CONDITIONAL RANDOM FIELD BASED PART OF SPEECH TAGGER FOR CHHATTISGARHI LANGUAGE

[1] Vikas Pandey,[2] Dr. M.V Padmavati, [3]Dr. Ramesh Kumar
[1]Dept. of Information Technology , Bhilai Institute of Technology, Durg,India
[2,3] Dept. of Computer Science and Engg., Bhilai Institute of Technology, Durg,India
vikas.pandey@bitdurg.ac.in,

ABSTRACT

Part of Speech Tagger (POS) is an important tool that is used to develop machine translation (MT) system. Chhattisgarhi is a low resource language for which POS tagger, morphological analyser and parser has not been developed. For doing conversion from Chhattisgarhi to Hindi and Hindi to Chhattisgarhi we need Chhattisgarhi POS tagger. In this paper, we are presenting conditional random field based part of speech tagger for Chhattisgarhi language. The system is constructed over corpus size 69,731 words which act as training data set. A new tag set has designed for Chhattisgarhi language in consultation with the linguistic expert. With the help of this tag set all the untagged Chhattisgarhi words are tagged after tokenization. The tag set consist of 34 different part of speech tags. The corpus is taken from various domains including stories and news articles etc. The system achieves an accuracy of 85%.

Keywords: Corpus, Chhattisgarhi, Machine Translation, Part of Speech Tagger, Tag set .

1. INTRODUCTION

Part of Speech (POS) tagging is a process of identifying the suitable class of tag for a word from a given tag set. It is very important task of pre-processing activity in machine translation. Machine translation systems take a source language and convert it into target language. Various tools are required in machine translation systems like tokenizer , POS tagger, morphological analyser and parser. POS tagger comes under pre-processing phase of machine translation system. 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 challenging task for these languages. In POS tagging process a specific grammar class which is called as tag to a word in the sentence from tag set. Tag set is a collection of grammar class which consist of English abbreviations like NN(Noun), VM(Verb), PP(Preposition) etc.[9].

Example 1: हमन दुनो रेलगाड़ी म बंबई जाबों

WORDS	हमन	दुनो	रेलगाड़ी	म	बंबई	जाबों
TAGS	PRP	N	N	PP	N	VM

Table 1: Chhattisgarhi words and its tags taken from Chhattisgarhi tag set

There are various approaches for POS tagging: Rule based approach, Statistical approach and Hybrid approach [8,9]. Accuracy factor is the most important factor in deciding the performance of POS tagger [5].

The Rule Based POS tagging approach is based on grammar rules that are framed by observing the grammatical structure of any language. These rules can be written in form of production grammar rules. Example:

"A proper noun is always followed by a noun" as in the Table (a) हमन (Pronoun)is followed by दुनो(Noun)

There are some limitations of rule based approach; the main limitation is the formation of rule base. In this a rule is formulated for each condition [8,9].

The Statistical Based POS tagging approach is based on two important factors. These are : Frequency and probability of occurrence of any word .In this approach most frequently used tag for a specific word in the annotated training dataset is used to tag that word in the unannotated dataset .The limitation 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[8].

In Hybrid Approach the probability theory of statistical method is used to train the corpus and then the set of production rules are applied on the testing corpus for tagging of testing corpus[8,9]. POS Tagging process is broadly classified into two models: Supervised Model and Unsupervised Model [6]. Classification of POS Tagging is shown in Figure 1 .

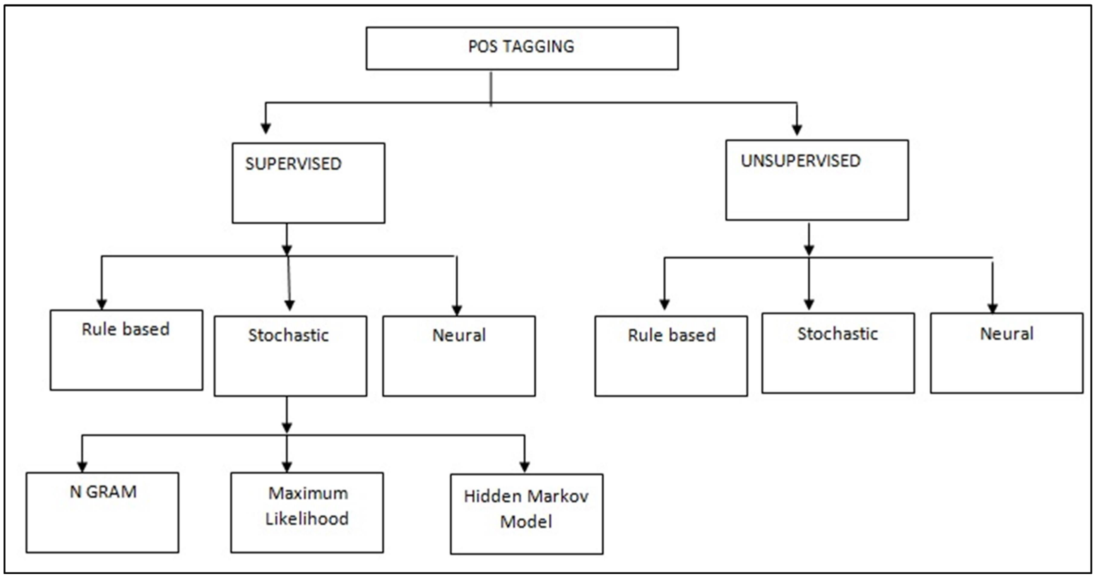


Figure.1. Classification of POS Tagging

Part of Speech (POS) tagging has many applications in natural language processing like semantic analysis, deep parsing, information retrieval etc..

2. LITERATURE SURVEY

Research has already been done in morphologically rich Indian languages like Hindi, Bengali, Telugu, Marathi, Tamil, Urdu, Gujarati, Kannada, Malayalam, Odia and Punjabi. There are some low resource languages in India like Awadhi, Magahi, Nemari, Bhojpuri, and Chhattisgarhi for which machine translation tools have not been developed yet.

A POS tagger was developed using conditional random field for Bengali language In this system contextual information of the words has been used to search different POS tags for various tokenized words. The system was evaluated over a corpus of 72,341 words with 26 different POS tags and system achieved the accuracy of 90.3% [1].

A POS tagger was developed using Hidden Markov Model for Hindi ,which uses a Naïve stemmer as a pre-processor based on longest suffix matching algorithm to achieve accuracy of 93.12% [2].

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% [3].

A POS tagger was developed using Hidden Markov Model for Hindi .They uses Indian language POS tag set to develop this tagger and achieved the accuracy of 92% [4]

3. METHODOLOGY

Conditional Random Field approach (CRF) has been adapted for the training of dataset which comes under supervised model of POS tagging and is based on theory of probability. In this method two data sets are used: training set and testing data. Pre-tagged models to automatically tag the testing data. The performance of tagger is based on the accuracy of tagging. As the size of corpus increases the accuracy of POS tagger also increases.

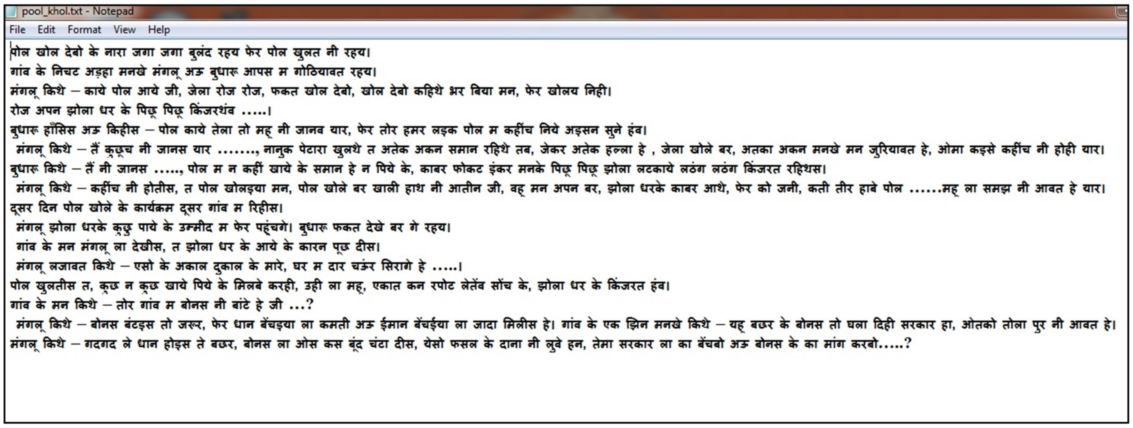
In Conditional Random Field approach CRF++ tool has been used which is a open source tool available for implementation of Conditional Random Fields (CRFs) for segmenting and labelling sequential data. This tool is based on the idea that the tag which is best for a given word is determined by the probability which occurs with the n-1 previous tags. The drawback of this tool is that sometimes it retrieves a correct tag for a given word but along with this it can also sometimes retrieve invalid sequences of tags.

CRF++ tool kit is a general purpose tool which is applied to a variety of NLP tasks, such as Named Entity Recognition, Information Extraction and Text Chunking.

Steps for tagging using CRF Method

Chhattisgarhi POS tagging is done using following steps:

1.For corpus, several Chhattisgarhi articles are considered. The contents of the articles needs to be converted into Unicode before processing. Figure.2 contains the Chhattisgarhi story called “Polkhol”.



```
पोल खोल देबो के नारा जगा जगा बुलंद रहय फेर पोल खुलत नी रहय।
गांव के निचट अड़हा मनखे मंगलू अऊ बुधारू आपस म गोठियावत रहय।
मंगलू किथे – काये पोल आये जी, जेला रोज रोज, फकत खोल देबो, खोल देबो कहिथे भर बिया मन, फेर खोलय निही।
रोज अपन झोला धर के पिछू पिछू किंजरथंव .....।
बुधारू हाँसिस अऊ किहीस – पोल काये तेला तो महू नी जानव यार, फेर तोर हमर लइक पोल म कहींच निये अइसन सुने हंव।
 मंगलू किथे – तैं कुछूच नी जानस यार ......., नानुक पेटारा खुलथे त अतेक अकन समान रहिथे तब, जेकर अतेक हल्ला हे , जेला खोले बर, अतका अकन मनखे मन जुरियावत हे, ओमा कइसे कहींच नी होही यार।
बुधारू किथे – तैं नी जानस ....., पोल म न कहीं खाये के समान हे न पिये के, काबर फोकट इंकर मनके पिछू पिछू झोला लटकाये लठंग लठंग किंजरत रहिथस।
 मंगलू किथे – कहींच नी होतीस, त पोल खोलइया मन, पोल खोले बर खाली हाथ नी आतीन जी, वहू मन अपन बर, झोला धरके काबर आथे, फेर को जनी, कती तीर हाबे पोल ......महू ला समझ नी आवत हे यार।
दूसर दिन पोल खोले के कार्यक्रम दूसर गांव म रिहीस।
 मंगलू झोला धरके कुछु पाये के उम्मीद म फेर पहूंचगे। बुधारू फकत देखे बर गे रहय।
 गांव के मन मंगलू ला देखीस, त झोला धर के आये के कारन पूछ दीस।
 मंगलू लजावत किथे – एसो के अकाल दुकाल के मारे, घर म दार चऊंर सिरागे हे .....।
पोल खुलतीस त, कुछ न कुछ खाये पिये के मिलबे करही, उही ला महू, एकात कन रपोट लेतेंव सोंच के, झोला धर के किंजरत हंव।
गांव के मन किथे – तोर गांव म बोनस नी बांटे हे जी ...?
 मंगलू किथे – बोनस बंटइस तो जरूर, फेर धान बेंचइया ला कमती अऊ ईमान बेंचईया ला जादा मिलीस हे। गांव के एक झिन मनखे किथे – यहू बछर के बोनस तो घला दिही सरकार हा, ओतको तोला पुर नी आवत हे।
मंगलू किथे – गदगद ले धान होइस ते बछर, बोनस ला ओस कस बूंद चंटा दीस, येसो फसल के दाना नी लुवे हन, तेमा सरकार ला का बेंचबो अऊ बोनस के का मांग करबो.....?
```

Figure.2. Sample Chhattisgarhi Corpus

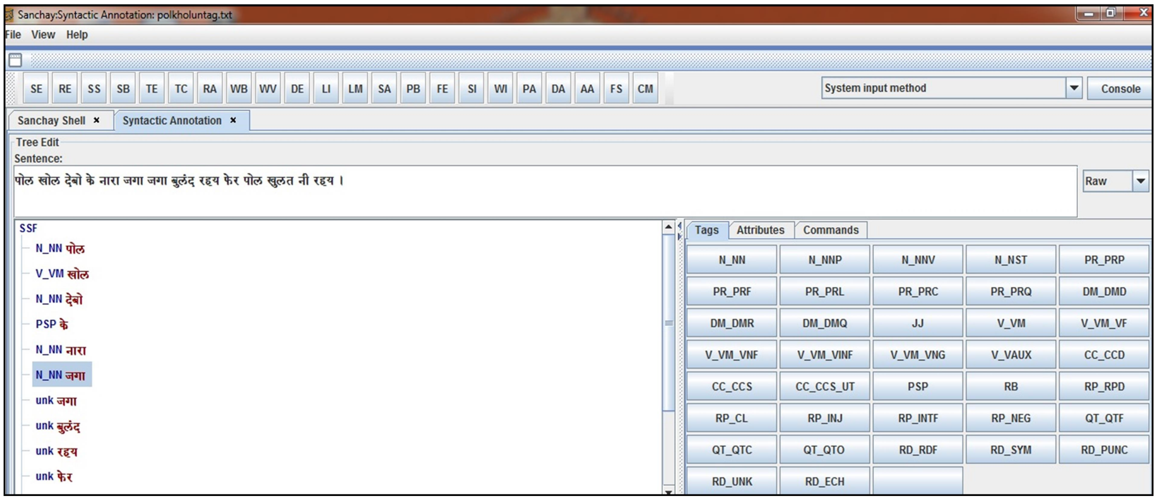
2. When a user gives the input path of the corpus , the Chhattisgarhi sentences are tokenised based on the technique of finding delimiter. In case of Chhattisgarhi sentences, Purnviram (‘|’) is the delimiter. The final untagged words of split sentences are stored in a separate file. The pseudo code for splitting and tagging id shown below:

```
Input Chhattisgarhi text (Unicode)
Read Chhattisgarhi Text T source
T Temp=T source text
Reg X Parse sentence parse [ ] =Split sentence ( )
For i = 0 to Regular sentence Parse length-l
a (i) = Reg Sentence Parse (i)
Sentence = at length
Words [ ] =split words ( )
Open file F .txt
Ib I [ ]=a I  [ ]
for i = 0 to Ibl-l
Sentence id = i+ 1
for j = 0 to word count -1
word id = (j+l)
word [ ]= words [j]
next
write word [ ] to F.txt
display F.txt
```

S is the user input the source text and i is the length of the sentence

3.By the help of Sanchay tool, tagging of each token is done in consultation with annotators.In order to tag the word Sanchay tool is used which is a open source platform made by Language Technologies Research Centre (LTRC) Hyderabad, for working on Indian languages, using computers and also for developing Natural Language Processing (NLP) based applications. It is used in syntactic annotation interface (used for Hindi dependency annotation), it has several other useful functionalities as well. Font conversion, language and encoding detection, n-gram generation are a few of them [7].

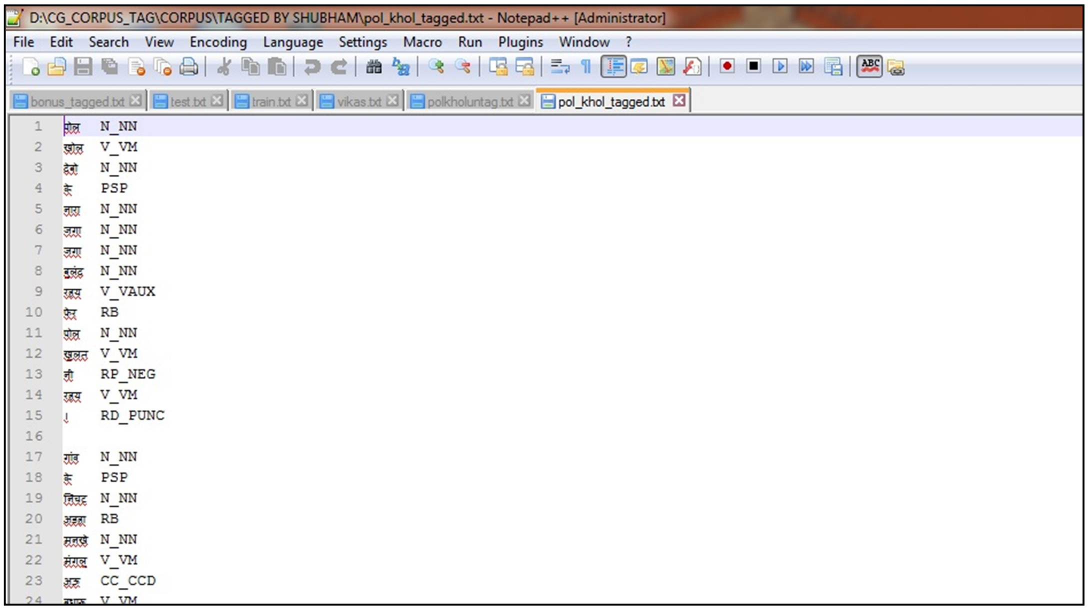
The screen shots of out after splitting and tokenizing is shown in Figures 2(a) and 2(b).


Sanchay:Syntactic Annotation: polkholuntag.txt
File View Help
SE RE SS SB TE TC RA WB WV DE LI LM SA PB FE SI WI PA DA AA FS CM
System input method
Console
Sanchay Shell
Syntactic Annotation
Tree Edit
Raw
Tags Attributes Commands
N_NN N_NNP N_NNV N_NST PR_PRP
PR_PRF PR_PRL PR_PRC PR_PRQ DM_DMD
DM_DMR DM_DMQ JJ V_VM V_VM_VF
V_VM_VNF V_VM_VINF V_VM_VNG V_VAUX CC_CCD
CC_CCS CC_CCS_UT PSP RB RP_RPD
RP_CL RP_INJ RP_INTF RP_NEG QT_QTF
QT_QTC QT_QTO RD_RDF RD_SYM RD_PUNC
RD_UNK RD_ECH

Sentence:
पोल खोल देबो के नारा जगा जगा बुलंद रहय फेर पोल खुलत नी रहय ।

```
SSF
N_NN पोल
V_VM खोल
N_NN देबो
PSP के
N_NN नारा
N_NN जगा
unk जगा
unk बुलंद
unk रहय
unk फेर
```

Figure.2(a). Tagging of each words in Sanchay


D:\CG_CORPUS_TAG\CORPUS\TAGGED BY SHUBHAM\pol_khol_tagged.txt - Notepad++ [Administrator]
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
bonus_tagged.txt
test.txt
train.txt
vikas.txt
polkholuntag.txt
pol_khol_tagged.txt

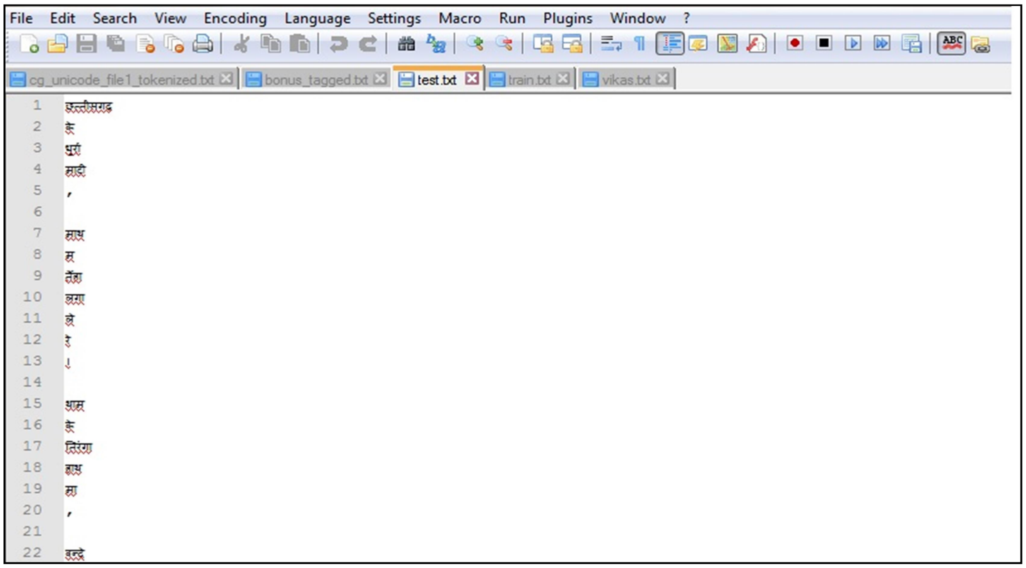
```
पोल   N_NN
खोल   V_VM
देबो   N_NN
के    PSP
नारा   N_NN
जगा   N_NN
जगा   N_NN
बुलंद  N_NN
रहय   V_VAUX
फेर   RB
पोल   N_NN
खुलत  V_VM
नी    RP_NEG
रहय   V_VM
।     RD_PUNC

गांव   N_NN
के    PSP
निपट  N_NN
अब्बड़  RB
मनखे  N_NN
मंगल  V_VM
अऊ   CC_CCD
      V_VM
```

Figure.2(b). Complete Tagged Training file

4. After the complete formation of training dataset in text file, this file is inserted in CRF++ tool kit

5. A test data file is taken which is again tokenized by the help of line splitter program , consist of untagged words which has to be trained by the help of CRF++ tool kit .


File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
cg_unicode_file1_tokenized.txt
bonus_tagged.txt
test.txt
train.txt
vikas.txt

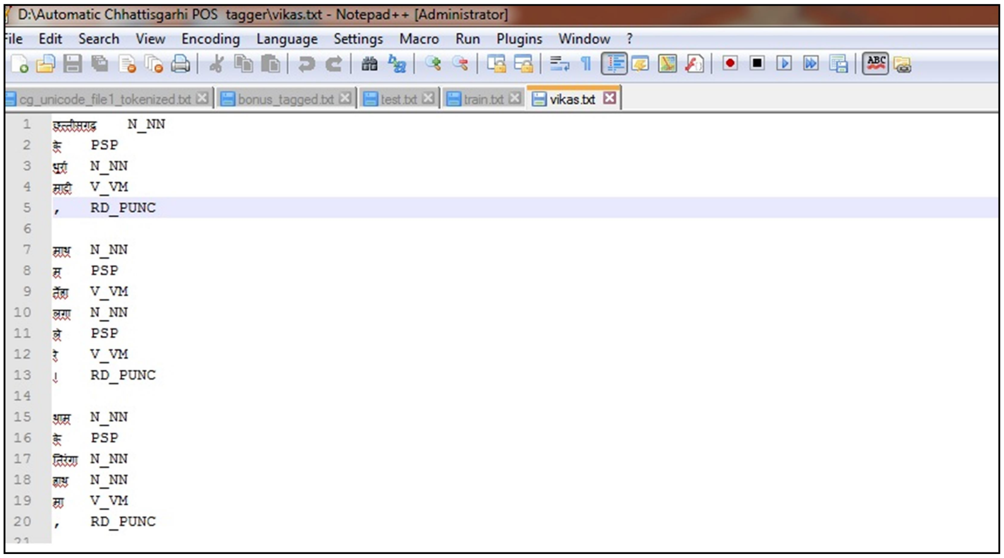
```
छत्तीसगढ़
के
धुर्रा
माटी
,

माथ
म
तैंहा
लगा
ले
रे
।

थाम
के
तिरंगा
हाथ
मा
,

वन्दे
```

Figure.3(a). Test data file without tagging


D:\Automatic Chhattisgarhi POS tagger\vikas.txt - Notepad++ [Administrator]
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
cg_unicode_file1_tokenized.txt
bonus_tagged.txt
test.txt
train.txt
vikas.txt

```
छत्तीसगढ़	N_NN
के	PSP
धुर्रा	N_NN
माटी	V_VM
,	RD_PUNC

माथ	N_NN
म	PSP
तैंहा	V_VM
लगा	N_NN
ले	PSP
रे	V_VM
।	RD_PUNC

थाम	N_NN
के	PSP
तिरंगा	N_NN
हाथ	N_NN
मा	V_VM
,	RD_PUNC
```

Figure.3(b). Test data file after training

4. RESULTS AND DISCUSSIONS

Input Chhattisgarhi Sentence

छत्तीसगढ़ के धुर्रा माटी, माथ म तैंहा लगा ले रे । थाम के तिरंगा हाथ मा, वन्दे-मातरम् गा ले रे।

Output Chhattisgarhi Sentence

Sno.	Chhattisgarhi Words	Tagging
1	छत्तीसगढ़ के धुर्रा माटी, माथ	<N_NN><PSP><N_NN><V_VM><RD_PUNC><N_NN>

2	म तैंहा लगा ले रे।	<PSP><V_VM><N_NN><PSP> <V_VM><RD_PUNC>
3	थाम के तिरंगा हाथ मा,	<N_NN><PSP><N_NN><N_NN>< V_VM> <RD_PUNC>
4	वन्दे-मातरम् गा ले रे।	<N_NN> <CC_CCD><N_NN><N_NN> <PSP><V_VM><RD_PUNC>

Table (b): Chhattisgarhi words and corresponding tags

5. CONCLUSION AND SCOPE OF FURTHER WORK

In Conditional Random Field technique POS tagging can be done where test dataset can be easily trained. A test data set of 106 untagged words gets trained by trained data set .In some cases exact categorization noun tag becomes difficult and differentiation between main verb and auxiliary verb is also an important issue. These problems can be resolved by adapting a hybrid approach in which rule based approach can be clubbed with neural network based approach which will enhance the performance of the tagger. In hybrid approach first training of test data set can be done after which we can formulate certain rule which will be helpful in resolving these issues.

REFERENCES

1. Ekbal, A., Haque, R, & Bandhopadha, S. (2007). Bengali part of speech tagging using Conditional Random Field. In Proc. of SPSAL2007. 131-136.
2. Shrivastav, M., & Bhattacharyya, P. (2008). Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge. In Proc.of ICON, 1-6
3. Sharia, N., Das, D., Sharma U. , Kalita, J. (2009). Part of Speech Tagger for Assamese Text . In Proc. of the ACL-IJCNLP, 33-36.
4. Joshi, N., Darbari, H., & Mathur, I. (2013). HMM Based POS Tagger For Hindi. In Proc.of CCSIT, SIPP, AISC, PDCTA, 341-349
5. Kumar, D., Josan, G., S. (2010). Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. International Journal of Computer Applications, Vol.6, 1-9.
6. Antony, P.,J.(2011). Parts Of Speech Tagging for Indian Languages: A Literature Survey. International Journal of Computer Applications, Vol 34, 22-29.

7. Agrawal, R. , Ambati, B., & Singh, A.Singh.(2012). A GUI to Detect and Correct Errors in Hindi DependencTreeban. In Proc.of Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, 1907-1911

8. Hasan, M., F.,Uzzaman, N., & Khan, M.(2006). Comparison of Different POS Tagging Techniques (n- grams, HMM and Brill's Tagger) for Bangla. International Conference on Systems, Computing Sciences and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering.

9. Kumawat, D., & Jain,V. (2015). POS Tagging Approaches: A Comparison. International Journal of Computer Applications ,vol 118,32-38