Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 6, Issue 05, May -2019

CURRENT STATE OF THE ART PARTS OF SPEECH TAGGING FOR INDIAN LANGUAGES –A STUDY

Pragya Bhagat[1], Dr.Piyush Pratap Singh[2]

[1,2]*Department of informatics and language Engineering,Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya,Wardha,Maharashtra, India*

Abstract — *parts-of speech tagging is a pipeline module for almost all application areas of natural language processing (NLP). POS Tagging is a very important preprocessing task for language processing activities. This paper reports about the parts of speech systems proposed for 15 Indian languages(Hindi, Panjabi, Marathi, Gujrati, Kannad, Tamil, Telgu, Malyalam, Manipuri, Konkani, Bengali, Assames, Odia, Sambalpuri, Sindhi). In this paper, all the approaches have been also briefly discussed which is used in POS tagging.*

Keywords- *Ambiguity, Tagset, Natural Language Processing, Part of Speech Tagging, Rule Based Approach, Statistical Approach, Hybrid Approach.*

I. INTRODUCTION

Natural language processing is a subfield of artificial intelligence and an interaction between the computer and human. It is also related to computational linguistics. The process in which each word of each sentence is categorized based on its properties, definition, context is called part of speech tagging or grammatical tagging, part of speech helps in deep parsing of a text.It helps in developing information extraction system, helps in semantic processing etc.

II. CLASSIFICATION OF POS TAGGER APPROACHES

The word 'attributable' of sentences entered into a language by the POS tagger software is tagged in its absolute grammatical categories (noun, pronoun, verb etc.), which is used in the following approaches: rule based, stochastic, neural network.

2.1. Rule Based Approach - In this approach, a set of linguistic rules written by hand to tag a word is used. These rules are also known as context frames rules.

2.2. Stochastic Approach - This method uses the statistics (frequency or probability) to tag input text, that is, in this method, the annotated training data provides the most frequently used tag for the particular word, which is unannoted in this method, there is a large number of corpus in this method. The correctness of this method is the parellal of the corpus size. The decision of this method depends on the calculation of frequency, probability or statistics of corpus.
The stochastic tagging model is classified into two parts, supervised and unsupervised .supervised POS tagging models require pre-tagged corpus, on the contrary, unsupervised tagging models do not require an already annotated corpus in which advanced computationl techniques are used. The stochastic tagger is based on the following different model: Hidden Markov Model (HMM), Maximum Likelihood Estimation, Decision Trees, N-Grams, Maximum Entropy, Support Vector Machines Or Conditional Random Fields.

2.2.1. Hidden Markov Model (HMM) –Hidden Markov Model large corpora to automatically reach the finite state and tagging by removing perfect tag for a particular word. The combined probability of the state emits the special symbol in the first probability then enter into 2nd state. Because this joint probability can not be observed, therefore this is called Hidden Markov Model. In this approach, untagged text is required only as lexicon and training data. The main purpose of this method is to develop a perfect language model in less effort. This Method only gives us the probabilistic function of state sequence. In the initial process of tagging, some initial tag probabilities are assigned, after which these tag probabilities are refine by the attributable training cycle. In this refining process, search algorithm (viterbi algorithm). The viterbi algorithm is used as a dynamic programming algorithm which is used in lexical calculations.

2.2.2 Maximum Entropy Approach (ME)- Maximum Entropy Model to find the probability of a specific category in context of a specific language, provide a specific approach for linking the pieces of all the evidence obtained in the context of that linguistic category. It is higher and accurate than hmm. Maximum entropy model is the probability of the applied constraints. They are obtained from the constraints training data and create relation between the properties and the result.

2.2.3. Conditional Random Field (CRF) - This is a discriminative probabilistic model that helps in sequence segmentation and labeling, meaning it is also an undirected graphical model, which provides a sequence of lebels with

the highest probability for input sequence. CRFs Can also be understood as a supervised learning model, which creates a set of feature functions and corresponding corresponding weights for classification. There are all the benefits of MEMMs without label bias problem.

2.2.4. Support Vector Machine (SVM) -SVM is a machine learning approach, it is commonly used in classification and regression. This approach is based on maximum marginal hyperplane (MMH) in which hyperplane is searched with the largest margin. This margin is divided between two classes. shows the largest separation. SVMtool is a simple, flexible generative model for sequential tagging, whose main purpose is to complete all of the requirements (simplicity, flexibility, robustness, portability and efficiency) of mordern NLP technology with accuracy. This is a machine learning for binary classification. The algorithm that is used to solve many practical problems including NLP is also used in pattern recognition, this approach is also used in text categorization.

2.3. Neural Network - Neural taggers are based on the neural network which learns the parameters of POS tagger from a representative training data set. The performance was better than stochastic taggers.

2.4. Hybrid approach - The hybrid model is basically a combination of rule based approach and stochastic models. The hybrid system, approach uses a combination of rule based and machine learning techniques from each method, which aims to make the POS tagging system more accurate.

III. CURRENT WORKS IN INDIAN LANGUAGES

The work of tagging of Indian languages is still limited. Its main reason is limited availability of explanatory corpora and morphological richness of Indian languages. The works of POS tagging are described in various universities and research centers.

3.1. HINDI

Hindi is our national language which is 4th in languages used in the world. Hindi morphological rich language and relatively free word order language.

Garg.Etal [1] developed systems for Hindi language using the rule based method, whose corpus used 26,149 words and 30 tags and the accuracy of their system is 87.55%. Modi Etal[2] developed a system for Hindi language using the rule based approach, in which he created 9,000 words of database and used 29 tags, out of which 27 tags for IIIT hydrabad tagset and 91.84% is indicates the accuracy of their system. Shrivastava Et al[3] developed the system for Hindi Based on the hidden markov model, , which employs the longest suffix matching (naive stemmer) as a pre-processor, created a corpus of 66,900 words for training and testing purpose and the system shows 93.12% accuracy. Joshi Et al[4] used the hidden markov model approach for built tagger in which he used POS Tagset manufactured by Bharti Et al. He developed training corpus of 15,200 (3,58,288 words)sentences and accuracy 92.13% of the system. Dalal et al [5] developed tagger and chunker for Hindi, for which they took the maximum entropy markov model (MEMM) in which 35,000 words used as corpus, using 29 tags for tagger and 6 chunk tags, and the accuracy of the tagger was 89.35% And the accuracy of the chunker is 87.40%. Khicha Et al[6] created pos tagger by assuming both the markov model and the rule based method (hybrid approach), for which 13,000 words were created as corpus and the system showed 87.55% accuracy. Mohnot Et al[7] also created a pos tagger using 80,000 words. The corpus has been developed and the system shows 89.9% accuracy. Narayan Et al[8] made use of artificial neural network for Hindi language, for which he created a corpus of 2,600 sentences (11,500 words) and the system shows 91.30% accuracy. Jabar H. Yousif Et al[9] used the neurosolution package to develop a multilayered network system developed by assuming the multilayered perceptron (mlp) concept base and the system shows 82.8% accuracy.

3.2. PANJABI

Punjabi is the language of indo-aryan family which is spoken in countries like india, pakistan, usa, canada, england etc. Punjabi is written in eastern panjab in "Gurukhi" script and written in "shahmukhi" script in western Punjab.

Singh M. Et al developed the first POS tagger for Punjabi as a module based on this rule based approach in which hand written rules created for removal of the ambiguity and accuracy of the system is 80.29%. Mittal S. Et al[11] has made Bi-gram model for Punjabi words , for which they produced corpus of 24,000 sentences (10,000words) from online resources and their system shows 92.16% accuracy. Dinesh Et al [12]Created a tagger for Punjabi language using the vector machine (SVM) approach, in which he created corpus of 27,000 words, his tagging work is completed in 3 steps, vectorization phase, training, classification . In vectorization phase, manually tagged Punjabi file is converted into SVM format. During training SVM files inputted from vectorization phase are trained, its output is that the model file is created for every tag. The last phase is the classification phase in which untagged file along with the model file created during the training phase is given as input and the tagged file will be generated as output. Singh U. Et al [13] developed the system for Punjabi language using the hybride approach (rule based + HMM) in which he used the Panjabi corpus of TDIL which used 49,319 sentences (approx.63,000 words) 35 tagset used in this system With the rule based system, they

also developed the trigram HMM based tagger. In the rule based system, 150 linguistic rules developed to tag text. in the full corpus, 44,387 sentences were for hmm system, which is 90% of total corpus and full system Accuracy is 93.33%. Kanwar S. Et al[14] has made tagger for Punjabi using the HMM approach. and the accuracy of the system is 86.2%. Sharma S.K. Et al[15] developed the system using a bi-gram HMM based model in which he trained 20,000 words of the annoted corpus and the accuracy of the system is 90.11%

3.3. MARATHI

Patil H.B. Etal[22] developed a system for Marathi language using the rule-based approach, for which 576 unique words developed for the corpus which 9 tags were used, in which they considered language specifications to overcome the ambiguity and system accuracy is 78.82%. Rathod S.Etal [23]developed the corpus of 17,197unique words by using the rule based method for which he used 10 document randomly selected 29 tagset and developed 141 rules for disambigution and the system also compared with the already made NLTK system and shallow parser. In which system correctly tag to 677 words from 811 words and accuracy of system 95.05%. Bagul P. Etal[17], gaikwad D.K. etal[20], govilkar S. Etal[24] developed a POS tagger for Marathi using the rule based method. Patil N. Etal[18] created a POS tagger for Marathi language using the supervised learning method, in which he used unigram, bi gram, trigram language model and using viterbi decoding algorithm to identify the most probable tag in the word sequence. Trained data for 12,000 sentences, testing data is 3,000 sentences and system 86.61% indicates accuracy.Singh J. Etal[19] developed the system using statistical method (trigram) whose corpus size is 2,000 sentences (48,635 words) and The accuracy of system 91.63%. singh J. Etal[21] Created a POS tagger using statistical Approachs Unigram, bigram, trigram and HMM, for which he developed a corpus of 1,000 sentences (25,744words) and the system's accuracy was 77.39% for unigram , 90.30% for bigram, 91.46 % for trigram, and HMM is 93.82%.

3.4. GUJRATI

Gujrati poor Resources Indo-Aryan language is the number of gujrati speaking people 46.1 million.

Patel C. Etal[25] created a POS tagger for the gujrati language using the CRF model (a machine learning algorithm) using 600 sentences (tagged) for learning and 5,000 sentences (untagged) using the tagset created by IL and 26 different tagset also created. For training corpus size 10,000 words and testing data 5,000 words and accuracy of system 92%. Hala SY., Etal[26] created tagger for the gujrati language using the hybrid approach (rule based + HMM) in which they used the tagset made by the BIS (bureau of indian standards) and used corpus of 10,000 sentences .Prajapati M. Etal [27] has developed a system for gujrati language using SVM approach that has a corpus of 1,700 words, the accuracy of the system is 97.40%.

3.5. MANIPURI

Manipuri (meiteilon) is one of the oldest languages of Southeast Asia, which has its own script (Meitei Mayak) and literature. At present, Manipuri was writing in Bengali script. Manipuri is widely spoken in Manipur, Assam, Tripura, Bangladesh and Myanmar, which has been included in the eight Schedules of the Indian Constitution since 1992. The interesting thing is that this is the first tibeto-burman language that has obtained its proper place and recognition in the Indian Constitution.

Kh Raju Singha Etal[29] created a POS tagger for the Manipuri language using a rule based approach, in which he made 25 hand written Linguistic rules in classes in 3 parts. 1) orthographic rules 2) morphological rules 3) disambiguation rules. He used the affix strippring technique to separate the affix from the root. In this system, he created lexicon of 1000 words and the accuracy of the system is 85%. Kh Raju Singha[30] used the HMM method to output of their rule based system to another System developed for which 2000 used lexicon for the corpus, the accuracy of the system is 92%. Kishorjit Nongmeikapam etal[28] done part of speech (POS) tagging of Bengali Script Manipuri text using the Conditional Random Field (CRF) which is then followed by the transliteration to Meitei Mayek, in which he created 30,000 words of corpus, in which 24000 words training file and 6000 words as testing file and the accuracy of the system is 86.04%. Youdaam doren singh etal[31] developed morphology driven POS tagger for Manipuri language, which included a morphology driven Manipuri POS tagger that uses three dictionaries containing root words, prefixes and suffixes. Designed and implemented using the affix 3,784 sentences in which 10,917 unique words are tested, in which the accuracy of the system is 69% and 23% of the remaining 31% of the words are unknown words and 8% are incorrectly tagged. Thudam Doren Singh etal [32] using both CRF and SVM approaches For the purpose of developed the system for Manipuri language, for which 63,200 tokens were collected in which used 39,449 tokens for training file and 8,672 tokens for testing file, in which system's accuracy is 72.04 % for CRF and 74.38% for SVM.

3.6. BENGALI

Bengali is one of the most widely used languages In terms of native speakers, it is the seventh popular language in the world, second in India and the national language of Bangladesh.

Asif ekbal et al[33] created a tagger for Bengali language using the statistical maximum entropy (ME) model. The system makes use of different types of POS classes.in which training file size 72,341 words testing file 20,000 Words taken and

the accuracy of the system is 88.2%, along with the performance of the system also compared with other HMM-based tagger, where improvment of 8% was detected. Asif Ekbal et al[34] developed Tagger using the maximum entropy (ME), conditional random field (CRF) and the support vector machine (SVM) framework, and then they combines in one order with weighted voting techniques that allow multiple classifiers to get higher accuracy . All models assembled in the final system have been evaluated only on the same database 72,341 developed corpus of tokens 27 tags used. 15,000 tokens of 72,341 tokens as development sets And the remaining tokens as a training set, the test set accuracy 81.91% for ME, 84.23% for CRF, 85.92% for SVM. Asif ekbal et al[35] developed the POS tagger for the Bengali language using the CRFs method whose accuracy is 90.3%. Sandipan dandapat et al[36] developed the system using the HMM and ME method, he described the approach of automatic stochastic tagger for Bangla language, using 3625 sentences (approximately 40000 words) as training data and tagger's performance improved. The morph analyzer is used to make the accuracy of the system is 76.8%.

3.7. SAMBALPURI

Sambalpuri Bhasha is the Eastern Indo-Aryan language, and is spoken in the south, Kosli, Koshal, Koshli, Western and other languages. Yah is a low-density valet Isme 70-76% lexical similarity is the standard language of the language.

Pitambar behera et al [37]two statistical method SVM and CRF ++ were used to describe the POS tagger. 1,21,210 words of corpus developed by 80,288 words for training data and testing data of 12,791 words. SVM's accuracy is 83% and CRF ++ accuracy is 71.56%.

3.8. KONKANI

Konkani is an Indo-European, morphologically rich language. It is one of the twenty two languages incorporated into the Eighth Schedule of the Indian Constitution. Konkani is a language that is spoken in the state of Goa with
Devanagari as the officially recognized script Konkani also uses loan words from Sanskrit, Perso-Arabic, Kannada, and Portuguese.

Diksha N.Prabhu Kharjuvenkar et al[38] tagged Konkani language using hmm method . konkani's pre-tagged corpus training data to translate and translate into Konkani's text testing data. The POS model has been divided into two layers. 1st visible layers in input words and 2nd layer hidden layer learnt by the system with respect to the tags.

3.9. SINDHI

Sindhis are an Indo-Aryan ethno-Linguistic group who speak the Sindhi language and are native to the Sindh province of Pakistan. sindhi language script is based on Persian, Arabic script Phonological system of sindhi language is like second indo-aryan languages; 43 distinct cosonant phonemes and 10 vowels in the sindhi language include 3 short vowels [a, i, u] and 5 long vowels [aa , ii, uu, e, o] sindhi is highly homographic language, sindhi language is written in life without diacritities, which is a significant cause of lexical and morphological ambiguity.

Javed Ahmed Mahar Et al [39] developed a rule-based semantic Part of Speech (POS) tagging system, which relies on wordNet to identify the analogical relationships between words. Two types of lexicons used one is for simple word and another for disambiguated word. corpus collection was done with the Sindhi dictionary. corpus as tagged Data (training) of 26,366 words and untagged data (testing) took 6,738. In testing, were corpora divided into 2 types. 1) Those words that are without ambiguity.2) Other words that are ambiguity in the absence of diacritics. the accuracy of 96.28% was achieved but after applying WordNet approach the accuracy of tagger was increased up to 97.14%.

3.10. ODIA

Odiya is a classical Indo-Aryan language spoken in the Indian state of Odisha .It is the official language in Odisha (Orissa) where native speakers constitutes 82% of the population, also West Bengal, Jharkhand, Chhattisgarh, and parts of the spoken areas. Andhra Pradesh It is the official language of Odisha and the second official language of Jharkhand.The language is also spoken by a large population of less than 1 million people in the state of Chhattisgarh.

Bishwa Ranjan Dasa et al[40] developed a POS tagger for the odia language using the support vector machine approach, for which they took 10,000 odia words as corpus and the accuracy of the system is 81%.

3.11.ASSAMESE

Assamese is an Indo European language that is spoken by 32 million people. It is a morphologically rich, free word order inflectional language.

Navanath Saharia et al [41] created tagger for the assamese language using the HMM method, as tags of the assamese language were not devlop, he also tagged 172 tags as manually asamiya pratidin from the assamese daily newspaper, manually tagged corpus of 3,00,000 words, of which 10,000 words Used as training data and used as testing and the accuracy of the system is 87%.

3.12. TAMIL

Tamil language belongs to Dravidian language family. It is spoken in sri lanka and south india. Tamil is low resourced, morphologically rich language. It is a complex grammatical structure language that is spoken in sri lanka and south india.

R.akilan et al[42] classical tamil POS tagger created using rule based method for texts This method is based on form aggrement . noun forms are type pattern, verb forms are token pattern. classical tamil tagset divide into 2 basic classifications noun morphology and verb morphology. Mokanarangan Thayaparanetal et al [43] used a graph-based semi supervised learning approach to classify unlabelled data in the tagging of the Tamil language. Using this word embedding in this approach both labelled and unlabelled to convert into vectors and weighted using mahalanobis distance After this, graph created classified the unlabelled data using semi-supervised learning algorithms; this work was done by talukdar pereira 2010's case study (different algorithms for classification in graph). They used 60,000 words for training data 20,000 words for testing data.accuracy 0.8743 as compare to CRF Tagger 0.7333 . Dhanalakshmi V. Et al [44] Created Tagger and chunker for Tamil language using machine learning technique support vector machine. 32 tegset used corpus size 2,25,000 words (1,65,000 traning set, 60,000 testing set) system accuracy for tagger 95.64% for chunker 95.82%. S. lakshmana pandian et al [45] used a corpus-based approach in the tamil POS tagging, in which a language model was created using morpheme componants of the words, using the generalized iterative scaling technology for a factor of the estimation of this model, the accuracy of the model 96 %.

3.13. TELGU

Telgu is classified as a dravidian language with heavy Indo -Aryan influence. It is a official language of Andhra Pradesh. Telugu grammatical rule is deduced from a Sanskrit Canon. Telugu uses too many words together, forming complex words.

G.Sindhiya Binulal et al[46] created the tagger for the telgu using SVMTool SVMTool software package consists of 3 main componants, namely the model learner (SVMTlearn) the tagger (SVMTagger) and the evaluator (SVMTeval). They specifically explained this How to use the binary classifier for multiclass classification problem, . 10 tags used for this system. Took the corpus of 25,000 words in which 20,000 words for training set and 5,000 words for testing set. System total accuracy 95%. Phani gadde et al[47] developed a POS tagger for Hindi and telgu using statistical approach (HMM) and their main characteristic how adding features to HMM improves its accuracy we also describe a method for effective handling of compound words in hindi. Following tools used for experiments on statistical POS tagging. 1) brant TnT (Brants, 2000), a HMM based tagger. 2) CRF ++, a CRF based tagger. In the entire tagger they used TnT but used both tools in compound word handling (for Hindi tagger). training and testing corpus size for hindi 1,85,000 words and 23,483 words respectively And training and testing corpus size for telgu 1,90,006 words and 21,351 words respectively. The maximum accuracy achieved with HMM based approach is 92.36% for hindi and 91.23% for telgu.and CRF accuracy 93.13% for telgu . Srinivasu Badugu et al[48] created a morphological based automatic tagger without any machine learning algorithm or training data for telugu language. According to them, the critical information needed for the tagging of inflectional and agglutinating languages is more than context, with the term internal structure. And also explained how a well-designed morphological analyzer can assign a correct tag and tag ambiguities to a great extent. lexical and semantic information that is useful or usful for syntactic parsing. 50 million words corpus size and testing data size 15 million words, and the accuracy of the system is 94.99%.

3.14. MALAYALAM

Malyalam language is spoken in India predominantly in kerala state. It is one of India's 22 scheduled languages, malyalam has official language status in the state of kerala and in the union territory Lakshadweep and Pondicherry belongs to the drividian family and 38 million is spoken by people in Malayalam, Malayalam is incorporated into many elements from Sanskrit through the age and over 80% of the vocabulary of Malayalam in scholarly usage is from Sanskrit.

Rajeev RR et al[49] tagging for the Malayalamalam language using SVMTool and TnT tagger.The svmtool support is implemented by the vector machine and tnt tagger is implemented by the Hidden Markov Model, using corpus of 200,000 malayalam words .svmtool's accuracy is 87.5 %,and the accuracy of tnt tagger is 80%. Bindu.MS et al [50] made the POS tagger and ambiguity resolver using the high order conditional random field approach, as in 80.65% words compounds words in malayalam text documents, sometimes for more than one morphological analysis and more than one parts of speech is available for single word. Currently available tag sets for the single word only give significance to morphological and syntactical properties, which they designate the taget they think of semantic features. For corpus text, created a corpus of 2,352 sentences from different areas of malayalam language, whose Testing on both word level and sentence level. The accuracy of the system is 91% on the sentence level and 95% on the word level. D. Muhammad Noorul Mubarak et al[51] made tagger keeping in mind the merits and demerits of the rule based method and stochastic based method. Used for the evaluation of words or lexicon and the structure of sentences used for texts created by IIIT hydrabad for tagset. Tagged corpus of 20,000 words made.tagging done with the probability of the dictionary entry. Ajees A P et al [52] created the tagar malayalam language using CRF method and Corpus size took 23,000 words in which 5,700 words as test set. System accuracy is 91.2%. Jisha P Jayan et al[53] used the POS tagger and chunker model

for the malayalam language using the proposed viterbi algorithm. Viterbi algorithm is the only algorithm that implements n-gram approach corpus size of 15,245 words. System accuracy tagging 90.05% and chunking 92%.

3.15. KANNADA

It is the Dravidian family of languages Within Dravidian, it is the South Dravidian group. Kannada or Canarese is one of 1652 mother tongues spoken in India. Forty three million people use it as their mother tongue. Kannada has 44 speech sounds Among them 35 are consonants and 9 are vowels.

Shambhavi.B. R et al [54] developed the system for the Kannada language using the maximum entropy approach, as the training data, 51,267 words were manually tagged words collection made from EMILIE corpus and the accuracy of the system was 81.6%. After that Shambhavi.B. R et al [55] again tagged for the kannada language using the second order HMM approach and CRF approach. system accuracy HMM and CRF respectively, 79.9% and 84.58%. MC PADMA et al [56]created a morpheme based POS tagger model, which collected the word from EMILIE corpus and system accuracy is 90%. Shriya Atmakuri et al[57] presented the CRF approach based model which used corpus of 17,175 sentences and used 2,18,530 tokens. The system accuracy is 89.1%. Pallavi et al [58] developed the CRF based system over the kannada language, which created a corpus of 8,000 words and the accuracy of the system is 92.04%

IV. CONCLUSION

Finally, the conclusion is that part of speech tagging is the most important activity for natural language based applications.generating the most efficient pos tagger is a challenging task in every research work. In this paper, we have tried to give a brief idea about the various approaches that the use of pos tagger tools for indian languages. The method used in the creation of various pos tagger developed for various indian languages has been presented and brief descriptions of their accuracy and methods that used in construction of pos tagger etc. This survey revealed that the pos tagger the accuracy depends on its corpus size. That is, the larger the corpus the system increases the accuracy.

REFERENCES

1. Navneet Garg,, Vishal Goyal, Suman Preet " Rule Based Hindi Part of Speech Tagger" Proceedings of COLING 2012: Demonstration Papers, pages 163–174. 2012
2. https://www.researchgate.net/publication/301720820
3. https://www.researchgate.net/publication/252060697
4. Nisheeth Joshi, Hemant Darbari and Iti Mathur " HMM BASED POS TAGGER FOR HINDI" Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA -pp. 341–349,. 2013
5. https://www.researchgate.net/publication/241211496
6. Vijeta Khicha, Vijeta Khicha, " Part-of-Speech Tagging of Hindi Language Using Hybrid Approach" International Journal of Engineering Technology Science and Research ISSN 2394 – 3386 Volume 4. 2017
7. Kanak Mohnot, Neha Bansal, Shashi Pal Singh, Ajai Kumar " Hybrid approach for Part of Speech Tagger for Hindi language" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 4, Issue 1. 2014
8. Ravi Narayan, S. Chakraverty,V. P. Singh "Neural Network based Parts of Speech Tagger for Hindi"Third International Conference on Advances in Control and Optimization of Dynamical Systems March 13-15. 2014
9. Jabar H. Yousif , Dinesh Kumar Saini "Hindi Part-of-Speech Tagger Based Neural Networks" JOURNAL OF COMPUTING, VOLUME 3, ISSUE 2, FEBRUARY,2011
10. Manjit kaur, Mehak Aggerwal, Sanjeev Kumar Sharma " Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set" International Journal of Computer Applications & Information Technology Vol. 7, Issue II . 2015
11. Sumeer Mittal , Mr Navdeep Singh Sethi, Sanjeev Kumar Sharma "Part of Speech Tagging of Punjabi Language using N Gram Model" International Journal of Computer Applications (0975 – 8887) Volume 100– No.19. 2014
12. Dinesh Kumar,Gurpreet Josan " Prediction of Part of Speech Tags for Punjabi using Support Vector Machines"The International Arab Journal of Information Technology, Vol. 13, No. 6. 2016

13. Umrinderpal Singh, Vishal Goyal " Punjabi Pos Tagger: Rule Based and HMM" International Journals of Advanced Research in Computer Science and Software Engineering ISSN: 2277-128X (Volume-7, Issue-7)

14. https://www.scribd.com/document/70848503/POS-TAGGING-OF-PUNJABI-LANGUAGE-USING-HIDDEN-MARKOV-MODEL

15. https://www.researchgate.net/publication/252017355

16. Antony P J,,Dr. Soman K P "Parts Of Speech Tagging for Indian Languages: A Literature Survey" International Journal of Computer Applications (0975 – 8887) Volume 34– No.8. 2011

17. Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar "Rule Based POS Tagger for Marathi Text" International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 1322-1326. 2014

18. Nita V. Patil " POS Tagging for Marathi Language using Hidden Markov Model" International Journal of Computer Sciences and Engineering Volume-6, Issue-1 E-ISSN: 2347-2693. 2018

19. Jyoti Singh, Nisheeth Joshi, Iti Mathur "PART OF SPEECH TAGGING OF MARATHI TEXT USING TRIGRAM METHOD" International Journal of Advanced Information Technology (IJAIT) Vol. 3, No.2. 2013

20. Gaikwad Deepali K., Naik Ramesh R. , C. Namrata Mahender 2018 "Rule Based Part-of-Speech Tagger for Marathi Language"International Journal of Scientific Research in Science and Technology [volume (4) issue 5 : 1607-1612]

21. Jyoti Singh, Nisheeth Joshi, Iti Mathur "Development of Marathi Part of Speech Tagger Using Statistical Approach DOI: 10.1109/ICACCI.2013.6637411 ·2013

22. H.B. Patil, A.S. Patil, B.V. Pawar "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora" International Journal of Computer Applications (0975 – 8887) Recent Advances in Information Technology, 2014

23. Shubhangi Rathod, Sharvari Govilkar ,Sagar Kulkarni "Part of Speech TAGGER for MARATHI Language" Sixth International Conference on Computational Intelligence and Information Technology . 2016

24. Sharvari Govilkar, Bakal J, W Shubhangi Rathod "Part of Speech Tagger for Marathi Language" International Journal of Computer Applications (0975 – 8887) Volume 119 – No.18. 2015

25. Chirag Patel ,Karthik Gali "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields"Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 117–122, Hyderabad, India. 2008

26. Sirajuddin Y. Hala, Sagar H. Virani "Improve accuracy of Parts of Speech tagger for Gujarati language" International Journal of Advance Engineering and Research Development Volume 2,Issue 5. 2015

27. Manisha Prajapati, Archit Yajnik, "POS Tagging of Gujarati Text using VITERBI and SVM" International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 43, 2019

28. Kishorjit Nongmeikapama,,Sivaji Bandyopadhyay"A Transliteration of CRF Based Manipuri POS Tagging" 2nd International Conference on Communication, Computing & Security [ICCCS-2012] Procedia Technology 6 582 – 589(2012)

29. Kh Raju Singha, Bipul Syam Purkayastha,Kh Dhiren Singha "Part of Speech Tagging in Manipuri: A Rule-based Approach"International Journal of Computer Applications (0975 – 8887) Volume 51– No.14. 2012

30. Kh Raju Singha, Bipul Syam Purkayastha ,Kh Dhiren Singha "Part of Speech Tagging in Manipuri with Hidden Markov Model" International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, ISSN (Online): 1694-0814. 2012

31. Thoudam Doren Singh, Sivaji Bandyopadhyay "Morphology Driven Manipuri POS Tagger" Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98,Hyderabad, India, January 2008

32. Thoudam Doren Singh, Asif Ekbal, Sivaji Bandyopadhyay "Manipuri POS Tagging using CRF and SVM: A Language Independent Approach"Proceedings of ICON-: 6th International Conference on Natural Language Processing2008

33. Asif Ekbal, Rejwanul Haque,Sivaji Bandyopadhyay "Maximum Entropy Based Bengali Part of Speech Tagging"Advances in Natural Language Processing and Applications Research in Computing Science 33, , pp. 67-782008

34. Asif Ekbal, Md. Hasanuzzaman, and Sivaji Bandyopadhyay "Voted Approach for Part of Speech Tagging in Bengali" 23rd Pacific Asia Conference on Language, Information and Computation, pages 120–129. 2009

35. https://www.researchgate.net/publication/228955119

36. Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu "Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario" Proceedings of the ACL 2007 Demo and Poster Sessions, pages 221–224,Prague. 2007

37. Pitambar Behera, Atul Kr. Ojha, Girish Nath Jha "Issues and Challenges in Developing Statistical POS Taggers for Sambalpuri" LNAI 10930, pp. 393–406, 2018

38. Diksha N,.Prabhu Khorjuvenkar, Megha Ainapurkar, Sufola Chagas "Parts of Speech Tagging For Konkani Language" International Journal of Engineering Research in Computer Science and Engineering Vol 5, Issue 2, 2018

39. Javed Ahmed Mahar , Ghulam Qadir Memon "Sindhi Part of Speech Tagging System Using Wordnet" International Journal of Computer Theory and Engineering, Vol. 2, No. 4, pp.1793-8201. 2010

40. Bishwa Ranjan Dasa, Smrutirekha Sahoob, Chandra Sekhar Pandac, Srikanta Patnaik "Part of speech tagging in odia using support vector machine" nternational Conference on Intelligent Computing, Communication & Convergence (ICCC-2014), Procedia Computer Science 48 (2015) 507 – 512

41. Navanath Saharia, Dhrubajyoti Das, Utpal Sharma, Jugal Kalita "Part of Speech Tagger for Assamese" Text Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 33–36

42. R.Akilan, E.R.Naganathan "POS TAGGING FOR CLASSICAL TAMIL TEXTS" International Journal of Business Intelligent volume: 1 No: 01 2012

43. https://www.aclweb.org/anthology/L18-1

44. Dhanalakshmi V, Anand kumar M, Rajendran S, Soman K P"POS Tagger and Chunker for Tamil Language"Proceedings of the 8th Tamil Internet Conference, Cologne, Germany (2009)

45. https://www.researchgate.net/publication/262624311_Morpheme_based_Language_Model_for_Tamil_Part-of-Speech_Tagging

46. G.Sindhiya, Binulal, P., Anand Goud, K.P.Soman "A SVM based approach to Telugu Parts Of Speech Tagging using SVMTool" International Journal of Recent Trends in Engineering, Vol. 1, No. 2. 2009

47. Phani Gadde, Meher Vijay Yeleti "Improving statistical POS tagging using Linguistic feature for Hindi and Telugu" ICON-2008: International Conference on Natural Language Processing (ICON-2008)

48. Srinivasu Badugu "Morphology Based POS Tagging on Telugu" International Journal of Computer Science Issues, Vol. 11, Issue 1, No 1. 2014

49. Rajeev R R , Jisha P Jayan , Dr. Elizabeth Serly "Tagging Malayalam Text with Parts of Speech -TnT and SVM Tagger Comparison" Proc. of Int. Conf. on Advances in Computer Science 2010

50. Bindu.M.S , Sumam Mary Idicula "High Order Conditional Random Field Based Part of Speech Taggar and Ambiguity Resolver for Malayalam -a Highly Agglutinative Language" International Journal of Advanced Research in Computer Science Volume 2, No. 5. 2011

51. D. Muhammad Noorul Mubarak, Sareesh Madhu, S A Shanavas "A NEW APPROACH TO PARTS OF SPEECH TAGGING IN MALAYALAM" International Journal of Computer Science & Information Technology (IJCSIT) Vol 7, No 5. 2015

52. Ajees A P, Sumam Mary Idicula "A POS Tagger for Malayalam using Conditional Random Fields" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018)

53. Jisha P Jayan, Rajeev R R "Parts Of Speech Tagger and Chunker for Malayalam-Statistical Approach" Computer Engineering and Intelligent Systems Vol 2, No.3

54. Shambhavi.B., R Ramakanth Kumar, P Revanth G "A Maximum Entropy Approach to Kannada Part Of Speech Tagging" International Journal of Computer Applications (0975 – 8887) Volume 41– No.13. 2012

55. Shambhavi B R,,Ramakanth Kumar P "Kannada Part-Of-Speech Tagging with Probabilistic Classifiers"International Journal of Computer Applications (0975 – 888) Volume 48– No.17. 2012

56. M. C. PADMA,R. J. PRATHIBHA "MORPHEME BASED PARTS OF SPEECH TAGGER FOR KANNADA LANGUAGE"International Journal of Management and Applied Science, Volume-2, Issue-7. 2016

57. Shriya Atmakuri, Bhavya Shahi, Ashwath Rao B, Muralikrishna SN "A comparison of features for POS tagging in Kannada" International Journal of Engineering & Technology, 7 (4) 2418-2421. (2018)

58. https://www.researchgate.net/publication/295908477

