

POS tagger model for Hindi Language Using Novel Rule Based Technique

Priyanka Lohe¹, Vikas Pandey²

¹Mtech Scholar

²Associate Professor

^{1,2} Dept of Information Technology

^{1,2}Bhilai Institute of Technology Durg

Date of Submission: 20-11-2020

Date of Acceptance: 06-12-2020

ABSTRACT: Natural Language Processing (NLP) is one of the interesting topics in the research field of Computer Science. Hindi is most precisely language spoke in India. POS tagging is a procedure in which we tag each word in a sentence present in a tagset. This paper presents Part of Speech Tagging for Hindi Language by using Rule Base Approach for proper tagging of words and CRF++ for training and testing the file and to calculate accuracy. The total dataset used for this implementation is 1530 words. The corpus is taken from various news, essays and stories. The system achieves an accuracy of 85.78%.

Keywords: Natural Language Processing, Part of Speech, Rule base Approach, CRF++.

I. INTRODUCTION

Corpus is "an assortment of language in a colossal and sorted out plan of data and used as a key for trademark language preparing". Corpus is generally in the modes of written text, printed text or sample of spoken words or combination of all. In annotation process, input, as well as output, is natural languages like English, Hindi, etc. There are different levels of corpus annotation like Morphological analysis, POS tagging, Chunk tagging, etc. POS tagging is a basic step for language processing and can work as the first phase in other language processing tasks. The work on Part-of-Speech (POS) tagging for natural language tagging has begun in the early 1960s[1].

Most of the regional languages are low resources language. Some Indian languages are called low resource language as grammatical rules and literary work related to these languages is not present in public domain. Pre-processing task like POS tagging is a difficult errand for these languages. In POS tagging process a specific grammar class which is called as tag is assigned to a word in the sentence from tag set. Tag set is an

assortment of language structure class which comprise of English shortenings like N(Noun), VM(Verb), PP(Preposition) and so forth.[2]. Parts of Speech (POS) tagging is a process of identifying the suitable class of tag for a word from a given tag set.

Tagging of text is a difficult task the same number of times we get words which have diverse label classes as they are utilized in various setting. This phenomenon is termed as lexical ambiguity[13]. The same word 'कुल' is given a different label in the two sentences. In the first case it is termed as a pronoun as it is referring to an object (All). In the second case it is termed as a noun as it is referring to an ancestry. This issue can be settled by taking a gander at the word/label mixes of the encompassing words concerning the uncertain word (the word which has various labels).

Morphology is the field of the linguistics that studies the internal structure of the words. Morphological Analysis and generation are basic strides in any NLP Application. Morphological investigation implies accepting a word as information and recognizing their stems and appends. Morphological Analysis is essential for Hindi it has a rich system of inflectional morphology as like other languages[6]. Morphological Analyzer and generator is a tool for analyzing the given word furthermore, generator for producing word given the stem and its highlights (like appends). Grammatical form tagger is a significant use of characteristic language handling. It is an important part of morphological analyzer. Grammatical form labeling is the way toward allocating a grammatical form like thing, action word, relational word, pronoun, intensifier, descriptor or other lexical class marker to each word in a sentence[7]. There are different techniques for POS Tagging[14]:

1. **Lexical Based Methods** — Assigns the POS tag the most frequently occurring with a word in the training corpus.
 2. **Rule-Based Methods** — Assigns POS tags based on rules. For instance, we can have a standard that says, words finishing with "ed" or "ing" must be allocated to a verb. Rule-Based Techniques can be utilized alongside Lexical Based ways to deal with permit POS Tagging of words that are absent in the training corpus yet are there in the testing information.
 3. **Probabilistic Methods** — This technique does out the POS tags dependent on the likelihood of a specific tag succession happening. Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) are probabilistic ways to deal with appoint a POS Tag.
 4. **Deep Learning Methods** — In POS Tagging Recurrent Neural Network is used.
- A CRF is a **Discriminative Probabilistic Classifiers**. In CRFs, the info is a set of highlights (genuine numbers) got from the information arrangement utilizing highlight works, the loads related with the highlights (that are found out) and the previous label and the assignment is to anticipate the current label. The weights of different feature functions will be determined such that the likelihood of the labels in the training data will be maximised.

Example 1: कम हानिकारक बनाने के लिए कहा

WORDS	कम	हानिकारक	बनाने	के	लिए	कहा
TAGS	INTF	JJ	VNN	PREP	PREP	VFM

II. LITERATURE SURVEY

Researches were still going on in POS Tagger using Hindi Language. Different approaches were being used like mostly Rule Based Approach one of the oldest technique, Stochastic approach and transformation based learning are used with some modifications and have been tried and implemented.

Ekbal, Haque and Bandyopadhyay has implemented a “POS Tagger using Bengali Language in Conditional Random field”. They had used variety of features along with different contextual information of the words. The tagger has been trained and tested with the 72,341 and 20K wordforms. The system has achieves an accuracy of 90.3%[3].

Garg, Goyal and Preet presented a “Rule Based Part of Speech Tagger for Hindi”. The system is evaluated on the different domains of Hindi Corpus .Also Trained and Tested over a corpus of 26,149 words with 30 different POS tags. The system is achieved an accuracy of 87.55%[4].

Vijeta and Mantosh presented a “Part-of-Speech Tagging of Hindi Language Using Hybrid Approach”. They build a Hybrid Approach using Hidden Markov Model and Rule Based Tagging on Hindi language. The system uses a corpus of 13,000 words. Achieved an accuracy of 96.01% of average precision and 89.32% of average accuracy[5].

Joshi, Darbari and Mathur presented a “HMM Based POS Tagger for Hindi”. For the development of this tagger they used IL POS tag. They developed a test corpus of 500

sentences(11720 words). They disambiguated right word-label mixes utilizing the relevant data accessible in the content. They achieved the accuracy of 92%.[6].

Singh, Gupta, Shrivastava and Bhattacharya presented a “Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi”. They establish a methodology of POS tagging which the resource disadvantaged languages can make use of. They have done the evaluation on the news domain with 4-fold cross validation of the corpora. They train their data of size 15,562 words. The system achieved the accuracy of 93.45%[8].

Aniket Dalal et al., 2006 built up a framework utilizing Maximum Entropy Markov Model for Hindi. Framework required an element work catching the lexical and morphological component of language and list of capabilities was shown up after an inside and out examination of a clarified corpus. The system was evaluated over a corpus of 15562 words with 27 different POS tags and system achieved the accuracy of 94.81%[9].

Sanjeev Kumar Sharma et al., 2011 built up a framework utilizing Hidden Markov Model to improve the precision of Punjabi Part of Speech tagger. A module has been built up that takes yield of the current POS tagger as info and relegate the right tag to the words having more than one tag. The system was evaluated over a corpus of 26,479 words and system achieved the accuracy of 90.11% [10].

Pranjal Awasthi et al., 2006 built up a framework utilizing a mix of Hidden Markov Model and mistake driven learning. Labeling measure comprises of two phases, an underlying factual labeling utilizing the TnT tagger, which is a subsequent request Hidden Markov Model (HMM) and apply a lot of change rules to address the mistakes presented by the TnT tagger. The framework was created utilizing 26 distinct POS labels and precision of framework is 79.66% utilizing the TnT tagger and changes in post handling improves the exactness to 80.74% [11].

A POS tagger was developed using Hidden Markov Model for Assamese. Unknown words were tagged using simple morphological analysis. The system was evaluated over a corpus of 10,000 words with 172 different POS tags and system achieved the accuracy of 87% [12].

III. SYSTEM DESCRIPTION

For our System development we had collected a hindi corpus from different domains and tokenize the sentence using Tokenizer and tag the words using Sanchay Software and unknown words are tag 'Unk'

3.1 Algorithm For Tagging and Tokenizing Hindi Words

1. Collect Hindi Corpus from Different Domains
2. Tokenize the Sentence using Tokenizer.
3. After Tokenizing the sentence, tag the sentence using Sanchay Software.
4. For better results apply Rule Base Technique which can tag the results using appropriate rules.

5. After tagging now split the corpus file into training and testing data.
6. Apply CRF technique to check accuracy of data.

2.2 Tokenizing of Data

The sentence gets tokenized into words, punctuation marks and other symbols were categorized as tokens. Tokens are separated by white-space characters, line breaks or punctuation markers. Special tokens are handled separately to avoid wrong Tokenization In software engineering, lexical investigation, lexing or tokenization is the way toward changing over an arrangement of characters, (for example, in a PC program or site page) into a grouping of tokens (strings with a doled out and along these lines recognized significance).

3.3 Tagging of Data

We automatically tag our data using Sanchay Software. The tagging module doles out labels to tokens and furthermore look for questionable words and as per their sort relegate some extraordinary images to them. Tagging is the way toward increasing a word in a book (corpus) as comparing to a specific grammatical feature, in light of the two its definition and its unique circumstance—i.e., its relationship with adjoining and related words in an expression, sentence, or section.

Sanchay is a collection of tools and APIs for Natural Language Processing (NLP) and Computational Linguistics (CL), specially tailored for South Asian languages.

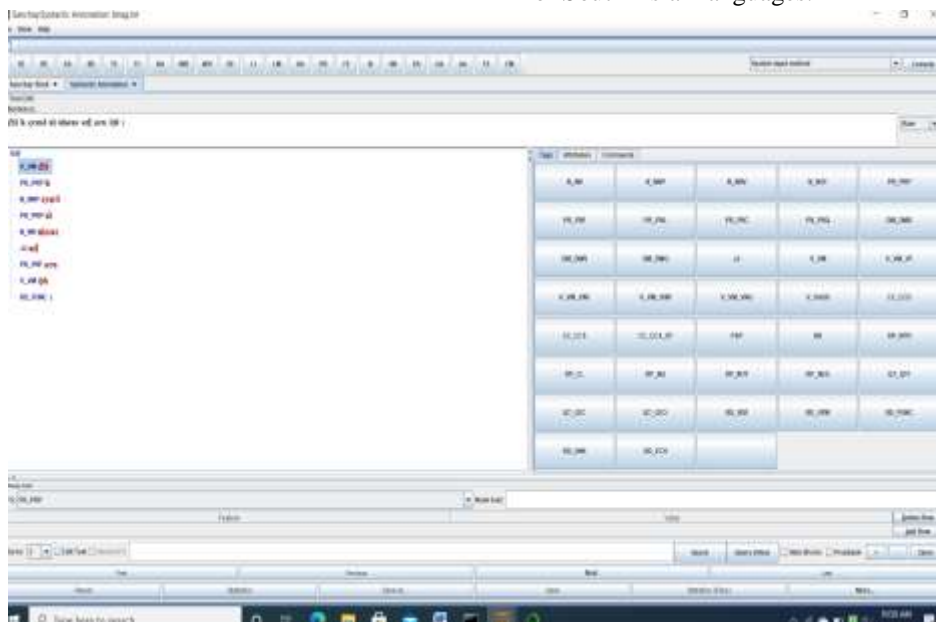


Figure 1. Tagging of Data

Now we are getting our Tagged data after tagging words with appropriate Tags using Sanchay Software

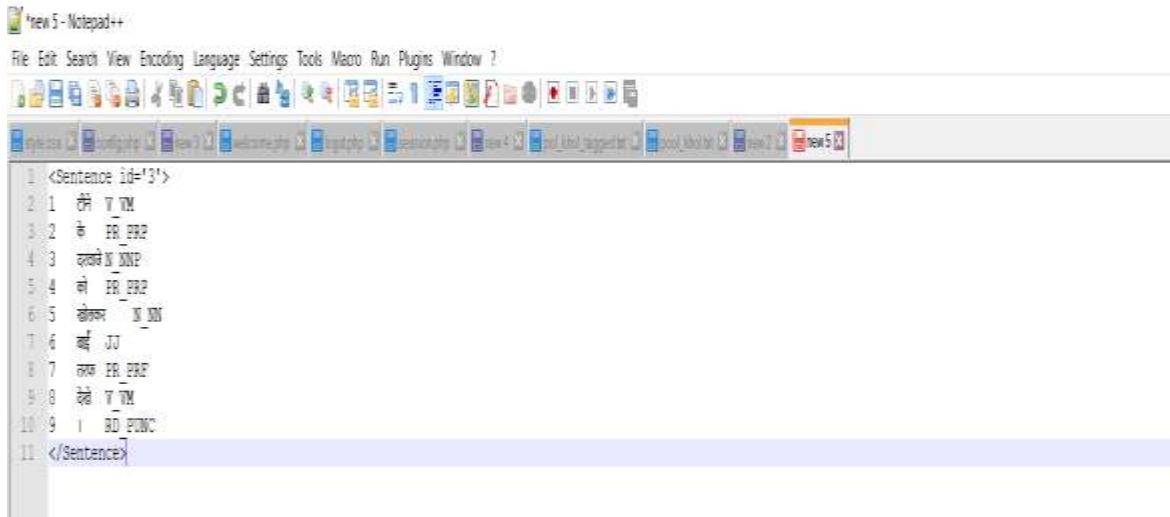


Figure 2. Tagged Data

Train and Test data using CRF++

Training

Use crf_learn command for training

crf_learn template train_crf.txt tagger_model

The learn template generates a trained model file.

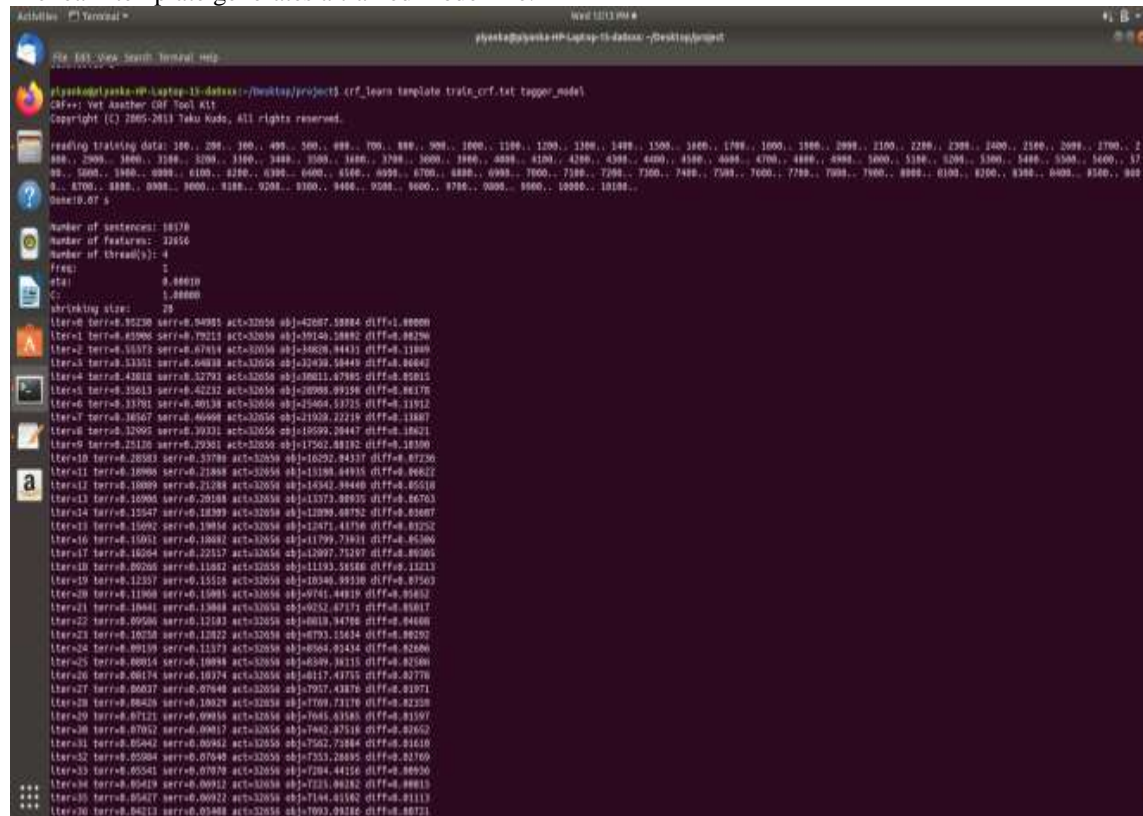


Figure 3. Training Data


- iter: number of iterations performed.
- terr: error rate with respect to tags. (# of error tags/# of all tag)
- serr: error rate with respect to sentences. (# of error sentences/# of all sentences)
- obj: current object value. At the point when this worth joins to a fixed point, CRF stops the emphasis.
- diff: relative distinction from the past item esteem.

Use crf_test command:

```
crf_test -m tagger_model test_crf.txt >
new_file.txt
```

In the testing, you don't have to indicate the layout document, in light of the fact that the model record has a similar data for the format. test_file is the test information you need to allocate consecutive tags. This document must be written in a similar configuration as preparing record.

Testing :



Index	Word	Actual Tag	Predicted Tag
1	अपने	IN	IN
2	सू	PREP	PREP
3	हैं	IN	IN
4	को	PREP	PREP
5	के	IN	IN
6	के	PREP	PREP
7	के	PREP	PREP
8	के	PREP	PREP
9	के	PREP	PREP
10	के	PREP	PREP
11	के	PREP	PREP
12	के	PREP	PREP
13	के	PREP	PREP
14	के	PREP	PREP
15	के	PREP	PREP
16	के	PREP	PREP
17	के	PREP	PREP
18	के	PREP	PREP
19	के	PREP	PREP
20	के	PREP	PREP
21	के	PREP	PREP
22	के	PREP	PREP
23	के	PREP	PREP
24	के	PREP	PREP
25	के	PREP	PREP
26	के	PREP	PREP
27	के	PREP	PREP
28	के	PREP	PREP
29	के	PREP	PREP
30	के	PREP	PREP
31	के	PREP	PREP
32	के	PREP	PREP
33	के	PREP	PREP
34	के	PREP	PREP
35	के	PREP	PREP
36	के	PREP	PREP
37	के	PREP	PREP
38	के	PREP	PREP
39	के	PREP	PREP
40	के	PREP	PREP
41	के	PREP	PREP
42	के	PREP	PREP
43	के	PREP	PREP
44	के	PREP	PREP
45	के	PREP	PREP
46	के	PREP	PREP
47	के	PREP	PREP
48	के	PREP	PREP
49	के	PREP	PREP
50	के	PREP	PREP

Figure 4. Testing Data

And finally we calculate the accuracy of our tagged data.

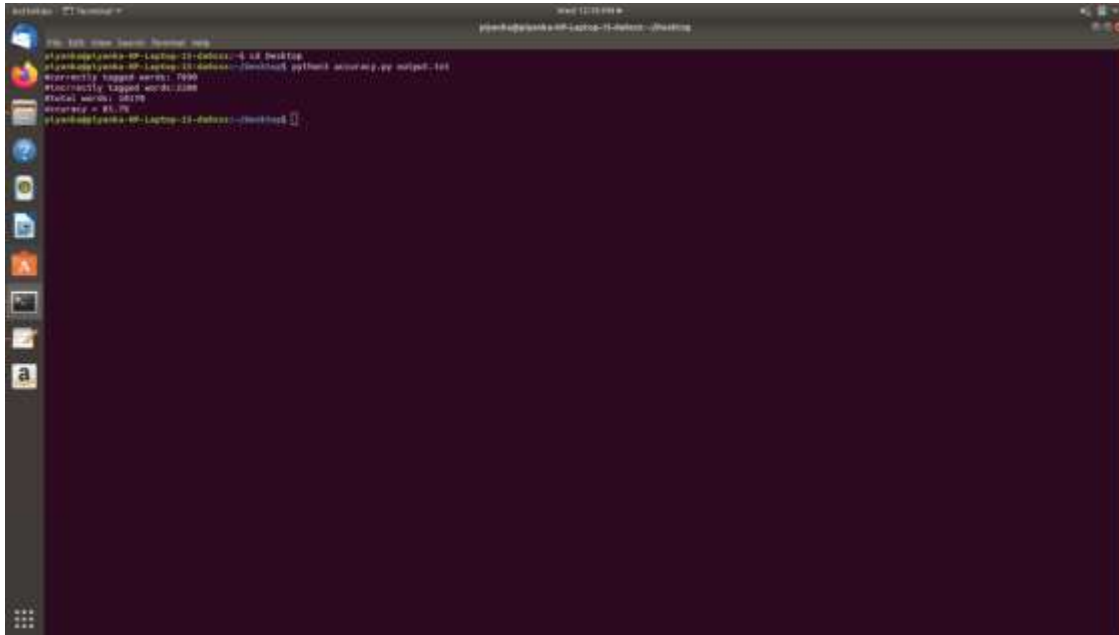


Figure 5. Accuracy of Data

3.4 Automatic Tagger

We have made a Automatic Hindi POS Tagger which tag our data present in our database.

Working of Automatic Tagger

1. Enter the input hindi text which are present in our database.
2. It break the sentence into meaningful words.

3. Find token singular or plural.
4. Find POS category for appropriate token.
5. If words are found in our database then system will tag it
6. Otherwise it will tag it 'UNK'.
7. At last we get our Tagged data.

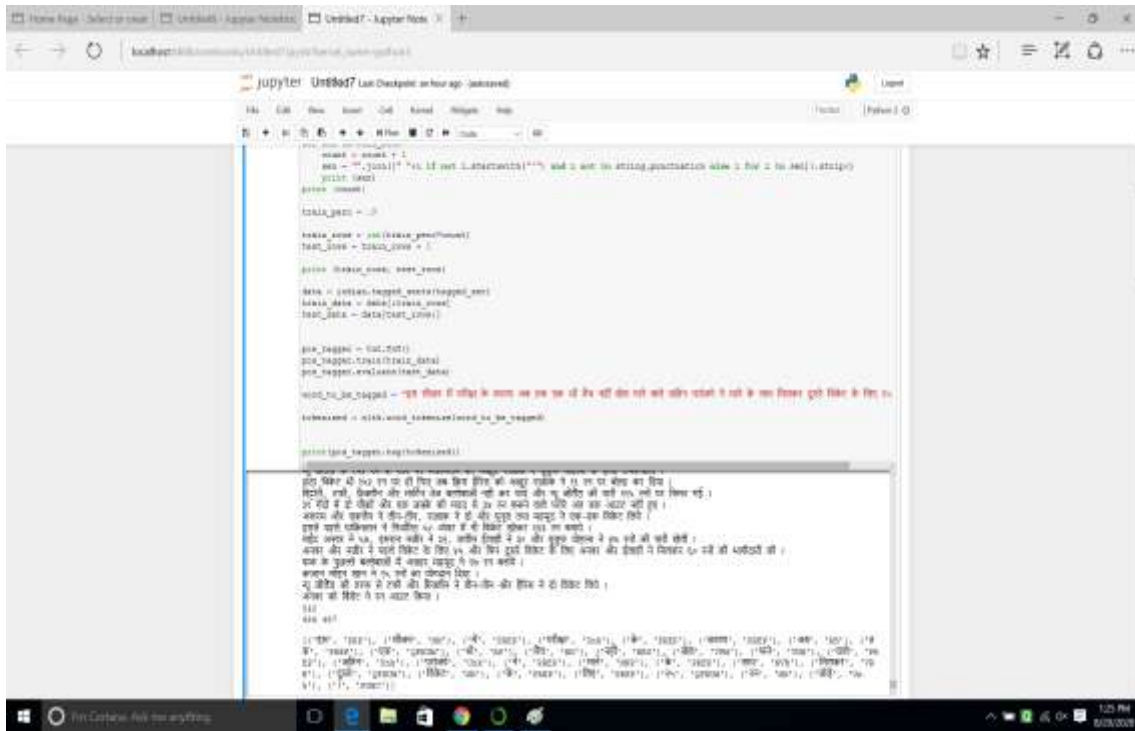


Figure 6. Automatic Tagger

IV. RESULT

Our system has validated through various data sets. For testing the performance of our system, we developed a test corpus of 1530 words. Our system achieves an accuracy of 85.79%. For standard performance of system we should calculate Precision, Accuracy.

$$\text{Precision} = \frac{\text{No. Of Correct POS Tags}}{\text{No. Of POS Tags}}$$
$$\text{Accuracy} = \frac{\text{No. Of Correct POS Tags}}{\text{No. Of POS Tags in Text}}$$

The system using Rule base Approach to increase Precision and Accuracy. This model also improves quality of tagging. The system yield 89.57% of average precision and 85.78% of average accuracy.

V. CONCLUSION

The system is developed with the help of Rule base and CRF base approach for accuracy. Corpus matching is applied while tagging known words. For obscure words labeling different Hindi sentence structure rules are applied. In future we would also increase our database size. We would also provide some additional functionality with POS tagging. In future work we hope to increase Precision and Accuracy of our system by increasing the size of tagged corpus.

REFERENCES

- [1]. Vijeta Khicha, Mantosh Manna.(2017). Part-of-Speech Tagging of Hindi Language Using Hybrid Approach. In Proc. Of International Journal of Engineering Technology Science and Research IJETSRSR, pp 737-741.
- [2]. Agrawal, R., Ambati, B., & Singh, A.Singh.(2012). A GUI to Detect and Correct Errors in Hindi Dependency Treebank. In Proc.of Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, 1907-1911.
- [3]. Ekbal, A., Haque, R, & Bandhopadha, S. (2007). Bengali part of speech tagging using Conditional Random Field. In Proc. of SPSAL2007. 131-136
- [4]. Garg,Goyal & Preet.(2012).Rule Based Hindi Part of Speech Tagger.In Proc. Of COLING 2012. 163-174.
- [5]. Vijeta & Mantosh.(2017).Part-of-Speech Tagging of Hindi Language using Hybrid Approach. In Proc. Of IJETSRSR 2017. 737-741.
- [6]. Beesley, K. and L. Karttunen. 'Finite State Morphology'. Stanford, CA: CSLI Publications, 2003.
- [7]. Dinesh Kumar and Gurpreet Singh Josan. (2010). Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey, International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010, pp.1-9
- [8]. Singh, Gupta, Shrivastava and Bhattacharya. Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi :Department of Computer Science and Engineering Indian Institute of Technology, Bombay Powai, Mumbai.
- [9]. Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke. (2006). Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach, In Proceeding of the NLP AI Machine Learning Competition, 2006.
- [10]. 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.
- [11]. Pranjal Awasthi, Delip Rao and Balaraman Ravindran. (2006). Part Of Speech Tagging and Chunking with HMM and CRF, In the proceedings of NLP AI Contest, 2006.
- [12]. Sharia, N., Das, D., Sharma U. , Kalita, J. (2009). Part of Speech Tagger for Assamese Text . In Proc. of the ACL-IJCNLP, 33-36.
- [13]. Joshi, Darbari and Mathur , 2013 HMM BASED POS TAGGER FOR HINDI In Proc. Of Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA pp 341–349.
- [14]. <https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields>