Part Of Speech Tagger for Low Resource Indian Language Using Machine Learning Approach

[1] Vikas Verma, [2] S.K.Sharma

[1]Research Scholar, [2]Associate Professor
[1,2] Department of Computer Science and Applications, DAV University, Jalandhar, India
[1] vikas2005verma@yahoo.co.in, [2] sanju3916@rediffmail.com

ABSTRACT

In Language Processing, Part of Speech tagger is one of the fundamental components that are used as a preprocessor for a number of natural language processing tools. For every language before developing the advance tools, POS tagger is developed at the early stage. Various approaches are used for the development of POS tagger. In this research article, a comparative analysis of various Punjabi POS taggers developed by various researchers has been provided and an architecture using an efficient Machine Learning technique is proposed to enhance the accuracy of POS tagger. As all the researchers have used their own test data and not all the developed POS taggers are available online, therefore it is not feasible to test all the POS taggers on common test data set. The claimed results show that POS tagger developed using hybrid approach performs better as compare to rule based technique and other statistic techniques like N-gram, bigram and HMM.

Keywords: Ambiguity, Part of speech, POS, Punjabi, Rule based approach, Statistical approach, Machine Learning,NLP.

INTRODUCTION

Part of speech tagging sometimes also called word sense disambiguation is the basic task performed in almost all the natural language engineering applications. In most of the natural language processing applications, part of speech tagger is the pre-processing activity. Some of the most common natural language applications that comprise the POS tagger as the essential component include; Grammar checker, sentence identification, clause boundary identification, phrase chunking etc. The performance of all these applications depends upon the accuracy of POS tagger. Various techniques are used for development of POS tagger. These techniques can be categorized into rule based technique and statistical techniques. In statistical techniques Conditional Random Field (CRF), Hidden Markov Model (HMM), N-gram and Support Vector Machine (SVM) are used. Beside these rule based and statistical techniques, a hybrid approach i.e. combination of two or more than two techniques is also used.

INTRODUCTION TO PUNJABI LANGUAGE

Punjabi is one of the Indo-Aryan languages spoken in a multilingual country i.e. India. Out of 22 language spoken in India, Punjabi is the 11th most spoken language. Punjabi falls in the category of Indo- Aryan languages and is spoken by 130 million people in the world with 113 million native speakers. Most of the Punjabi speakers live in Punjab state of India and Pakistan. Punjabi

language is written in two scripts Gumukhi script (Used in India) and Shahmukhi (used in Pakistan). Moreover there are many Punjabi spoken migrants in Australia, Canada and UK. It is official language of Punjab, additional language of Haryana, Delhi and J&K.

A. Natural Language processing in Punjabi language

Researchers are working on the technical development of Punjabi language. A number of basic and advanced tools for Punjabi language processing have been developed and many are under development. Punjabi University Patiala, Thapar University, C-DAC Mohali and TDIL are some of the Universities and research organizations which are working on the technical development of Punjabi language. Existing tools for Punjabi language processing include machine translation system, grammar checker, phrase chunker, clause boundary identification system, morphological analyzer, summarization system, machine transliteration system and many more. Most of these resources include POS tagger as an essential component. Therefore efforts have been done to improve the accuracy of Punjabi POS tagger by using different approaches. In the following section a brief description of various POS taggers developed by various researchers have been discussed.

VARIOUS TECHNIQUES USED FOR DEVELOPMENT OF POS TAGGER

Part of speech (POS) is performed after applying morphological analyzer (MA) in order to remove disambiguates. It is one of the fundamental components that are developed at the initial stage of every NLP tool. There are basically three techniques used for its development. These techniques includes rule based technique used for Kannad language (Vijayalaxmi .F. Patil, 2010) [12]; for Telugu language (Sreeganesh, 2006) [14]; for Portuguese language (Nogueira Dos Santos et al., 2008) [15]; for English language (Wilson and Heywood, 2005 and Brill, 1995) [16-17]. Further Statistical technique is used for Marathi (Jyoti Singh et al., 2013) [18]; for Hindi language (Nidhi Mishra and Amit Mishra, 2011) [19]; for Bangla Language (Hammad Ali 2010) [20]; for Malayalam (Antony P.J. et al., (2010) [21]; for Hindi (Aniket Dalal et.al., 2006 and Agarwal Himashu et al., 2006) [22-23]; for Bengali (Ekbal and S. Bandyopadhyay, 2008) [24]; for Tamil (V.Dhanalakshmi et al., 2008) [25]; for Telugu language (M Anandkumar et.al., 2008 and Bindhiya Binulal et al., 2009) [26]; for Malayalam language (Antony P.J et al., 2010) [27]; for Hindi language (Shrivastava & Pushpak Bhattacharyya, 2008, Aniket Dalal et al., 2006 and Himanshu A., 2007) [28,33,35]; for Malayalam (Manju K. et al., 2009) [29]; for Assamese (Navanath Saharia et.al., 2009) [30]; for Punjabi (Sharma, S.K. et al., 2011) [31]; for Bengali (Ekbal, S. et al., 2007, Ekbal and Bandyopadhyay, 2008) [32] [34]. Further Neural network based technique has been used by Ankur Parikh (2009) for Hindi [36]; Hybrid based approach is used by Arulmozhi P. et al. (2006) for Tamil [37]; Chirag Patel and Karthik Gali (2008) for Gujarati [38]. Elba et al., (2006). [12] ran tests on various linguistic corpora and compared the results to those of other prominent techniques as well as a standard dynamic programming algorithm and found that their algorithms, as well as some of its components, could be used to represent a new set of cutting-edge processes for complex tagging scenarios.

EXISTING POS TAGGING SYSTEMS FOR PUNJABI LANGUAGE

Various POS tagging system have been proposed/developed by different authors are:

A. Rule based POS tagger

This POS tagger has been developed by (Lehal and Singh, 2009) [1]. This was the first part of speech (POS) tagger developed for Punjabi language. This tagger was developed to be used as an essential component in Punjabi Grammar Checker. The tagset having 630 fine grade tags was used to annotate the text. These 630 tags were used to annotate 12 word classes i.e. Noun, Pronoun (Personal, Reflexive, Demonstrative, Instrumental, Relative, Interrogative), Adjective (Inflected and Un-inflected), Ordinal, Cardinal, main verb, Auxiliary verb, Adverb, Post position (Inflected and Un-inflected), Conjunction, Particle (Un-inflected and Vocative) and Verb part. This POS tagger was reported to give accuracy of 80.29% including unknown words and of 88.86% excluding unknown words. This POS tagger is online available at http://punjabi.aglsoft.com/?show=tagger [41] and http://pgc.learnpunjabi.org/#Tagger [42].

B. HMM based POS tagger

(Sharma and Gill, 2011) [2] used Hidden Markov Model (HMM) based technique to develop statistical POS tagger. Viterby algorithm was used to implement the concept of HMM. Author used bigram model and used a corpus of 8 million words to train the system. The emission and transition probabilities were used by Maximum likelihood method. The author used two datasets containing 10000 words each to test the system and reported an accuracy of 84.9% and 87.6% respectively for each set.

(Kanwar et al., 2011) [8] sought to improve the existing Punjabi POS tagger's accuracy. The ambiguity of compound and complex statements was not resolved by this POS tagger. The challenge of part of speech tagging was solved using a Bi-gram Hidden Markov Model. The HMM parameter was trained and estimated using an annotated corpus. The parameter was estimated using the maximum likelihood technique. The Viterby algorithm was used to implement this HMM approach. A corpus of writings from various genres was used to evaluate the suggested tagger. The results were manually analyzed to determine which tag assignments were correct and which were wrong. A total of 20,000 words were selected at random from a corpus of 4 million Punjabi words and manually categorized into two categories having accuracies of 84.9% and 87.6% respectively.

C. N-gram based POS tagger

This technique has been used by (Sumeer mittal et al., 2014) [3] for development of statistical Punjabi POS tagger. Author used bi-gram model along with TDIL (Technical development of Indian Languages) proposed tagset. The corpus used was collected from various online resources like Punjabi e-paper, Punjabi stories etc. the corpus was annotated with TDIL proposed tagset using semi-automated approach. Bi-gram probabilities were calculated using an annotated corpus. System was tested on a test data containing 2400 sentences having approximately 10000 words

andoutput an accuracy of 92.16% (Excluding Unknown words).

D. Reduced tagset POS tagger

In order to further improve the accuracy of HMM POS tagger (Manjit kaur et al., 2014) [4] used small size tagset. Author observed that the previously developed HMM based Punjabi POS tagger used a tagset containing more than 630 tags and the main problem with this tagset was data sparseness. This problem of data sparseness can be reduced either by using a very large amount of corpus for training or by using a tagset have small number of tags. Therefore in this research author used a reduced tagset containing 36 tags (Proposed by TDIL). Result showed a significant improvement by increasing the accuracy to 92-95% as compare to 85-87% (reported in previously developed HMM POS tagger).

E. SVM based POS tagger

(Kumar and Josan , 2016) [9] proposed SVM based POS tagging system for Punjabi language. Author developed his own tagset for machine learning approach containing 38 tags. A corpus containing 27000 words was collected from online sources and was used for training and testing purpose. A rule based method and decision tree based methods were used for identification of unknown words. The author claimed an accuracy of 89.86% including unknown and ambiguous words.

F. Bigram POS tagger

(Sanyam Sood et al., 2014) [7] proposed a method to assign the POS tags to unknown words and ambiguous Punjabi words. Author used a bi-gram model for this task and used the tagset proposed by TDIL. This tagger equipped with unknown tag guesser component shows an accuracy of 92-94%. Further author clarified that introduction of unknowntag guesser component play major role in improving the system. The unknown word tag guesser component also gavean accuracy of 88-94%.

G. Neural network based POS tagger

(Kashyap and Josan, 2013) [5] presented a novel approach of POS tagging of Punjabi text using neural networks. Author proposed a multi-layer perceptron neural network tagger with fixed context for tagging of Punjabi text. For learning, author used back propagation learning algorithm. A randomize function was used to divide the corpus into training and testing data. By using the neighboring context of current word in the training data a feature vector was generated. Trigram model was used for generating this feature vector. Author claimed an accuracy of 88.95%.

H. Hybrid POS tagger

(Sharma and Lehal, 2011) [2] used HMM in combination with rules to improve the existing rule based POS tagger. To implement the HMM author used an annotated corpus of 20,000 words.

From this annotated corpus, transition, emission and initial parameters were calculated. A rule based tagger followed by HMM based tagged was used to implement hybrid approach. This module was tested on the corpus containing 26,479 words. An accuracy of 90.11% was observed using manual approach.

I. GA based POS tagger

(Kamaljot Singh, 2015) [6] used Genetic Algorithm for POS tagging of Punjabi language. Author used training set of 20k words and testing set of 6k words. For the comparison with existing Taggers, author used the Gene_SIZE 4, Population_SIZE 64, Generation_COUNT 20, Mutation_RATE 10%. Author claimed an accuracy of 90.63%. It had been observed by the author that the sentence level tagging worked better than the word level tagging methods.

J. Punjabi POS tagger Rule based and HMM

(Singh and Goyal, 2017) [39] developed Punjabi POS tagger using two different approached i.e. rule based approach and HMM based model. A tagset proposed by Technical development of Indian languages (TDIL) having 35 tags was used for POS tagger development. In case of rule based tagger authors used 150 rules for tagging unknown words as well as to resolve the ambiguity of tags. For training HMM model, authors used a data set of 49 thousand words from tourism domain and health domain. This training data was collected from Indian Language Corpora Initiative (ILCI) corpus available at http://sanskrit.jnu.ac.in/ilci/index.jsp [43]. On testing both the systems using two test sets authors observed F1 score 0.922 and 0.924 in case of rule based system and 0.927 and 0.933 for HMM based system. This POS tagger is available online at http://punjabipos.learnpunjabi.org/ [40].

Apart from formulating a POS tagger, few researchers performed innovativeness by making changes in tagset of Punjabi like (Kumar and Josan, 2012) [10] created a POS tagset for capturing morphosyntactic aspects of Punjabi using coarse-grained granularity and devised the various tags for the proposed tagset. After then, the proposed tagset was compared to the Indian Languages existing tagsets. Their tagset features 38 tags in comparison to Indian Language tagsets, notably Punjabi tagsets and visualized that he suggested tagset would be utilised to create a Punjabi POS tagger based on machine learning. However, such alterations are beyond the scope of our current discusions.

COMPARATIVE ANALYSIS OF VARIOUS PUNJABI POS TAGGERS

As discussed above and as per existing literature till date, there are ten different types of part of speech taggers that have been developed for Punjabi language. Out of these nine, four part of speech taggers [1],[2],[3] and [5] are developed for large tagset proposed by (Gill and Lehal, 2009) [1] (to be used as a component of Punjabi Grammar Checker) i.e. tagset having more than 630 fine grade tags. Further four POS taggers are developed for small tagset proposed by TDIL (Technical Development of Indian Languages) [11] i.e. tagset having 36 tags, One tagger is

developed for tagset having 38 tags [6]. In table 1, comparative analysis of all the ten POS taggers irrespective of thetagset used is provided. In table 2,

comparative analysis of POS tagger developed for tagset having more than 630 tags is provided and further in the table 3, comparative analysis of POS tagger developed for tagset having 35, 36 or 38 tags is provided

Sr. No.	Technique used	% age Accuracy
1.	excel	80.91
2.	HMM based	84.90
3.	N-gram based	92.16
4.	HMM with Reduced tagset	93.5
5	SVM based	89.86
6.	Bi-gram based	93
7.	Hybrid	90.11
8.	Neural Network based	88.95
9.	GA based	90.63
10.	HMM and Rule based	F1-score 0.933 on Tourism domain and 0.927 on Health domain

Table (1)- Comparative analysis of various Punjabi POS taggers

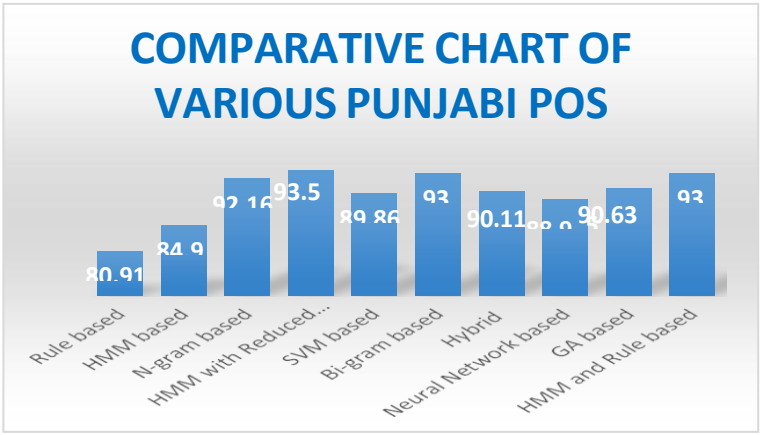

Figure – 1. Comparative chart of various Punjabi POS taggers

Sr. No.	Technique used	Tagset Size	% age Accuracy
1.	Rule based	630	80.91
2.	HMM based	630	84.90
3.	Hybrid	630	90.11
4.	Neural network based	630	88.95

Table (2) - Comparative analysis of Punjabi POS tagger with same tagset having 630 tags

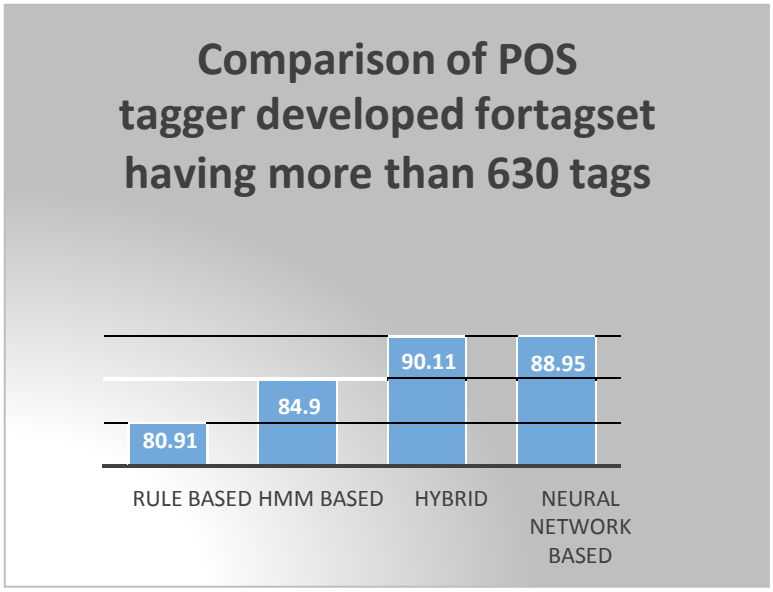

Figure 2 - Comparative chart of various Punjabi POS taggers developed for tagset having 630 tags

Sr. No.	Technique used	Tagset Size	% age Accuracy
1	N-gram based	**36**	92.16

2	HMM with Reduced tagset	**36**	93.5
3	Bi-gram based	**36**	93
4.	GA	**38**	90.63
5	SVM	**38**	89.86
6	HMM and Rulebased	**35**	93

Table (3) - Comparative analysis of Punjabi POS tagger with same tagset having 36 tags

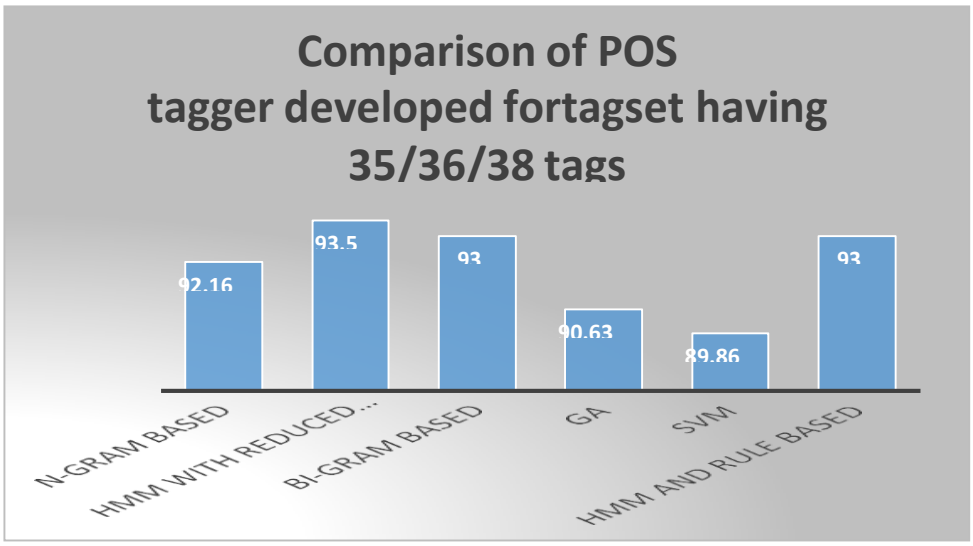

Figure 3: Comparative chart of various Punjabi POS taggers developed for tagset having 36 tags

III. ARCHITECTURE OF AN EFFICIENT MACHINE LEARNING POS TAGGER FORPUNJABI

Research community working in the domain of NLP has been working extensively in improving the efficiency of POStagger so as to ultimately improve the overall efficiency of system in terms of false alarms particularly for grammar checkers.

We are motivated to use “Hybrid” model involving “Machine Learning” techniques keeping in consideration the workdone in other Indian scheduled and Indo-Aryan languages. This approach is novel one as till date research has not been processed for Punjabi language using such techniques. Work done for other languages is summarized as under.

(Todi et al., 2018) [44] developed a Kannada POS Tagger Using Machine Learning and Neural Network Models. The authors used a window feature of [-2,+2] to get an F1-score of 0.92 in Conditional Random Field model, and in the neural network model, they employed character embeddings combined with pre-trained word embeddings to get an F1-score of 0.92. They further concluded that the neural networks' accuracy can be increased by training them on a larger dataset, and the word embeddings employed could also be trained on a larger dataset to reduce Out-of- Vocabulary words. Furthermore, they suggested that given annotated data, the neural network model could also be used for chunking and at last stated, in case, the correct word POS is provided, chunking models tend to be more accurate.

(Sayami et al., 2019) [45] compared and implemented various deep learning-based POS taggers for Nepali, including the Simple Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bi-directional Long Short Term Memory (Bi-LSTM). In a corpus of Nepali tag sets from a corpus of tag size 40, several techniques were trained and tested by them. The authors concluded that Bi-directional LSTM outperformed the other three techniques as simple RNN, LSTM, GRU, and Bi-directional LSTM all had accuracy of 96.84 percent, 96.48 percent, 96.86 percent, and 97.27 percent, respectively. Further, they proposed that to improve efficiency, the vocabulary size and tag set could be expanded and stated that reinforcement learning could be used to improve trainingefficiency.

(Kumar et al., 2019) [46] used Deep learning sequential models to tag Malayalam Twitter data by created a tagset with 17 coarse tags. 9915 tweets were manually annotated (85404 tokens). Sequential deep learning methods such as Recurrent Neural Network (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Bidirectional LSTM were used to analyse the tagged data (BLSTM). The model was trained on a word-by-word and character-by-character basis. The authors concluded that the GRU-based deep learning sequential model had the greatest f1-measure of 0.9254 at the word level, and the BLSTM-based deep learning sequential model had the highestf1-measure of 0.8739 at the character level. They experimented with different numbers of hidden states, such as 4, 16,32, and 64, and executed the training for each and discovered that increasing the number of hidden states enhances the tagger model.

(Prabha et al., 2018) [47] proposed a Deep learning-based POS tagger with 43 tags for Nepali text, with Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), Gated Recurrent Unit (GRU), and its bidirectional variations. The model was evaluated using performance criteria such as accuracy, precision, recall, and F1-score. With binary cross entropy as the loss function, bi-directional versions of RNN, LSTM, and GRU achieved the highest performance ratings. The system's accuracy improved as the size of the word embedding vector grew larger. The results revealed that the author’s algorithm improved significantly and exceeded state-of-the-art POS taggers with greater than 99 percent accuracy.

To conduct POS tagging for Marathi text, (Deshmukh and Kiwelekar, 2020) [48] presented a Deep learning model with a bidirectional long short-term memory (Bi-LSTM) model. A 1500-

sentence POS-tagged corpus from a Marathi e-Newspaper was produced, with 32 tags. Each phrase was of 100 words long. Every word was represented by a 100-point vector. A total of 80% of the dataset was used for training and 20% for testing. Validation uses 20% of the data collected during training. 'Categorical Cross Entropy' was the loss function utilised. 'RMSprop' was the optimizer that was employed. The deep learning model was 85 percent accurate, whereas the Bi-LSTM model was 97 percent accurate. Creating a deep learning model, building the Bi-LSTM model, and comparing machine learning techniques for the same dataset were the efforts of authors. To enhance efficiency they also suggested useage a large POS-taggedcorpus.

With the above stated notion in context of Indian languages, we hereby propose state-of-the-art, novel model for Punjabi POS Tagger using Machine Learning concept integrated with existing approach. The proposed algorithm is given below:

Proposed Algorithm (for POS tagging):

i. NLTK (natural language Tool kit i.e. one of the python library is used)
ii. Two arrays i.e. one is tagged sentence array and second is sentence array is used.
iii. Annotated training file is opened using UTF-8 encoding
iv. Each sentence from the training file is extracted in sentence array.
v. Total Number of sentences in training.
vi. Total Number of words in training.
vii. Numpy package is imported
viii. From the training sentences (tagged sentences) tag and words are separated.
ix. Total data (words and tags) is split into train and test data
x. Indexing of word and tags is done. So the words and tags are converted to numeric as the input cannot betext.

IV. ACKNOWLEDGEMENT

We'd want to express our gratitude to DAV University for making the Punjabi corpus available to researchers. We are grateful to TDIL, ILCI, Punjabi University, and the Government of India for taking the initiative to create resources for Indian languages and making them available to the NLP community.

V. CONCLUSION AND FUTURE WORK

From table 1, table 2 and table3 it can be concluded that part of speech tagger developed using hybrid approach performs best when a tagset containing more than 630 fine grade tags is used. When a small tagset containing 36 or 38 tags is used then HMM perform better as compare to others. Also, that smaller the tagset better will be the performance of statistical POS tagger.

Further, to improve the efficiency and resolution of “False Alarm”, a “Hybrid” model is proposed which uses the basic functionality of traditional approach in combination with Machine Learning

method. The proposed architecture can be implemented using languages like Python by importing suitable packages and libraries. The following metrics can be utilised to evaluate the experiments: precision, recall, f1-measure, and accuracy.

REFERENCES

[1]. Gill, M. S., Lehal, G. S., & Joshi, S. S. (2009). Part of speech tagging for grammar checking of Punjabi. *The Linguistic Journal*, *4*(1), 6-21.

[2]. Sharma, S. K., & Lehal, G. S. (2011, June). Using Hidden Markov Model to improve the accuracy of Punjabi POS tagger. In *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on* (Vol. 2, pp. 697-701). IEEE.

[3]. Mittal, S., Sethi, N. S., & Sharma, S. K. (2014). Part of Speech Tagging of Punjabi Language using N Gram Model. *International Journal of Computer Applications*, *100*(19).

[4]. Kaur, M., Aggerwal, M., & Sharma, S. K. (2014). Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set. *International Journal of Computer Applications & Information Technology*, *7*(2), 142.

[5]. Kashyap, D. K., & Josan, G. S. (2013, October). A trigram language model to predict part of speech tags using neural network. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 513-520). Springer, Berlin, Heidelberg.

[6]. Singh, K. (2015). Part-of-Speech Tagging using Genetic Algorithms. *International Journal of Simulation-- Systems, Science & Technology*, *16*(6).

[7]. Sood, S., Arora, V., & Sharma, S. K. (2014). Word Class Prediction of Ambiguous and Unknown Words of Punjabi Language Using Bi-gram Methods. *International Journal of Computer Applications & Information Technology*, *7*(2), 152.

[8]. Kanwar S.,Ravishankar, Sharma, S.K. (2011) POS tagging of Punjabi language Using Hidden Markov Model. Research Cell: International Journal of Engineering Sciences. pp 98-106.

[9]. Kumar, D., & Josan, G. (2016). Prediction of Part of Speech Tags for Punjabi using Support Vector Machines. *International Arab Journal of Information Technology (IAJIT)*, *13*(6).

[10]. Kumar D. and Josan G., "Developing a tagset for machine learning based POS tagging in Punjabi," international Journal of Applied Research on Information Technology and Computing, vol. 3, no. 2, pp. 132-143, 2012.

[11]. http://tdildc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf (Accessed on Oct 5, 2021).

[12]. Vijayalaxmi .F. Patil (2010), "Designing POS Tagset for Kannada, Linguistic Data Consortium for Indian Languages (LDC-IL), Organized by Central Institute of Indian Languages, Department of Higher Education Ministry of Human Resource Development, Government of India, March 2010.

[13]. E. Alba, G. Luque, L. Araujo, Natural language tagging with genetic algorithms, Information Processing Letters 100 (5) (2006) pp. 173 – 182.

[14]. Sreeganesh, T. (2006). Telugu parts of speech tagging in WSD. Language of India, 6.
[15]. Milidiú, R. L., Santos, C. N., & Duarte, J. C. (2008). Phrase chunking using entropy guided transformation learning. Proceedings of ACL-08: HLT, 647-655.
[16]. Wilson, G., & Heywood, M. (2005, June). Use of a genetic algorithm in brill's transformation-based part-of-speech tagger. In Proceedings of the 7th annual conference on Genetic and evolutionary computation (pp. 2067-2073). ACM.
[17]. E. Brill, "Some advances in rule based part of speech tagging", In Proceedings of The Twelfth 5ational Conference on Artificial Intelligence (AAAI94), Seattle, Washington, 1994.
[18]. Singh, J., Joshi, N., & Mathur, I. (2013, August). Development of Marathi part of speech tagger using statistical approach. In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on (pp. 1554-1559). IEEE.
[19]. Mishra, N., & Mishra, A. (2011, June). Part of speech tagging for Hindi corpus. In Communication Systems and Network Technologies (CSNT), 2011 International Conference on (pp. 554-558). IEEE.
[20]. Ali, H. (2010). An unsupervised parts-of-speech tagger for the bangla language. Department of Computer Science, University of British Columbia, 20, 1-8.
[21]. Antony, P. J., Mohan, S. P., & Soman, K. P. (2010, March). SVM based part of speech tagger for Malayalam. In Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conferenceon (pp. 339-341). IEEE.
[22]. Dalal, A., Nagaraj, K., Sawant, U., & Shelke, S. (2006). Hindi part-of-speech tagging and chunking: A maximum entropy approach. Proceeding of the NLPAI Machine Learning Competition.
[23]. Agarwal, Himashu., and Mani,A. (2006), Part of Speech Tagging and Chunking with Conditional Random Fields. In the proceedings of NLPAI Contest, 2006.
[24]. Ekbal, A., & Bandyopadhyay, S. (2008). Web-based Bengali news corpus for lexicon development and POStagging. Polibits, (37), 21-30.
[25]. V Dhanalakshmi, M Anandkumar, MS Vijaya, R Loganathan, KP Soman, and S Rajendran. 2008. Tamil part-of-speech tagger based on svmtool. In Proceedings of the COLIPS International Conference on natural language processing (IALP), Chiang Mai, Thailand
[26]. Binulal, G. S., Goud, P. A., & Soman, K. P. (2009). A SVM based approach to Telugu parts of speech tagging using SVMTool. International Journal of Recent Trends in Engineering, 1(2), 183.
[27]. Antony, P. J., Mohan, S. P., & Soman, K. P. (2010, March). SVM based part of speech tagger for Malayalam. In Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conferenceon (pp. 339-341). IEEE.
[28]. Shrivastava, M., & Bhattacharyya, P. (2008, December). Hindi pos tagger using naive

stemming: Harnessingmorphological information without extensive linguistic knowledge. In International Conference on NLP (ICON08), Pune, India.

[29]. Manju, K., Soumya, S., & Idicula, S. M. (2009, October). Development of a POS tagger for Malayalam-an experience. In Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on (pp. 709-713). IEEE.

[30]. Saharia, N., Das, D., Sharma, U., & Kalita, J. (2009, August). Part of speech tagger for Assamese text. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 33-36). Association for Computational Linguistics.

[31]. Sharma, S. K., & Lehal, G. S. (2011, June). Using hidden markov model to improve the accuracy of punjabi pos tagger. In Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conferenceon (Vol. 2, pp. 697-701). IEEE.

[32]. Ekbal, A., Mondal, S., & Bandyopadhyay, S. (2007). POS Tagging using HMM and Rule-based Chunking.The Proceedings of SPSAL, 8(1), 25-28.

[33]. Dalal, A., Nagaraj, K., Sawant, U., & Shelke, S. (2006). Hindi part-of-speech tagging and chunking: A maximum entropy approach. Proceeding of the NLPAI Machine Learning Competition.

[34]. Ekbal, A., & Bandyopadhyay, S. (2008). Web-based Bengali news corpus for lexicon development and POStagging. Polibits, (37), 21-30.

[35]. Agrawal, H. (2007). POS tagging and chunking for Indian languages. Shallow Parsing for South Asian Languages, 37.

[36]. Parikh, A. (2009). Part-of-speech tagging using neural network. Proceedings of ICON.

[37]. Arulmozhi, P., & Sobha, L. (2006). A Hybrid POS Tagger for a Relatively Free Word Order Language. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages (pp. 79-85).

[38]. Patel, C., & Gali, K. (2008). Part-of-speech tagging for Gujarati using conditional random fields. InProceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

[39]. Singh, Umrinder and Goyal, Vishal (2017). Punjabi POS tagger: Rule Based and HMM. International journalof computer science and software Engineering.

[40]. http://punjabipos.learnpunjabi.org/ (Accessed on Oct 5, 2021).

[41]. http://punjabi.aglsoft.com/punjabi/?show=tagger (Accessed on Oct 5, 2021).

[42]. http://pgc.learnpunjabi.org/#Tagger (Accessed on Oct 5, 2021).

[43]. http://sanskrit.jnu.ac.in/ilci/index.jsp. (Accessed on Oct 5, 2021).

[44]. Todi, K. K., Mishra, P., & Sharma, D. M. (2018). Building a kannada pos tagger using machine learning andneural network models. *arXiv preprint arXiv:1808.03175*.

[45]. Sayami, S., Shahi, T. B., & Shakya, S. (2019). *Nepali POS Tagging Using Deep Learning Approaches* (No. 2073). EasyChair.

[46]. Kumar, S., Kumar, M. A., & Soman, K. P. (2019). Deep learning based part-of-speech tagging for Malayalam Twitter data (Special issue: deep learning techniques for natural language processing). *Journal of IntelligentSystems*, *28*(3), 423-435.

[47]. Prabha, G., Jyothsna, P. V., Shahina, K. K., Premjith, B., & Soman, K. P. (2018, September). A deep learning approach for part-of-speech tagging in nepali language. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1132-1136). IEEE.

[48]. Deshmukh, R. D., & Kiwelekar, A. (2020, March). Deep learning techniques for part of speech tagging by natural language processing. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 76-81). IEEE.