Parts Of Speech Tagging for Indian Languages: A Literature Survey

Antony P J
Research Scholar
Computational Engineering and Networking (CEN),
Research Centre,
Amrita Vishwa Vidyapeetham
University, Coimbatore, India

Dr. Soman K P
Professor and Head
Computational Engineering and Networking (CEN),
Research Centre,
Amrita Vishwa Vidyapeetham
University, Coimbatore, India

ABSTRACT
Part of speech (POS) tagging is the process of assigning the part of speech tag or other lexical class marker to each and every word in a sentence. In many Natural Language Processing applications such as word sense disambiguation, information retrieval, information processing, parsing, question answering, and machine translation, POS tagging is considered as the one of the basic necessary tool. Identifying the ambiguities in language lexical items is the challenging objective in the process of developing an efficient and accurate POS Tagger. Literature survey shows that, for Indian languages, POS taggers were developed only in Hindi, Bengali, Panjabi and Dravidian languages. Some POS taggers were also developed generic to the Hindi, Bengali and Telugu languages. All proposed POS taggers were based on different Tagset, developed by different organization and individuals. This paper addresses the various developments in POS-taggers and POS-tagset for Indian language, which is very essential computational linguistic tool needed for many natural language processing (NLP) applications.

Keywords
Ambiguity, Tagset, Information Retrieval, Data Driven System, Foreign Languages

1. INTRODUCTION
Part of speech tagging (POS tagging) has a crucial role in different fields of natural language processing (NLP) including machine translation. Parts-of-speech-tagging is defined as the process of assigning to each word in a sentence, a label which indicates the status of that word within some system of categorizing the words of that language according to their morphological and/or syntactic properties. A part-of-speech is a grammatical category, commonly including verbs, nouns, adjectives, adverbs, determiner, and so on.

Some classic examples for POS Taggers available in English are Bril tagger, Tree tagger, CLAWS tagger, online tagger ENGTWOL. Most of these methods used rule based, stochastic or morphological inputs. The Fig. 1 shows the development of various corpus and POS taggers using different approaches.

The analysis of languages is a complex procedure in India. In Indian languages, most of natural language processing work has been done in Hindi, Tamil, Malayalam and Marathi. These languages have several part-of-speech taggers that use different mechanisms. Research on part-of-speech tagging has been closely tied to corpus linguistics. Earlier work in POS tagging for Indian languages was mainly based on rule based approaches. But the fact that rule-based method requires expert linguistic knowledge and hand written rule. Due to the morphological richness of Indian languages, researchers faced a great difficulty to write complex linguistic rules and the rule based approach did not result well in many cases. Later, researchers shifted to stochastic and other approaches and developed some better POS taggers in various Indian languages. Even though stochastic methods need very large corpora to be effective, many successful POS were developed and used in various natural language processing tasks for Indian language.

In most of the Indian languages, the ambiguity is the key issue that must be addressed and solved while designing a pos tagger. For different context words behave differently and hence the challenge is to correctly identify the POS tag of a token appearing in a particular context. This paper gives a survey on developments of various POS taggers in Indian languages. The following sections of this chapter are organized as follow. The first section gives a brief description about various attempts in POS taggers in Indian languages. The second section is about the different Tagset developed for Indian languages.

2. POS TAGGING APPROACHES
POS taggers are broadly classified into three categories called rule based, Empirical based and Hybrid based .In case of rule based approach hand-written rules are used to distinguish the tag ambiguity. The empirical POS taggers are further classified into Example based and Stochastic based taggers. Stochastic taggers are either HMM based, choosing the tag sequence which maximizes the product of word likelihood and tag sequence probability, or cue-based, using decision trees or maximum entropy models to combine probabilistic features. The stochastic taggers are further classified in to supervised and unsupervised taggers. Each of these supervised and unsupervised taggers are categorized into different groups based on the particular algorithm used. The Fig. 2 shows the classification of parts of speech approaches.

2.1 Rule Based POS tagging
The rule based POS tagging models apply a set of hand written rules and use contextual information to assign POS tags to words. These rules are often known as context frame rules. For example, a context frame rule might say something like: *"If an ambiguous/unknown word X is preceded by a Determiner and*

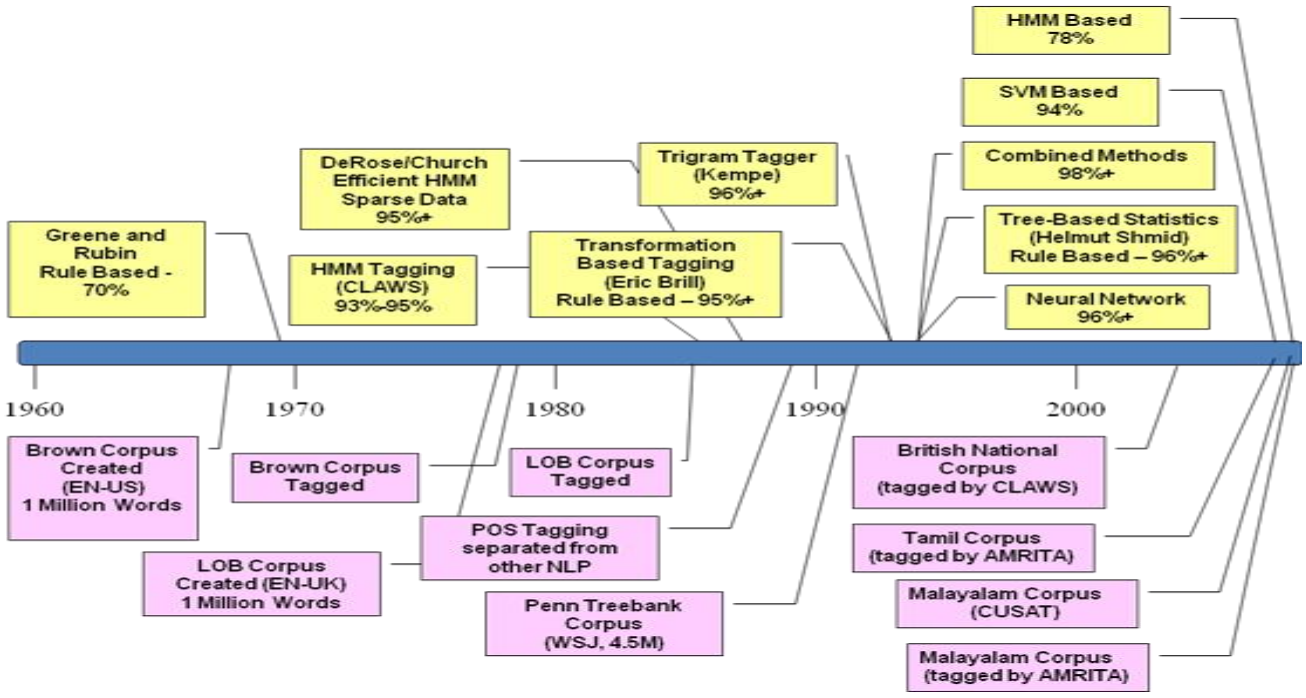

Fig 1: Various Corpus and POS taggers

followed by a Noun, tag it as an Adjective". One of the first and widely used English POS-taggers employs rule based algorithms is "Brill's tagger".

The earliest algorithms for automatically assigning part-of-speech were based on two-stage architecture. The first stage used a dictionary to assign each word a list of potential parts of speech. The second stage used large lists of hand-written disambiguation rules to bring down this list to a single part-of-speech for each word. The ENGTWOL tagger is based on the same two-stage architecture, although both the lexicon and the disambiguation rules are much more sophisticated than the early algorithms.

2.2 Empirical Based POS tagging

The relative failure of rule-based approaches, the increasing availability of machine readable text and the increase in capability of hardware (CPU, memory, disk space) with decrease in cost are some of the reasons, researchers to prefer corpus based pos tagging. The empirical approach of parts speech tagging is further divided in to two categories: Example-based approach and Stochastic based approach. Literature shows that majority of the developed POS taggers belongs to empirical based approach.

2.2.1 Example-Based techniques

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

2.2.2 Stochastic based POS tagging

The stochastic approach finds out the most frequently used tag for a specific word in the annotated training data and uses this information to tag that word in the unannotated text. A stochastic approach required a sufficient large sized corpus and calculates frequency, probability or statistics of each and every word in the corpus. The problem with this approach is that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules of a language.

The use of probabilities in tags is quite old; probabilities in tagging were first used in 1965, a complete probabilistic tagger with Viterbi decoding was sketched by Bahl and Mercer (1976), and various stochastic taggers were built in the 1980's (Marshall, 1983; Garside, 1987; Church, 1988; DeRose, 1988).

Supervised and unsupervised are two broad categories of stochastic based approach.

Supervised POS tagging: The supervised POS tagging models require pre-tagged corpora which are used for training to learn information about the tagset, word-tag frequencies, rule sets etc. The performance of the models generally increases with the increase in size of this corpus. The following are the two familiar examples for supervised POS taggers.

Hidden Markov Model (HMM) based POS tagging: An alternative to the word frequency approach is known as the n-gram approach that calculates the probability of a given sequence of tags. It determines the best tag for a word by calculating the probability that it occurs with the n previous tags, where the value of n is set to 1, 2 or 3 for practical purposes. These are known as the Unigram, Bigram and Trigram models. The most common algorithm for implementing an n-gram approach for tagging new text is known as the HMM's Viterbi Algorithm. The Viterbi algorithm is a search algorithm that avoids the polynomial expansion of a breadth first search by trimming the search tree at each level using the best 'm' Maximum Likelihood Estimates (MLE) where 'm' represents the number of tags of the following word. For a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes as in formula 1:

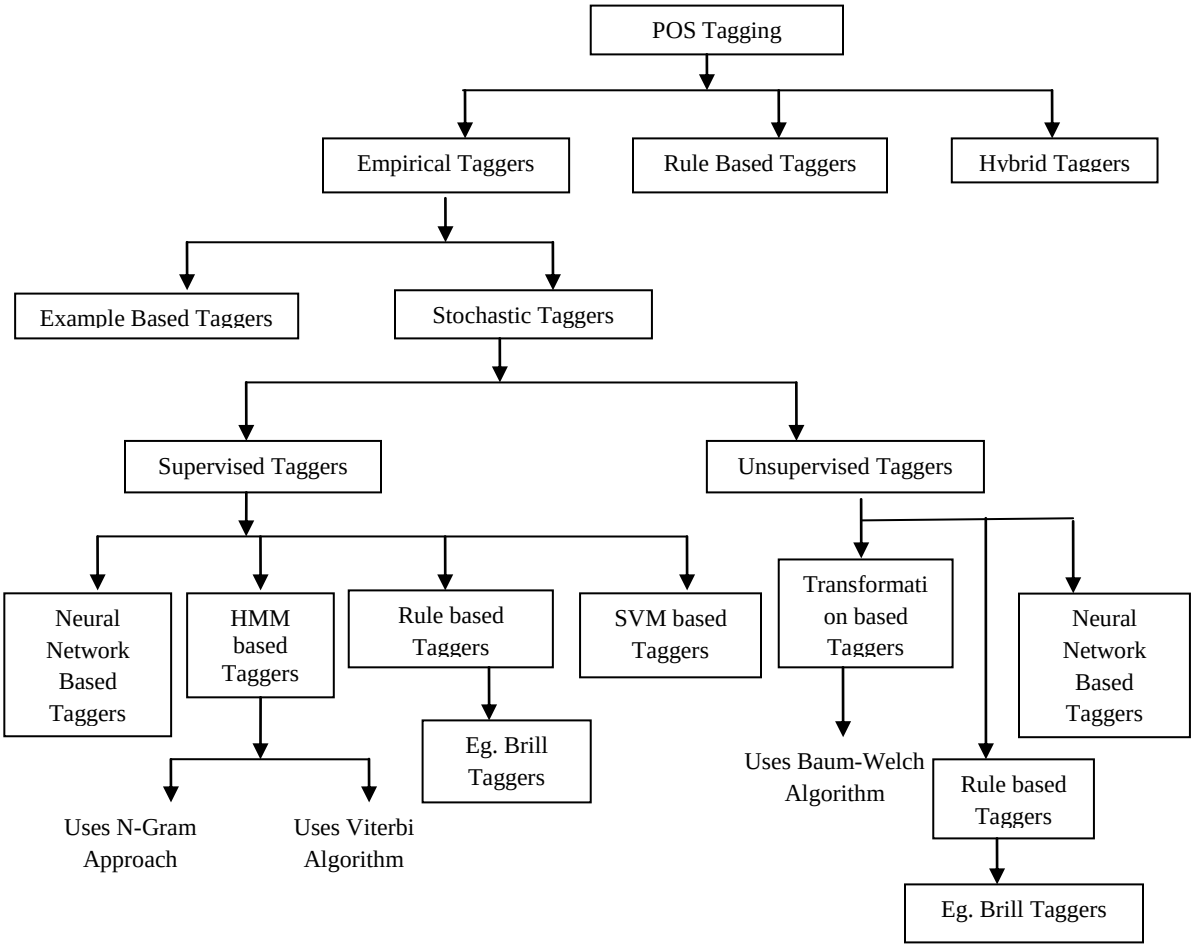

Fig 2: Classification of POS tagging Approaches

$$P(\text{word} \mid \text{tag}) \text{ X } P(\text{tag} \mid \text{previous n tags}) \quad (1)$$

A bigram-HMM tagger of this kind chooses the tag t_i for word w_i that is most probable given the previous tag t_{i-1} and the current word w_i:

$$t_i = \arg\max_j P(t_j \mid t_{i-1}, w_i) \quad (2)$$

Support Vector Machines: SVM is a machine learning algorithm for binary classification, which has been successfully applied to a number of practical problems, including NLP. Let $\{(x_1, y_1) \ldots (x_N, y_N)\}$ be the set of N training examples, where each instance x_i is a vector in $\mathbf{R}^N$ and $y_i \in \{-1,+1\}$ is the class label. In their basic form, a SVM learns a linear hyperplane, that separates the set of positive examples from the set of negative examples with maximal margin (the margin is defined as the distance of the hyperplane to the nearest of the positive and negative examples). This learning bias has proved to have good in terms of generalization bounds for the induced classifiers.

The SVMTool is intended to comply with all the requirements of modern NLP technology, by combining simplicity, flexibility, robustness, portability and efficiency with state–of–the–art accuracy. This is achieved by working in the Support Vector Machines (SVM) learning framework, and by offering NLP researchers a highly customizable sequential tagger generator.

Unsupervised POS Tagging: Unlike the supervised models, the unsupervised POS tagging models do not require a pre-tagged corpus. Instead, they use advanced computational methods like the Baum-Welch algorithm to automatically induce tagsets, transformation rules etc. Based on the information, they either calculate the probabilistic information needed by the stochastic taggers or induce the contextual rules needed by rule-based systems or transformation based systems.

2.2.3 Transformation-based POS tagging

In general, the supervised tagging approach usually requires large sized pre-annotated corpora for training, which is difficult for most of the cases. But recently, good amount of work has been done to automatically induce the transformation rules. One approach to automatic rule induction is to run an untagged text through a tagging model and get the initial output. A human then goes through the output of this first phase and corrects any erroneously tagged words by hand. This tagged text is then submitted to the tagger, which learns correction rules by comparing the two sets of data. Several iterations of this process are sometimes necessary before the tagging model can achieve considerable performance. The transformation based approach is

similar to the rule based approach in the sense that it depends on a set of rules for tagging.

Transformation-Based Tagging, sometimes called Brill tagging, is an instance of the Transformation-Based Learning (TBL) approach to machine learning (Brill, 1995) and draws inspiration from both the rule-based and stochastic taggers. Like the rule-based taggers, TBL is based on rules that specify what tags should be assigned to a particular word. But like the stochastic taggers, TBL is a machine learning technique, in which rules are automatically induced from the data.

3. LITERATURE SURVEY FOR INDIAN LANGUAGES

Compared to Indian languages, foreign languages like English, Arabic and other European languages have many POS taggers [1]. Literature shows that, for Indian languages, POS taggers were developed only in Hindi, Bengali, Panjabi and Dravidian languages. As per our knowledge, no other publically available attempts are available in other Indian languages.

3.1 POS Taggers for Hindi Language

A number of POS taggers were developed in Hindi language using different approaches. In the year 2006, three different POS tagger systems were proposed based on Morphology driven, ME and CRF approaches respectively. There are two attempts for POS tagger developments in 2008, both are based on HMM approaches and proposed by Manish Shrivastava and Pushpak Bhattacharyya. Nidhi Mishra and Amit Mishra proposed a Part of Speech Tagging for Hindi Corpus in 2011. In an another attempt, a POS tagger algorithm for Hindi was proposed by Pradipta Ranjan Ray, Harish V., Sudeshna Sarkar and Anupam Basu.

i) In the first attempt, Smriti Singh proposed a POS tagging methodology which can be used by languages having lack of resources [2]. The POS tagger is built based on hand-crafted morphology rules and does not involve any sort of learning or disambiguation process. The system makes use of locally annotated modestly-sized corpora of 15,562 words, exhaustive morphological analysis backed by high-coverage lexicon and a decision tree based learning algorithm (CN2). The system uses Lexicon lookup for identifying the other POS categories. The performance of the system was evaluated by a 4-fold cross validation over the corpora and found 93.45% accuracy.

ii) Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke, proposed a POS tagger based on Maximum Entropy (ME) based approach [2]. To develop a POS tagger based on ME approach requires feature functions extracted from a training corpus. Normally a feature function is a boolean function which captures some aspect of the language which is relevant to the sequence labeling task. The experiment showed that the performance of the system depend on size of the training corpus. There is an increase in performance till it reaches 75% of the training corpus after which there is a reduction in accuracy due to over fitting of the trained model to training corpus. The least and best POS tagging accuracy of the system was found to be 87.04% and 89.34% and the average accuracy over 10 runs was 88.4%.

iii) The third POS tagger is based Conditional Random Fields developed by Agarwal Himashu and Amni Anirudh in 2006 [2]. This system makes use of Hindi morph analyzer for training purpose and to get the root-word and possible POS tag for every word in the corpus. The training is performed with CRF++ and the training data also contains other information like suffixes, word length indicator and special characters. A corpus size of 1, 50,000 words were used for training and testing purposes and accuracy of the system was 82.67% .

iv) The HMM based approach was intended to utilize the morphological richness of the languages without resorting to complex and expensive analysis [2]. The core idea of this approach was to explode the input in order to increase the length of the input and to reduce the number of unique types encountered during learning. This idea increases the probability score of the correct choice and at the same time decreasing the ambiguity of the choices at each stage. Data sparsity also decreases by new morphological forms for known base words. Training and testing was performed with an exploded corpus size of 81751 tokens which was divided into 80% and 20% parts respectively.

v) 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%.

vi) Nidhi Mishra and Amit Mishra proposed a Part of Speech Tagging for Hindi Corpus in 2011 [4]. In the proposed method, the system scans the Hindi corpus and then extracts the sentences and words from the given corpus. Also the system search the tag pattern from database and display the tag of each Hindi word like noun tag, adjective tag, number tag, verb tag etc.

vii) Based on lexical sequence constraints, a POS tagger algorithm for Hindi was proposed by Pradipta Ranjan Ray, Harish V., Sudeshna Sarkar and Anupam Basu [5]. The proposed algorithm acts as the first level of part of speech tagger, using constraint propagation, based on ontological information, morphological analysis information and lexical rules. Even though the performance of the POS tagger has not been statistically tested due to lack of lexical resources, it covers a wide range of language phenomenon and accurately captures the four major local dependencies in Hindi

3.2 POS Taggers for Bengali

A substantial amount of work has already done in POS tagger developments for Bengali language using different approaches. In the year 2007, two stochastic based taggers were proposed by Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu using HMM and Maximum Entropy (ME) approaches. Also Ekbal Asif developed a POS tagger for Bengali language using Conditional Random Fields (CRF). In 2008, Ekbal Asif and Bandyopadhyay S developed another machine learning based POS tagger using SVM algorithm. An Unsupervised Parts-of-Speech Tagger for the Bangla language was proposed by Hammad Ali in 2010. Debasri Chakrabarti of CDAC Pune proposed a Layered Parts of Speech Tagging for Bangla in 2011.

i) In the first attempt three different types of stochastic POS taggers were developed. In this attempt a supervised and

semi supervised bigram HMM & a ME based model was explored based on tagset of 40 tags [1][2]. The first model called as HMM-S makes use of the supervised HMM model parameters where as the second uses the semi supervised model parameters and is called HMM-SS. A manually annotated corpus of about 40,000 words was used for both supervised HMM and ME model. For testing a set of randomly selected 5000 words have been used for all three cases and the results showed that, the supervised learning model outperforms over other models. They also showed that further improvement can be achieved by incorporating a morphological analyzer for any model.

ii) The second POS tagger is based Conditional Random Fields (CRF) framework where features selection plays an important role in the development of POS tagger [6][2]. A tagset of 26 tags were used to develop the POS tagger. In this approach the system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various POS classes. For training purpose a corpus size of 72,341 tagged words were used. The system was tested with 20000 words selected from out of corpus and achieved 90.3%.

iii) The third POS tagger for Bengali is based on statistical approach using a supervised machine learning algorithm called SVM [1][2]. The earlier CRF based corpus was used for training and testing the POS tagging system using SVM based algorithm. The entire training corpus was divided in to two different set of sizes 57,341 and 15000 words each and used as training and development set. The test data of CRF model was used to evaluate the performance of the SVM based system and reported 86.84% accuracy.

iv) In the year 2010, Hammad Ali proposed an unsupervised POS tagger for the Bangla language based on a Baum-Welch trained HMM approach [8]. The proposed Layered Parts of Speech Tagger is a rule based system, with four levels of layered tagging [6]. The tagset used in the POS tagger was based on common tag set for Indian Languages and IIIT tagset guidelines. In the first level, a universal category containing 12 different categories are identified which is used to assign ambiguous basic category of a word. Followed by the first level, disambiguation rules are applied in the second level with more detail morphological information. The third and fourth levels are intended to tagging of multi word verbs and local word grouping. The proposed rule based approach shows better performance.

3.3 POS Taggers for Punjabi Language

There is only one publically available attempt proposed in POS tagger for Panjabi language [2]. Using rule based approach, a Panjabi POS tagger developed by Singh Mandeep, Lehal Gurpreet, and Sharma Shiv, in 2008. The fine–grained tagset contain around 630 tags, which consists of all the tags for the various word classes, word specific tags, and tags for punctuations. This tagger is different from the other in such a way that only handwritten linguistic rules are used to disambiguate the part-of-speech information for a given word, based on the context information. Using the rule based disambiguation approach a database was designed to store the rules. To make the structure of verb phrase more understandable four operator categories have been established. Also a separate database was maintained for marking verbal operator. The performance of the system was manually evaluated to mark the correct and incorrect tag assignments and the system reports an accuracy of 80.29% including unknown words and 88.86% excluding unknown words.

3.4 POS Taggers for South Dravidian Languages

Some noticeable attempts were done in Dravidian languages like Tamil, Telugu, Malayalam and Kannada language. There are six different attempts for POS taggers developments in Tamil language. There are three different attempts in Telugu and two attempts in case of Malayalam languages. There is only one corpus based POS tagger was developed in Kannada language.

3.4.1 POS Taggers for Tamil

There are six different attempts for the development in POS tagger for Tamil language. Vasu Ranganathan proposed a Tamil POS tagger based on Lexical phonological approach. Another POS tagger was prepared by Ganesan based CIIL Corpus and tagset. An improvement over a rule based Morphological Analysis and POS Tagging in Tamil were developed by M. Selvam and A.M. Natarajan in 2009. Dhanalakshmi V, Anand Kumar, Shivapratap G, Soman KP and Rajendran S of AMRITA university, Coimbatore developed two POS taggers for Tamil using their own developed tagset in 2009.

i) Vasu Ranganathan developed a POS tagger for Tamil called 'Tagtamil' based on Lexical phonological approach [9]. Morphotactics of morphological processing of verbs was performed using index method. The advantages of Tagtamil POS tagger is that, it handle both tagging and generation.

ii) The second Tamil POS tagger was based on CIIl corpus and proposed by Ganesan [9]. He used his own tagset, and he tagged a portion of CIIL corpus by using a dictionary as well as a morphological analyzer. Manual correction was performed and trained the system repeatedly in order to increase the performance of the system. The tags are added morpheme by morpheme. Its efficiency in other corpora has to be tested.

iii) The third POS tagger system was proposed by Kathambam using heuristic rules based on Tamil linguistics for tagging, without using either the dictionary or the morphological analyzer [9]. The system used twelve heuristic rules and identifies the tags based on PNG, tense and case markers. Using a list of words in the tagger, the system check for standalone words. Unknown words are tagged using 'Fill in rule' by using bigram approach.

iv) Using Projection and Induction techniques, an improved rule based morphological analysis and POS Tagging in Tamil was proposed by M. Selvam and A.M. Natarajan in 2009 [10]. Rule based techniques cannot address all inflectional and derivational word forms. There for improvement of rule based morphological analysis and POS tagging through statistical methods like alignment, projection and induction is essential. The proposed idea was based on this purpose and achieved an improved accuracy of about 85.56%. Using an alignment-projection techniques and categorical information, a well organized POS tagged sentences in Tamil were obtained for the Bible corpus. Through alignment, lemmatization and induction processes, root words were induced from English to Tamil. Root words obtained from POS projection and morphological induction, further improved the accuracy of the rule based POS tagger.

v) Dhanalakshmi V, Anand Kumar, Shivapratap G, Soman KP and Rajendran S of AMRITA University, Coimbatore developed a POS tagger for Tamil using Linear Programming approach [11]. They have developed their own POS tagset consists of 32 tags and used in their POS tagger model. They have proposed a SVM methodology, based on Linear Programming for implementing automatic Tamil POS tagger. A corpus of twenty five thousand sentences is trained with linear programming based SVM. The testing was performed using 10,000 sentences and reported an overall accuracy of 95.63%.

vi) In another attempt they have developed a POS tagger using machine learning techniques, where the linguistical knowledge is automatically extracted from the annotated corpus [12]. The same tagset was used here also to develop POS tagger. This is a corpus based POS tagger and annotated corpus size of two hundred and twenty five thousand words was used for training (1, 65,000 words)) and testing (60,000 words) the accuracy of the POS tagger. Support vector machine algorithms were used to train and test the POS tagger system and reported an accuracy of 95.64%.

3.4.2 POS Taggers for Telugu Language

NLP in Telugu language is better position when compared with other South Dravidian and many of other Indian languages. There are three noticeable POS tagger developments in Telugu, based on Rule-based, Transformation based learning and Maximum Entropy based approaches [2]. An annotated corpus of 12000 words was constructed to train the transformation based learning and Maximum Entropy based POS tagger models. The existing Telugu POS tagger accuracy was also improved by a voting algorithm by Rama Sree, R.J. and Kusuma Kumari P in 2007.

i) The rule based approach uses various functional modules which works together to give tagged Telugu text [2]. Tokenizer, Morphological Analyzer, Morph-to-POS translator, POS disambiguator, unigram, bigram rules and Annotator are the different functional modules used in the system. The function of Tokenizer is to separates pre-edited input text into separate sentences and each sentence to words. These words are then given to MA for analysis of each word. Pattern rule based Morph-to-POS translator then converts morphological analysis into their corresponding tags. This is followed by handling the disambiguation problem by the POS disambiguator which reduces the problem of POS ambiguity. Using unigram and bigram rules ambiguity is controlled in the POS tagger system. Annotator is used to produce the tagged words in a text and reported accuracy of the system was 98%.

ii) In the second attempt, Brill transformation rule based Learning (TBL) was used to build a POS tagger for Telugu language [2]. The Telugu language POS tagger system consists of three phases of Brill tagger: Training, Verification and Testing. The reported accuracy of the proposed POS tagger is 90%.

iii) Another Telugu POS tagger was also developed based on Maximum Entropy approach [2]. The idea behind the ME approach is similar to the general principles used in other languages. The proposed POS tagger was implemented using publically available Maximum Entropy Modeling toolkit [MxEnTk] and the reported accuracy is 81.78%.

3.4.3 POS Taggers for Malayalam

There are two separate corpus based POS tagger for Malayalam language was proposed as follow:

i) In 2009, Manju K., Soumya S. and Sumam Mary Idicula, proposed a stochastic Hidden Markov Model (HMM) based part of speech tagger [4][2]. A tagged corpus of about 1,400 tokens were generated using a morphological analyzer and trained using the HMM algorithm. An HMM algorithm in turn generated a POS tagger model that can be used to assign proper grammatical category to the words in a test sentence. The performance of the developed POS Tagger is about 90% and almost 80% of the sequences generated automatically for the test case were found correct.

ii) The second POS tagger is based on machine learning approach in which training, testing and evaluations are performed with Support Vector Machine (SVM) algorithms developed by Antony P.J, Santhanu P Mohan and Dr. Soman K.P of AMRITA university Coimbatore in 2010 [13][2]. They have proposed a new AMRITA POS tagset and based on the developed tagset a corpus size of about 180,000 tagged words were used for training the system. The performance of the SVM based tagger achieves 94 % accuracy and showed an improved result than HMM based tagger.

3.4.4 POS Taggers for Kannada Language

Antony P J and Soman KP of Amrita University, Coimbatore proposed statistical approach to build a POS tagger for Kannada language using SVM [14]. We proposed a tagset consisting of 30 tags. The architecture of the proposed POS tagger in Kannada language is based on corpus based and supervised machine learning approach. The Part-Of-Speech tagger for Kannada language was modeled using SVM kernel. We have conducted a linguistic study to determine the internal linguistic structure of the Kannada sentence and based on this we developed a suitable tagset. A corpus size of fifty four thousand words was used for training and testing the accuracy of the tagger generators. From the experiment we found that accuracy increased with increasing the number of words in the corpus.

3.5 Generic POS Taggers for Hindi, Bengali and Telugu Languages

Many different attempts were done for developing POS tagger for three different languages namely Hindi, Bengali and Telugu in Shallow Parsing Contest for South Asian Languages in 2007 [8]. All the participant in this contest were given corpus of 20000 and 5000 words respectively for training and testing based on the IIT POS tagset which consists of 24 tags. In this contest, participants proposed eight different POS tagger development techniques. Half of the these ideas are based on HMMs technique and others used Two Level Training based, Naive Bayes, Decision Trees to Maximum Entropy Model and Conditional Random Fields for developing POS tagger. Even though all the HMM based approaches used Trigrams'n'Tags or the TnT tagger for POS tagging, there was a considerable differences in the accuracies. The noticeable fact is that no one used rule based approach to develop POS tagger in their contribution. The following section gives a brief description about each and every proposed POS tagger system.

i) G.M. Ravi Sastry , Sourish Chaudhuri and P. Nagender Reddy used HMMs for developing POS tagging [8]. They used a Trigrams'n'Tags or the TnT tagger for their proposed system. The advantage of TnT is that it is not optimized for a particular language and the system incorporates several methods of smoothing and of handling unknown words which improved the POS tagger performance. The second HMM based generic POS tagger was developed by Pattabhi and his team. They used linguistic rules to tag words for which the emission or transition probabilities are low or zero instead smoothing. Another HMM based approach was proposed by Asif and team. They are also avoided smoothing and for unknown words, the emission probability was replaced by the probability of the suffix for a specific POS tag. The final HMM based generic POS tagger was developed by Rao and Yarowsky. They have developed HMM based POS tagger along with other systems using different approaches. They used TnT based HMM, compared the result with other systems and found that HMM based system outperforms.

ii) Naive Bayes Classifier, A suffix based Naive Bayes Classifier and QTag are the other three approaches used by Rao and Yarowsky to develop generic POS tagger. A suffix based Naive Bayes Classifier uses a suffix dictionary information for handling unseen words.

iii) For modelling the POS tagger, Sandipan and team used Maximum Entropy approach and result shows that this approach is best suited to Bengali language. They used contextual features covering a word window of one and suffix and prefix information with lengths less than or equal to four. The output of the tagger for a word is restricted by using a manually built morphological analyser.

iv) In another attempt, Himanshu and his team used a CRF based approach to develop the POS taggers. In their system, they used a feature set including a word window of tow, suffixes information with length less than or equal to four, word length and flag indicating special symbols. A knowledge database was used to handle data sparsity by picking word & tag pairs which are tagged with high confidence by the initial model over a raw corpus of 150,000 words. Similar to the ME proposed by Sandipan, a set of tags listed in the knowledge database and the training data are used to restrict the output of the tagger for each word instead. The experiment results shows that the CRF approach is well suited for Bengli and Telugu and not performed well for Hindi.

v) A two level training approach based POS tagger model was proposed by Avinesh and Karthik. In this approach a Transformation Based Learning (TBL) was applied on top of a CRF based model. Morphological information like root word, all possible categories, suffixes and prefixes are used in the CRF model along with exhaustive contextual information with a window size of 6, 6 and 4 for Hindi, Bengali and Telugu respectively. The system performance is good for Hindi and Telugu when compare with Bengali.

vi) Using a Decision Forests approach, Satish and Kishore proposed POS tagging with some innovative features based on subwords (syllables, phonemes and Onset-Vocal-Code for syllables) for Indian languages like Hindi, Bengali and Telugu. The subwords are an important source of information to determine the category of the word in indian language and the performance of the system is encouraging only for Telugu.

4. DEVELOPMENT OF POS TAGSET FOR INDIAN LANGUAGES

A number POS tagsets are developed by different organization and persons based on the general principles of tagset design strategy. However, most of the tagsets are language specific and some of these tagset are constructed by considering general peculiarities of Indian languages. A few tagset attempts were based on the feature of South Dravidian languages and other aim to a particular language. The following section gives a brief description of tagsets developed for Indian languages.

i) In a major work, IIT Hyderabad developed a Tagset in 2007, after consultations with several institutions through two workshops [15]. The aim was to create a general standard tagset suitable for all the Indian languages. The tagset also consist of a detail description of the various tags used and elaborates the motivations behind the selection of these tags. The total number of tags in the tagset is 25.

ii) The 6th Workshop on Asian Languae Resources, 2008 was intended to design a common POS-Tagset framework for Indian Languages [16][17]. It was a shared work from experts from various organizations like Microsoft Research-India, Delhi University, IIT- Bombay, Jawaharlal Nehru University- Delhi, Tamil University- Thanjavur and AU-KBC Research Centre, Chennai. There are three levels of tagsets were proposed and the top level consists 12 universal categories for all Indian languages and hence these are obligatory for any tagsets. The other levels consist of tags which are recommended and optional categories for verbs and participles.

iii) Dr.Rama Sree R.J, Dr.Uma Maheswara Rao G and Dr. Madhu Murthy K.V proposed a Telugu tagset by carefully analyzing the two tagset developed by IIIT, Hyderabad and CALTS, Hyderabad in 2008[18]. The proposed tagset was developed based on the argument that an inflectional language needs additional tags. They proposed some additional tags over the existing tagset to capture and provide finer discrimination of the semantic content of some of the linguistic expressions.

iv) Dhanalakshmi V, Anand Kumar, Shivapratap G, Soman KP and Rajendran S of AMRITA university, Coimbatore developed a tagset for Tamil in 2009, called AMRITA tagset which consists of 32 tags [11].

v) Vijayalaxmi .F. Patil developed a POS tagset for Kannada language in 2010 which consists 39 tags [7]. This tagset was developed by considering the morphological as well as syntactic and semantic features of the Kannada language.

vi) Antony P J, Santhanu P Mohan and Soman KP of AMRITA University, Coimbatore developed a tagset for Malayalam language in 2010. The developed tagset is based on AMRITA tagset which consists of 29 tags [Antony POS].

vii) Central Institute of Indian Language (CIIL) proposed a tagset for Hindi language based on Penn tagset [19]. This tagset was designed to includes more lexical categories than IIIT-Hyderabad and containing 36 tags.

viii) IIT- Karagpur developed a tagset for Bengali language which consists of 40 tags [20]. Another tagset called CRBLP tagset which consists of a total of 51 tags, where 42 tags are general POS tags, and 9 other tags are intended for special symbols [20].

ix) Antony P J and Soman KP of AMRITA University, Coimbatore developed a tagset for Kannad language in 2010. The developed tagset is based on AMRITA tagset which consists of 30 tags [Antony POS].

5. CONCLUSION

In this paper work, we have presented a survey on developments of different POS tagger systems as well as POS tagsets for Indian languages. Additionally we tried to give a brief idea about the existing approaches that have been used to develop POS tagger tools. From the survey we found out that almost all existing Indian language POS tagging systems are based on statistical and hybrid approach. The main effort and challenge behind each and every development is to design the system by considering the agglutinative and morphological rich features of language.

6. ACKNOWLEDGMENTS

We acknowledge our sincere gratitude to Mr. Benjamin Peter (Assistant Professor, MBA Dept, St.Joseph Engineering College, Mangalore, India) and Mr. Rakesh Naik (Assistant Professor, MBA Dept, St.Joseph Engineering College, Mangalore, India) for their valuable support regarding proof reading and correction of this survey paper.

7. REFERENCES

[1] Akshar Bharathi and Prashanth R. Mannem (2007), "Introduction to the Shallow Parsing Contest for South Asian Languages", Language Technologies Research Center, International Institute of Information Technology, Hyderabad, India 500032.

[2] Dinesh Kumar and Gurpreet Singh Josan,(2010), "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887) Volume6–No.5, September, 2010, www.ijcaonline.org/ volume6/number5 /pxc3871409 .pdf..

[3] 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. Proceeding of the ICON 2008.

[4] Nidhi Mishra Amit Mishra (2011), "Part of Speech Tagging for Hindi Corpus", International Conference on Communication Systems and Network Technologies.

[5] Pradipta Ranjan Ray, Harish V., Sudeshna Sarkar and Anupam Basu, "Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi",Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur, INDIA 721302. www.mla.iitkgp.ernet.in/papers/hindipostagging.pdf.

[6] Debasri Chakrabarti (2011), "Layered Parts of Speech Tagging for Bangla", Language in India www.languageinindia.c o m, M a y 2 0 1 1, Special Volume:Problems of Parsing in Indian Languages.

[7] 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..

[8] Hammad Ali (2010), "An Unsupervised Parts-of-Speech Tagger for the Bangla language", Department of Computer Science, University of British Columbia. 2010.

[9] S. Rajendran (2006), "Parsing in Tamil", LANGUAGE IN INDIA www.languageinindia.com Volume 6: 8 August, 2006.

[10] M. Selvam, A.M. Natarajan (2009), "Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques", International Journal of Computers, Issue 4, Volume 3, 2009.

[11] Dhanalakshmi V1, Anand Kumar1, Shivapratap G1, Soman KP1 and Rajendran S (2009), "Tamil POS Tagging using Linear Programming", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[12] Dhanalakshmi V1, Anand kumar M1, Rajendran S2, Soman K P., "POS Tagger and Chunker for Tamil Language".

[13] Jabar Hassan Yousif , Tengku Mohd Tengku Sembok, "Arabic part-of-speech tagger based support vectors machines".

[14] Antony P J, Santhanu P Mohan and Soman K P (2010), "SVM Based Parts Speech Tagger for Malayalam", International Conference on-Recent Trends in Information, Telecommunication and Computing (ITC 2010).

[15] A Part of Speech Tagger for Indian Languages (POS tagger), Tagset developed at IIIT - Hyderabad after consultations with several institutions through two workshops, 2007. shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf.

[16] G.M. Ravi Sastry , Sourish Chaudhuri and P. Nagender Reddy, "An HMM based Part-Of-Speech tagger and statistical chunker for 3 Indian languages", www.cs.cmu.edu/~schaudhu/publications.html.

[17] Pattabhi R K Rao T, Vijay Sundar Ram R, Vijayakrishna R and Sobha L (2007), "A Text Chunker and Hybrid POS Tagger for Indian Languages", AU-KBC Research Centre, MIT Campus, Anna University, Chromepet, Chennai, 2007 . shiva.iiit.ac.in/SPSAL2007/final/aukbc.pdf.

[18] Asif Ekbal, Samiran Mandal and Sivaji Bandyopadhyay (2007), "POS Tagging Using HMM and Rule-based Chunking", Workshop on shallow parsing in South Asian languages, shiva.iiit.ac.in/SPSAL2007/proceedings.php.

[19] Sathish Chandra Pammi and Kishore Prahallad (2007), "POS Tagging and Chunking using Decision Forests", Workshop on shallow parsing in South Asian languages, 2007. shiva.iiit.ac.in/SPSAL2007/proceedings.php.

[20] Mona Parakh, Rajesha N. and Ramya M (2011), "Sentence Boundary Disambiguation in Kannada Texts", Language in India www.languageinindia.c o m 1 1 : 5 M a y 2 0 1 1 Special Volume:Problems of Parsing in Indian Languages, Pages 17-19.

[21] Delip Rao and David Yarowsky (2007), "Part of Speech Tagging and Shallow Parsing of Indian Languages", Department of Computer Science, Johns Hopkins University, USA, 2007. The proceedings of the wworkshop on "Shallow Parsing in South Asian Languages" shiva.iiit.ac.in/SPSAL2007/final/iitmcsa.pdf.