go to step 6.
    C. If the multiple POS category of token is found then go to step 7.
6. Assign the POS category to token by applying the rules on the basis of previous and next token and go to step 8.
7. Does a single category assign to previous and next token
    A. If yes, then assign the category by applying the HMM with the help of Hindi NER Database (using category_frequency table, tag_frequency table, word_frequencytable) and go to step 8.
    B. If no, then assign the No Tag Category to token i.e. POS category of token cannot be found and go to step 8
8. Display the result on the screen.

B. Following Rules are applied to identify different Tags

Rule 1: If current token is post position then there is high probability that previous token will be noun.

e.g. उसने पानी में पत्थर फेंका।

Rule 2: If token is adjective then there is high probability that next token will be noun.

e.g. राम को कच्चा आम पसन्द है|

Rule 3: If word ends with तर (tar), तम (tam), इक (ik) etc. postfix then token is tagged as adjective.

For Example: - लघुतर, विशालतम, प्रामाणिक

Rule 4: If current token is not tagged and next token tagged as an auxiliary verb, then there is high probability that currentn token will be main verb.

For Example: - वह खेल रहा है।

III. Evaluation and Result

Sentence	Hybrid POS Tagger
आइए जानतेहैं दिल्ली में हुए इस उलटफेर की कुछ वजहें	आइए - Verb / जानतेहैं - Verb / दिल्ली - Noun / में - Postposition / हुए - Verb/ इस - Pronoun / उलटफेर - Noun / की - Postposition / कुछ - Adjective / वजहें – Noun
केंद्रसरकार के खिलाफ आमलोगों का गुस्सा शीला सरकार को झेलना पडा	केंद्रसरकार - Noun / के - Postposition / खिलाफ - Prep/Conj / आमलोगों - Noun / का - Postposition / गुस्सा - Adjective / शीला - Noun / सरकार - Noun / को - Postposition / झेलना - Verb / पडा-Verb
दोपहर आते-आते ढोल नगाड़े यहां पहुंच गए और समर्थकनाच गाकर जश्नमनाने लगे	दोपहर- Noun / आते-आते - Verb / ढोल - Adjective / नगाड़े - Noun / यहां- Adverb / पहुंच - Noun / गए - Verb / और - Conjunction / समर्थकनाच - Verb / गाकर - Noun / जश्न - Noun/ मनाने - Noun / लगे - Verb /

Table 1:-Test Cases

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

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
Volume 4, Issue 1, February 2014

A. Precision

Precision is the ratio of the number of items of a certain named entity type correctly identified to all items that were assigned that particular type by the system.

$$P = \frac{\text{Number of correct tags assigned}}{\text{Total number of tags assigned}}$$

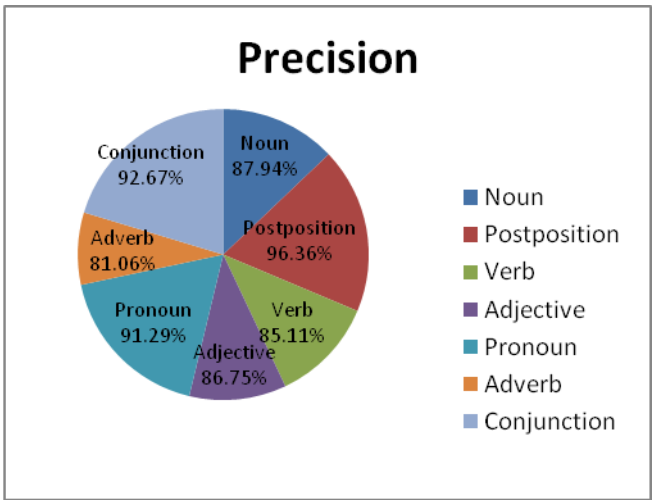

Figure 2: Precision Pie Chart

B. Recall

Recall measures the number of items of a certain named entity type correctly identified, divided by the total number of items of this type. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

$$R = \frac{\text{Number of correct tags assigned}}{\text{Total number of tags in the annotated test corpus}}$$

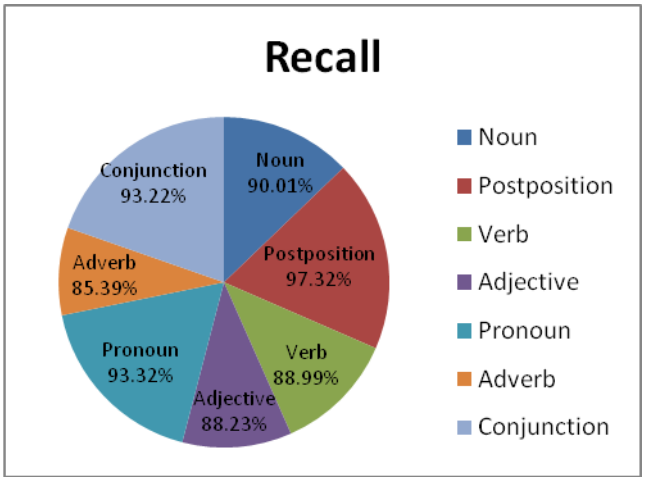

Figure 3: Recall Pie Chart

C. F-measure

The F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision *P* and the recall *R* of the test to compute the score. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0. It combines Recall (R) and Precision (P) using the formula .The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

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

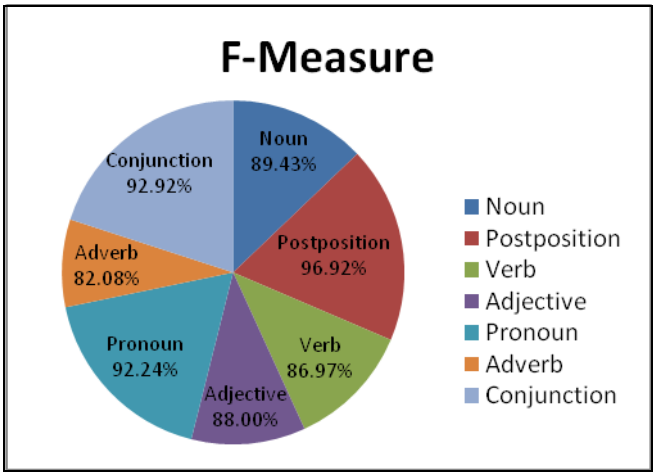

Figure 4: F-Measure Pie Chart

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
Volume 4, Issue 1, February 2014

IV. CONCLUSION

At last we conclude that Part of Speech tagging is the most important activity of any Natural Language based applications. The accuracy of any NLP tool is dependent on the accuracy of POS tagger. Different approaches have been used by authors for the development of part of speech tagger for Indian Languages.

We have presented a part-of-speech tagger for Hindi which uses hybrid framework. We have shown that such a system has good performance with an average accuracy of 89.9% for POS tagging. We believe that further error analysis and more language specific features would improve the system performance.

REFERENCES

[1] Uma Parameshwari Rao G, Parameshwari K: CALTS, University of Hyderabad, 'On the description of morphological data for morphological analyzers and generators: A case of Telugu, Tamil and Kannada'.

[2] Beesley, K. and L. Karttunen. 'Finite State Morphology'. Stanford, CA: CSLI Publications, 2003.

[3] Aduriz1.,Agirre E., 'A word-grammar based morphological analyzer for agglutinative languages', University of lheBasqtloCotlnlry, Basque Country.

[4] Koskenniemi .K , Two –Level Morphology: A general Computational; Model for Word Recognition and Production, , University of Helsinki, Helsinki,1983.

[5] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", In Proceedings of the IEEE, 77 (2), p. 257-286February 1989.Available at: http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf

[6] Language: A Hybrid Approach" International Journal of Computational Linguistics (IJCL), Volume (2) : Issue (1) : 2011.Available at: http://cscjournals.org/csc/manuscript/Journals/IJCL/volume2/Issue1/IJCL-19.pdf

[7] Sanjeev Kumar Sharma and Gurpreet Singh Lehal. (2011). *Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger*, Computer Science and Automation Engineering(CSAE), 2011 IEEE International Conference on June 2011, pp. 697-701.

[8] Dinesh Kumar and Gurpreet Singh Josan. (2010). *Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey*, International Journal of ComputerApplications (0975 – 8887) Volume 6– No.5, September 2010, pp.1-9.

[9] Nidhi Mishra and Amit Mishra. (2011). *Part of Speech Tagging for Hindi Corpus*, In the proceedings of 2011 International Conference on Communication systems and Network Technologies, pp.554-558.

[10] Andrew Borthwick. 1999. "Maximum Entropy Approach to Named Entity Recognition" Ph.D. thesis, New York University.

[11] F. Jelinek. 1997. Statistical Methods for Speech Recognition. MIT Press.

[12] NavneetGarg, Vishal Goyal, SumanPreet. "Rules Based Part of Speech Tagger" in the proceedings of COLING 2012:, Mumbai, December 2012.

Ms. Kanak Mohnot is a Research Scholar and doing her internship from C-DAC Pune under M.Tech, Information Technology IInd year Curriculum pursued by Banasthali University, Rajasthan, India. She has completed her B.Tech degree in Electronics & Communication from Technical University of Rajasthan.

Ms. Neha Bansal is a Research Scholar and doing her internship from C-DAC Pune under M.Tech, Computer Science IInd year Curriculum pursued by Banasthali University, Rajasthan, India. She has completed her B.Tech degree in Electronics & Communication from Technical University of Rajasthan.

Mr. Shashi Pal Singh is working as STO, AAI Group, C-DAC, Pune. He has completed his B.Tech and M.Tech in Computer Science & Engg. and has published various national & international papers. He is specialised in Natural Language Processing (NLP), Machine assisted Translation (MT), Cloud Computing and Mobile Computing.

***Mr.* Ajai Kumar** is working as Associate Director and Head, AAI Group, C-DAC, Pune. He is handling various projects in the area of Natural Language Processing, Information Extraction and Retrieval, Intelligent Language Teaching/Tutoring, Speech Technology [Synthesis & Recognition ASR], Mobile Computing, Decision Support Systems & Simulations and has published various national & international papers.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
Volume 4, Issue 1, February 2014