Natural Language Chhattisgarhi: A Literature Survey

Rijuka Pathak[1], Somesh Dewangan[2]

[#1] M.Tech, Scholar, Department CSE, DIMAT,India

[#2] Reader, Department CSE, DIMAT,India

***Abstract*— Chhattishgarhi is a official language in the Indian state of Chhattisgarh. Spoken by 17.5 million people. In this paper we will see the work has been done in the field of natural language processing (NLP) using Chhattisgarhi language and other state languages .main goal of NLP is to create machine learning, create translator, create dictionary and create POS tagger. POS tagger is one of the important tools that are used to develop language translator and information extraction so that computer based be compatible for natural language processing. Part-of-speech tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence. There are different types POS tagger are exist, are based on probabilistic approach and some based on morphological approaches. So in this paper we will see various developments of POS tagger and the major work has been done using Chhattishgarhi and other Indian state languages.**

***Keywords*— POS Tagger,Chhattisgarhi.**

I. INTRODUCTION

Chhattisgarhi (Devnagri) is official language in the Indian state of Chhattisgarh.Here the means of devanagri is compound of "Deva" and "Nagari" is an abugida(abuggida means is a segmental writing system in which consonant – vowel sequence are written as a unit based on a consonant letter and vowel notation is secondary[27]) alphabet of Indian and Nepal it is written from left to right does not have distinct letter cases (along with most other north Indic script exception Gujrati and Oriya) by horizontal line that runs along the top of full letters[28].Devnagri(since 19th century it has been the most commonly used script for writing sanskrit) is used to write hindi,marthi , nepali among other language and dialects. Chhattishgarhi is the eastern Hindi language with heavy vocabulary and linguistic features from "Munda" and Dravidian language. According to the Indian government Chhattisgarhi is eastern dialect of Hindi, but it is classified as separate in ethnology [27]

II. NLP

Natural language processing is an area of research and application that explores how computer can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim together knowledge on how Human beings understand and use language so that appropriate tools and techniques can be developed to understand and manipulate natural language to perform the desire tasks.[1], and goal of NLP is to find relation with words and identify its meaning from the language. A NLP has five major levels. Which are as follows

A. Phonology

Phonology is analysis of spoken language. There for it deals with the speech recognition and generation [30]

B. Morphological Analysis

Morphology deals with the word formation and it analysis, its punctuation and suffix. [30]

C. Lexicon

Lexicon deals with the validity of words and they belong to which category like Noun, pronoun, Verb, adverb so on[30]

D. Syntactic Analysis

Syntactic deals with the grammar of language and analysis them with the help of two phrasing techniques like top-down and bottom up approaches [30]

E. Semantic Analysis

The semantic analysis deals with language structure its meaning. [30]

F. Discourse integration

The Discourse is the collection of sentence for analysis understanding so on[30].

G. Pragmatic Analysis

The pragmatic level is relation between the language and context of use. Identify and how they related to people so on. [30]

III. POS TAGGER

A part of speech tagger is nothing but a software an application of Natural Language Processing used for assigning parts of speech in the natural languages, here the means of natural language are Hindi, English Gujrati, Marthi ,Bangali Punjabi ,Chhattisgarhi so on. the natural language which we speak write and understand use them for our day to day communication. so these language are known as natural language and when we process these language with help of the computer technology. its come under the field of natural language processing .while processing on any particular language assigning correct part of speech according to its respective grammar with the help of Software and that particular software is known as parts of speech tagger.

IV. POS TAGGING

Parts of speech tagging is process of tagging, assigning or labelling correct part of speech in the entered sentence of any language in POS tagger software. as we know language is made up of grammar and every language has their own grammatical rules and parts of speech as well. here the means of labelling correct parts of speech in the entered sentence .first we analysis the sentence an identify which is Noun, pronoun Adverb ,adjective ,verb, preposition conjunction , gender, number so on and label them correctly.

V. CLASSIFICATION OF POS TAGGER APPROACH

From the above section we under stand the part of speech tagger are software and the POS tagging is a process which is applicable in the POS tagger. For every different language we need separate POS tagger which made according to respective language grammatical rules. So the development of POS tagger is a major task and POS tagging is another critical task of the POS tagger. For the development of POS tagger few approaches are there which is classified in supervised and unsupervised category and contain few algorithms also which is shown on Fig. 1

A. *Supervised model*

The supervised POS tagging model requires pre tagged or pre annotated (annotated corpora serve as an important tool for investigators of natural language processing, speech recognition and other related areas [7]).Further divided into three parts rule based, stochastic and neural and also contain different POS tagging techniques like Brill, N-gram Maximum entropy HMM.

B. *Unsupervised Model*

The unsupervised POS tagging models do not require a pre annotated corpus .any likewise supervised method it also contain three types Further divided into three parts rule based, stochastic and neural and also contain different POS tagging techniques like Brill, N-gram Maximum entropy HMM.
HMMs are very simple stochastic models and present themselves with ease to modifications [8].HMM model is very simple any easy model to implement.

- Rule based
- Stochastic
- Neural

These three method of POS tagging are common in both supervised any unsupervised POS tagger Model but major difference between them occur they belong from which category supervised or unsupervised

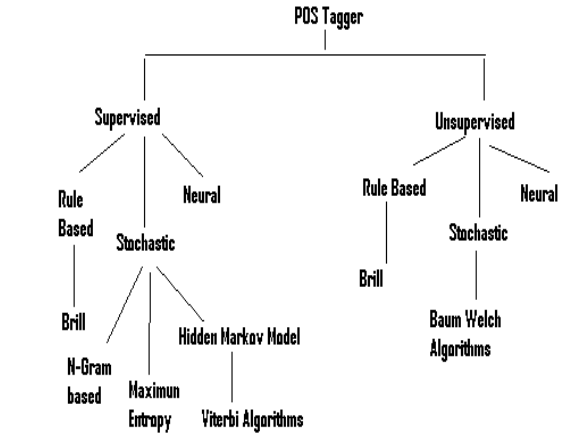

Fig.1 classification of POS tagging method

VI. LITERATURE SURVEY FOR CHHATTISHGARHI AND OTHER INDIAN STATE LANGUAGE

As we known India has bunch of Different languages is which is spoken by million people of Indian various POS tagger were developed in different language using different methods. now we will see earlier work has been done part of speech tagging for various Indian language.

1. POS tagger and Chunking with Conditional Random Fields developed by Himanshu Agrawal Anirudh Mani[10]. this system presents CRF (Conditional Random Fields) based part of speech tagger and chunker for hindi. Apart from CRF based learning using the CRF package "CRF++, Yet Another CRF Package", a morph analyzer is used to provide extra information like root word and possible PoS tags for training. With training on 21000 words with the best feature set, the CRF based POS tagger is 82.67% accurate, while the chunker performs at 90.89%.
2. POS Tagging and Chunking using Decision Forests Sathish Chandra Pammi,KishorePrahallad[12]They presents the building of POS Tagger and Chunk Tagger using Decision Forests and also focuses on the investigation towards exploring different methods for parts-of-speech tagging of Indian languages using sub-words as units. The two models POS Tagger and Chunk Tagger were tested with 3 different Indian languages Hindi, Bengali, Telugu and achieved the accuracies as 69.92%,70.99%, 74.74% and 69.35%, 60.08%,77.20% respectively
3. English—Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009 RejwanulHaque, SandipanDandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy

Way [13]In there project they presents English—Hindi transliteration in the NEWS 2009 Machine Transliteration Shared Task adding source context modeling into state-of-the-art log linear phrase based statistical machine translation(PB-SMT). Source context features enable to exploit source similarity in addition to target similarity, as modeled by the language model. They use a memory-based classification framework that enables efficient estimation of these features while avoiding data sparseness problems. they carried out experiments both at character and transliteration unit (TU) level. Position-dependent source context features produce significant improvements in terms of all evaluation metrics.

4. Using Rich Morphology In Resolving Certain Hindi-English Machine Translation Divergence,R. Mahesh K. Sinha [14] Identification and resolution of translation divergence (TD) is very crucial for any automated machine translation (MT) system. In there project, they present a technique that exploits the rich morphology of Hindi to identify the nature of certain divergence patterns and then invoke methods to handle the related translation divergence in Hindi to English machine translation. We have considered TDs encountered in Hindi copula sentences and those arising out of certain gaps in verb morphology.
5. Evaluating Stemmers and Retrieval Fusion Approaches for Hindi: UNT at FIRE 2010,Miguel E. RuizBharathDandala[15].In there work they describes the experiments conducted by the University of North Texas team as part of our participation in the Forum for Information Retrieval (FIRE). they concentrated on comparing the results using two morphological stemmers (YASS and Morfessor), studying the effect of using a part of speech tagger (Combined Random Fields) to weight the contribution of words with noun phrases, and to use a data fusion approach to improve performance of the system by combining these methods. conducted study using Hindi and explore the cross language retrieval performance from English to Hindi using Google translations. results show that using the YASS stemmer yields a small increase in retrieval performance. Fusion of results also showed to be effective and improved results 5% in the experiments.
6. Improving statistical POS tagging using Linguistic feature for Hindi and Telugu, PhaniGadde, Meher Vijay YeletiIn[16] they describe strategies for improving statistical POS tagging using Hidden Markov Models (HMM) for Hindi and Telugu. also describe a method for effective handling of compound words in Hindi. Experiments show that GNP1 and category information of a word are crucial in achieving better results. The maximum accuracy achieved with HMM based approach is 92.36% for Hindi and 91.23% for Telugu. result improvement of 1.85% in Hindi and 0.72%in Telugu over the previous methods.
7. Building Feature Rich POS Tagger for MorphologicallyRichLanguages:Experiences in Hindi AniketDalalKumarNagarajUmaSawant[17] They present a statistical part-of-speech(POS) tagger for a morphologically rich language: Hindi. there tagger employs the maximum entropy Markov model with a rich set of features capturing the lexical and morphological characteristics of the language. The feature set was arrived at after an exhaustive analysis of an annotated corpus.. The system was evaluated over a corpus of 15,562 words developed at IIT Bombay. Performed 4-fold cross validation on the data, and our system achieved the best accuracy of 94.89% and an average accuracy of 94.38%. result shows that linguistic features play a critical role in overcoming the limitations of the baseline statistical model for morphologically rich languages.
8. HMM Based Chunker for Hindi ,AkshaySingh,SushmaBendre.RajeevSangal[18] They presents an HMM-based chunk tagger for Hindi. Contextual information is incorporated into the chunk tags in the form of part- of-speech (POS) information. This in formation is also added to the tokens themselves to achieve better precision. Error analysis is carried out to reduce the number of common errors. It is found that for certain classes of words ,using the POS information is more effective than using a combination of word and POS tag as the token.
9. An HMM based Part-Of-Speech tagger and statistical chunker for 3 Indian languages ,G.M. Ravi Sastry ,SourishChaudhuri ,P. Nagender Reddy[19]In there project, they describe building an HMM based Part-Of-Speech (POS) tagger and statistical chunker for 3 Indian languages-Bengali,Hindi and Telugu. They employ the TnT tagger model for POS tagging of the corpus. The POS tagging accuracies for Bengali, Hindi and Telugu are 74.58, 78.35 and 75.37 respectively.
10. Large-Coverage Root Lexicon Extraction for Hindi,C ohanSujayCarlos,Monojit Choudhury Sandipan Dandapat[20] They describes a method using morphological rules and heuristics, for the au-omatic extraction of large-coverage lexicons of stems and root word-forms from a raw text corpus. the problem of high-coverage lexicon extraction as one of stemming followed by root word form selection. Examine the use of POS tagging to improve precision and recall of stemming and thereby the coverage of the lexicon.
11. A Text Chunker and Hybrid POS Tagger for Indian Languages Pattabhi R K Rao T, Vijay Sundar Ram R, Vijayakrishna R and SobhaL[21] Part-of-Speech (POS) tagging can be described as a task of doing

automatic annotation of syntactic categories for each word in a text document. This paper presents a generic hybrid POS tagger for Indian languages. Indian languages are relatively free word order, morphologically productive and agglutinative languages. In hybrid implementation used combination of statistical approach (HMM) and rule based approach. thetagset developed by IIIT, Hyderabad consisting of 26 tags. presents a transformational-based learning (TBL) approach for text chunking. In this technique of chunking, a single base rule (or a few base rules) is provided to the system, and the other rules are learned by system itself during the training phase for reorganization of the chunks

12. Word Sense Disambiguation in English to Hindi Machine Translation[22]Word Sense Disambiguation is the most critical issue in machine translation. Machine readable dictionaries have been widely used in word sense disambiguation. The problem with this approach is that the dictionary entries for the target words are very short. WordNet is the most developed and widely used lexical database for English.. The WordNet database can be converted in MySQL format and we have modified it as per our requirement. Sense's definitions of the specific word, "Synset" definitions, the "Hypernymy" relation, and definitions of the context features (words in the same sentence) are retrieved from the WordNet database and used as an input of Disambiguation.
13. Part-of-Speech Tagging and Chunking with Maximum Entropy Model, SandipanDandapat [23]There project is based on POS tagging and chunking for Indian Languages, for the SPSAL shared task contest. Maximum Entropy (ME) based statistical model. The Since only a small labeled training set is provided (approximately 21,000 words for all three languages), a ME based approach does not yield very good results. The tagger has the overall accuracy on development data of about about 83% for Hindi. The best accuracy achieved for chunking by there method on the development data 79.88% for Hindi on per word basis.
14. Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages Fahim Muhammad Hasan, NaushadUzZaman,Mumit Khan.[24]In there work Different methods of automating the process have been developed and employed for English and other Western languages.. They experimented with some of the widely-used approaches for POS Tagging on three South Asian languages, Bangla, Hindi and Telegu, using corpora of different sizes. The result performance of the approaches and found the Brill's transformation based tagger's performance to be superior to the other approaches.
15. Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources DebasisMandal, SandipanDandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar[25] In there project they experimented on two cross-lingual and one monolingual English text retrievals at CLEF1 in the ad-hoc track. The cross-language task includes the retrieval of English documents in response to queries in two most widely spoken Indian languages, Hindi and Bengali.to build statistical lexion Automatic Query Generation and Machine Translation and they are mostly dependent upon phoneme-based transliterations to generate equivalent English query from Hindi and Bengali topics. Other language-specific resources included a Bengali morphological analyzer, a Hindi stemmer and a set of 200 Hindi and 273 Bengali stop-words. Lucene framework was used for stemming, indexing, retrieval and scoring of the corpus documents. The CLEF results suggested the need for a rich bilingual lexicon for CLIR involving Indian languages. The best MAP values for Bengali, Hindi and English queries for experiment were 7.26, 4.77 and 36.49 respectively.
16. Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database Vishal Chourasia ,SamudravijayaK,ManoharChandwani [26] This paper they reports on methodology used in the generation of a phonetically rich Hindi text corpus. sThe corpus will be used as a resource for creation of a continuous speech, multi-speaker, and large vocabulary speech database for Hindi Language. This paper describes the design, structure and phonetic analysis of text corpus for Hindi. An analysis of the phonetic richness of sentences designed by this method is provided.
17. Pardeep Kumar, Vishal Goyal"Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments"[27]In there project paper, problem is "development of Hindi-Punjabi parallel corpus using existing Hindi to Punjabi machine translation system and using sentence alignment". The alignment based on the length based technique, location based technique and lexical techniques. They use Hindi-Punjabi machine translation system (i.e h2p.learnpunjabi.org). These tasks are need to Hindi-Punjabi parallel corpus. Sentence alignment is useful to developing Hindi-Punjabi parallel corpus and Hindi-Punjabi dictionary. The accuracy is basically depending upon the complexity of the corpus, more the complexity less the accuracy. Complexity means how to distribution of sentence in the target file. If any of these categories 1:1, 1:2, 2:1, 1:3, 3:1 sentences occur simultaneously in a paragraph
18. An improved Hindi POS tagger was developed by employing a naive (longest suffix matching) stemmer

as a pre- processor to the HMM based tagger [3]. Apart from a list of possible suffixes, which can be easily created using existing machine learning techniques for the language, this method does not require any linguistic resources. The reported performance of the system was 93.12%.[8][4]

VII. CONCLUSIONS

In this paper we have seen the development of POS tagger and the work has been carried for different Indian language. We found that most of work based on statistical approach,HMM model ,maxium entropy model are used in the development.we found that no work has been carried out in the Chhattisgarhi language.

REFERENCES

[1] Gobinda G.Chowdhary.*"Natura language processing"*, ,Dept of computer and information science,university of strathclyde,Glasgow G1 1XH,UK.

[2] Artificial intelligence by Elaine rich and Kevin knight.

[3] .Department of Information Technology Ministry of Communications & Information Technology Govt. of India "Unified Parts of Speech (POS) Standard in Indian Languages - Draft Standard –Version 1.0 "

[4] Antony P J ,Research Scholar, Computational Engineering ,and Networking (CEN), *"Parts Of Speech Tagging for Indian Languages: A Literature Survey"*

[5] fahim Muhammad hasan."comparison of unigram,bigram,HMM and brill's POS tagging approaches for some south asian languages".,

[6] GeorgiGeorgiev and Valentin ZhikovPetyaOsenova and KirilSimov,"Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian"

[7] .Apart of speech tagger for Indian language(POS tagger) tagset developed at IIT-Hyderabad.

[8] Manish Shrivastava and Pushpak Bhattacharyya (2008), *"Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge"*, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay.

[9] juananotoniop'erez-ortiz and Mikel L.Forcada,"*part of speech tagging with recurrent Neural networks*"universitatd'Alacant,Spain 2002

[10] HimanshuAgrawal,Anirudh Mani" Part of Speech Tagging and Chunking with Conditional Random Fields."

[11] Smriti Singh, Kuhoo Gupta"*Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi"*Department of Computer Science and EngineeringIndian Institute of Technology, BombayPowai, Mumbai

[12] Sathish Chandra Pammi,KishorePrahallad" POS Tagging and Chunking using Decision Forests"

[13] RejwanulHaque, SandipanDandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy Way" *English—Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009* "

[14] R. Mahesh K. Sinha" using rich morphology in resolving certain hindi-english machine translation divergence

[15] Miguel E. RuizBharathDandala" Evaluating Stemmers and Retrieval Fusion Approaches forHindi: UNT at FIRE 2010"

[16] PhaniGadde, Meher Vijay YeletiIn" Improving statistical POS tagging using Linguistic feature for Hindiand Telugu"

[17] AniketDalalKumarNagarajUmaSawant" .Building Feature Rich POS Tagger for Morphologically Rich Languages:Experiences in Hindi"

[18] AkshaySingh,SushmaBendre.RajeevSangal*" HMM Based Chunker for Hindi"*

[19] G.M. Ravi Sastry ,SourishChaudhuri ,P. Nagender Reddy" .An HMM based Part-Of-Speech tagger and statistical chunker for 3 Indian languages"

[20] Cohan SujayCarlos,MonojitChoudhurySandipanDandapat" Large-Coverage Root Lexicon Extraction for Hindi"

[21] .Pattabhi R K Rao T, Vijay Sundar Ram R, Vijayakrishna R and Sobha" .A Text Chunker and Hybrid POS Tagger for Indian Languages"

[22] "Word Sense Disambiguation in English to Hindi Machine Translation"

[23] SandipanDandapat" Part-of-Speech Tagging and Chunking with Maximum Entropy Model SandipanDandapat"

[24] Fahim Muhammad Hasan, NaushadUzZaman,Mumit Khan" Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages"

[25] DebasisMandal, SandipanDandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar " Bengali and Hindi to English Cross-languageText Retrieval under Limited Resources"

[26] Vishal Chourasia ,SamudravijayaK,ManoharChandwani" Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database

[27] http://en.wikipedia.org/wiki/Chhattisgarhi_language

[28] http://en.wikipedia.org/wiki/Devanagari

[29] ardeep Kumar, VishalGoyal"Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments"

[30] natural language processing by ela kumar