International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 14015-14028

http://www.ripublication.com

Improved Tagging Approach for Part-of-Speech in Tamil Language Using an Ensemble

P.Iswarya*, Dr.V.Radha
Research scholar, Professor*
Department of Computer science
Avinashilingam Institute for Home Science and Higher Education for Women, India.
** iswaryacbe333@gmail.com, radhasrimail@gmail.com*

Abstract

Part Of Speech (POS) tagging is an important task in many language related processing activities. Feature selection is significant pre-processing step, and it is used to select smallest feature subset that increases accuracy of the model. The effectiveness of feature selection methods in POS tagging are analyzed by presenting five strategies namely F-Score based feature selection, Linear Discriminate Analysis, Decision tree, proposed Ensemble feature selection and no feature selection. To implement automatic Tamil POS tagging new heterogeneous ensemble classifier model is used, that consists of Support Vector Machine (SVM) and SVM with Wavelet Neural Network (WNN).The tagged and untagged corpus of 40,263 and 25,690 sentences used for training and testing respectively. The performance measures such as accuracy, precision, recall is used for evaluation. The tagging result suggests that in the proposed methodology ensemble feature selection with heterogeneous ensemble classifier combination performs better than other strategies with maximum accuracy of 97.84%.

Keywords: Ensemble, Feature Selection, Part Of Speech, Support Vector Machine, Tagging, Wavelet Neural Network.

1. Introduction

Part Of Speech (POS) tagging is a task that reads set of texts and assigns part of speech label to each of them. The first step in preprocessing of any language sentence is to retrieve part of speech information that helps in processing many language related activities [1]. Some of the activities include natural language parsing, natural language understanding, speech recognition, summarization, question answering system, machine translation and information retrieval etc. Tagging can be achieved through different approaches, the rule based POS tagging approach consists of set of

handwritten rules and it uses dictionary or lexicon for tagging. Transformation based learning approach are combination of both rule based and machine learning models. The stochastic supervised learning models are widely used in previous research works and it is easier to maintain than other approaches. Also supervised machine learning model requires vast amount of pre annotated corpora for training. The languages like English, western, European languages with machine learning had reached more than 96% of accuracy in POS tagging [2]. The Hidden Markov Model (HMM) estimates parameter model using labeled corpora but Indian languages lack in such large POS labeled corpora, therefore simple HMM does not work well with small sized corpus [3]. In POS tagging the combination of multiple classifiers may yield better result than performance of individual classifiers. The ensemble classifier is divided into two categories, the classifier that applies single learning algorithm and also different learning algorithms over a dataset is known as homogeneous and heterogeneous classifiers respectively. In heterogeneous or homogeneous ensemble, the set of individual classifiers decision are combined using either majority voting or weight based voting schemes. The weight based voting depends on the error rate or performance of individual classifier. Through investigation of related works in Tamil POS tagging the Support Vector Machine (SVM) can be considered as an efficient tool which provides greater than 90% of accuracy, but it takes lot of time in processing the data. The Wavelet Neural Network (WNN) is a powerful nonlinear tool that provides better approximation facility than multilayer perceptron and radial basis function networks [4]. In this paper SVM and WNN classifiers are used to form Heterogeneous Ensemble Classifier (HEC).

Feature selection is an important step after feature extraction process, and its main goal is to find subset from full feature set which helps in improving classification accuracy. The most common feature selection approaches used in this paper are decision tree, F-Score and Linear Discriminate Analysis. Each feature selection method is implemented with ensemble classifier over a dataset to analyze their efficiency. The proposed Ensemble Feature Selection (EFS) gives more accuracy than individual feature selection methods, and EFS that determines essential feature subset by integrating all feature selection outputs using union operator. The proposed ensemble feature selection with heterogeneous ensemble system produces the maximum accuracy of 97.84% for Tamil language.

The paper is organized as follows. Section 2 discusses about some of the related works in POS tagging. Section 3 tells about a methodology of the proposed model and Section 4 lists the set of tags and features that were during this work. Section 5 describes about the various feature selection strategies. The proposed heterogeneous ensemble classification and their base classifiers are described in Section 6. In Section 7 experimental results are presented and discussed. Finally conclusion is presented in Section 8.

2. Related Works

Many research works were carried out in POS tagging for Indian languages using different approaches; here some of the previous works are discussed briefly. The

morpheme based language model approach [5] was used to tag Tamil words by extracting word features like stem type, last morpheme and previous to last morpheme. Their corpus consists of 4,70,910 words tagged with 35 POS category labels in semi-automatic manner using available morphological analyzer and the estimation of contribution factor is done through iterative scaling technique. The system is tested with 73,017 words to obtain accuracy of 95.92%. In this study [1] automatic Tamil POS tagging has been developed by implementing Support Vector Machine based on linear programming model. The customized POS tag set was created, which consists of 32 tags by considering only the grammatical characteristics and excluding grammatical features. The size of corpus is about 25000 sentences, out of which 15000 sentences are taken for training and remaining 10000 sentences are used for testing; then this system achieves the overall accuracy of 95.63%.

In [6] hybridization of rule based POS tagging approach with projection and Induction technique used to tag Tamil words. The system requires aligned parallel English-Tamil Bible corpus, root word dictionary, and alignment dictionary. The tag set has 600 tags; test case consists of 1000 sentences that provide the accuracy of 92.48%. The paper [7] presents hybridization of Hidden markov model with rule based approach. The tag set consists of 17 basic tags, 31 sub tags and for testing 6000 words were used to produce an accuracy of 97% in Tamil POS tagging.

The POS tagging is considered as an optimization problem, to solve this problem Multi Objective Optimization (MOO) technique called Archived Multiobjective Simulated Annealing (AMOSA) is used. The two base classifiers used in forming ensemble such as Support Vector Machine and Conditional random field, and it uses Simulated Annealing (SA) based method to find an appropriate weight of vote obtained in each classifier output classes. AMOSA approach shows the accuracy of 90.45% and 89.88% for Bengali and Hindi language respectively [3].

Through the literature survey in POS tagging, few of the works have implemented feature selection approaches in tagging application. Many of works have been carried out using SVM classifier, and achieve better accuracy in Tamil POS tagging. The tagging idea of AMOSA based Ensemble classification efficiency was reported in Bengali and Hindi language. SVM and AMOSA based works were compared with proposed methodology to analyze their performance.

3. Proposed Methodology

The first step in the proposed methodology is preprocessing the document which involves tokenization, punctuation removal (except full stop), Case folding and Normalization. After preprocessing, feature extraction process has been carried out to extract relevant features from the dataset. The performance of the system will get reduce due to the presence of noisy features in the feature set. The feature selection is an important step to select best feature subset from entire set and their techniques include Decision tree, F-Score, LDA and proposed Ensemble Feature Selection (EFS) methods. The proposed EFS method combines the merits of all three feature selection approaches. In the last step, all selected features obtained from different feature selection strategies given as input to the heterogeneous SVM-WNN ensemble

classification system for tagging. The proposed methodology steps are shown in Figure 1.

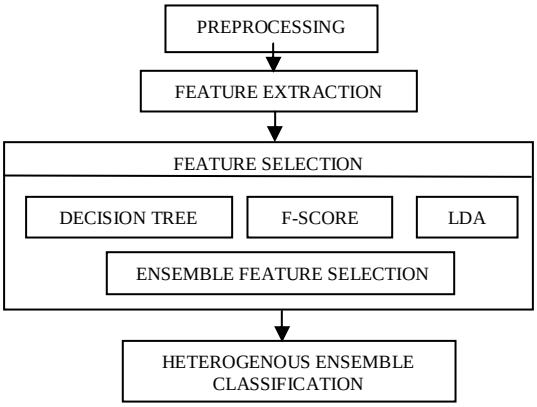

Figure 1: Proposed Methodology

4. Tag Set and Feature Extraction

The standard tag set was designed for Indian languages annotation task consists of 26 POS tags that are listed in Table 1, and it is developed at IIIT Hyderabad. Feature Extraction is an important step for all tasks; the features that are extracted from corpus are listed in Table 2.

Table 1: POS Tag set

S.No	TAG DESCRIPTION	TAG	S.No	TAG DESCRIPTION	TAG
1	Noun	<NN>	14	Quantifier	<QF>
2	Noun Location	<NST>	15	Cardinal	<QC>
3	Proper Noun	<NNP>	16	Ordinal	<QO>
4	Pronoun	<PRP>	17	Common Noun	<CL>
5	Demonstrative	<DEM>	18	Intensifier	<INTF>
6	Verb Finite	<VM>	19	Interjection	<INJ>
7	Verb Aux	<VAUX>	20	Negation	<NEG>
8	Adjective	<JJ>	21	Quotative	<UT>
9	Adverb	<RB>	22	Symbol	<SYM>
10	Postposition	<PSP>	23	Compounds	<C>
11	Particles	<RP>	24	Reduplicative	<RDP>
12	Conjunction	<CC>	25	Echo	<ECH>
13	Question Words	<VQ>	26	Unknown	<UNK>

Table 2: Description of Features for Tamil POS Tagging

Feature	Description
Context words	(w-1,w+1)(w-2,w+1)(w-2,w+2)(w-1,w+2)
Prefixes	P(2), P(3),P(4)
Suffixes	S(1),S(2),S(3)…….S(15)
POS features	POS(-1), (POS(-1),POS(-2))
Lexicalized features	NN, VV, ADJ, PR, UN
Digit feature	If NUM set to(1), else set(0)
Symbol feature	If SYM set to (1) else set (0)
Length feature	L(w) > 4 set to 1; else L(W)=0
Frequency	If word in Frequency word list set F=1 ; else F=0
Function word list	If word in function word list set G=1 ; else G=0
Sentence info	Punctuation (.)
Chunk info	Chunk information of current word

5. Different Feature Selection Strategies

Feature selection is focused on extracting most relevant features from the feature set which helps to improve classification accuracy. Five strategies used in this research work are elaborated below.

5.1 No Feature Selection

This strategy directly use the full feature set for ensemble classification, without applying any of the feature selection approaches. When POS tagged corpus is given as input to learn package component, it trains the model using the extracted features. The trained model contains features of tagged words, lexicon and merged models. The tagger component uses the trained model to tag the untagged corpus. Then the evaluator components evaluate the tagging output in terms of accuracy and it is very useful component to tune the system parameters.

5.2 Decision Tree (DT) Approach

The decision tree algorithm is introduced by Quinlan and their variables values representation is in the form of tree, which consists of root, branches, and leaves. The first step in construction of root node is to select a best test attribute done through a measure called information gain, calculated for all attributes, and the node having highest information gain is considered as root node. The splitting criterion is applied from root node to produce branches or subsets, and this splitting procedure is repetitively iterated to divide the subsets further. This procedure will continue until all tuples in subset belong to same class or if split reaches to a possible extent [8]. The bottom nodes in the decision tree are called leaves, then for each leaf decision rule will provide a unique path to reach a class label. The decision tree induction algorithm is greedy in nature that constructs tree in top-down fashion recursively. The criteria to select best test attribute for each node is using Information Gain which finds expected

reduction in entropy or uncertainty, and also Entropy is a common way to measure the impurity. Here higher Entropy gives more the information content and it is presented in Equation. (1).

$$(S) = \text{Entropy} \sum_i - p_i \log 2\ p_i \quad (1)$$

Where p_i is the probability of class i, in Eq. (1).

$$InformationGain(S, A) = E(S) - \sum_{values(A)} \frac{|S_V|}{|S|} E(S_V) \quad (2)$$

In Equation (2) ,E is an entropy , Values(A) represents the set of all possible values of attribute A, and S_v the subset of S for which attribute A has value v , S_v = {s in S | A(s) = v}.

To construct the decision tree, best attribute is selected from all the attributes based on heuristic algorithm with maximum information gain. Suppose D consists of 50, 30, 10 and 10 samples belong to noun, verb, adjective, and adverb respectively. The original entropy of above samples is defined in below implementation.

$$H[D] = -\frac{50}{100}\log_2\frac{50}{100} - \frac{30}{100}\log_2\frac{30}{100} - \frac{10}{100}\log_2\frac{10}{100}\log_2 - \frac{10}{100}\log_2\frac{10}{100}$$

$$\left.\begin{array}{l} Gain\ (D, context) \equiv H[D] - H_{context}[D \\ Gain\ (D, prefix) \equiv H[D] - H_{prefix}[D] \\ . \\ . \\ . \\ Gain\ (D, sen\ \text{inf}\ o) = H[D] - H_{sent\ \text{inf}\ o}[D] \end{array}\right\} Max\ (Gain)$$

For attribute A_i with v feature values, the root of the current tree will partition D into v subsets D_1, D2…..D_v. Then the expected entropy for partitioning each A_i is determined (i.e. $H_{context}$, H_{prefix},H_{suffix}, $H_{POSfeature}$, $H_{lexfeature}$, H_{Digit}, H_{Symbol},H_{length}, $H_{frequency,}$ $H_{function}$,$H_{seninfo}$). Then the attribute with highest information gain is selected as branch or root or split tree. A decision tree can be converted to a set rules and finding the best tree is using NP-Hard. The partitioning stops when there is no remaining attributes for further split.

5.3 F-Score Approach

F-score is simple but effective technique which determines discriminative value for each feature in feature subset, and a feature with higher F-Score value has more discriminative power. Given training samples X_k, k=1,2….m, the number of positive and negative instances is given by n^+ and n^- respectively then the i^{th} feature vector of F-Score is calculated using Equation. (3).

$$F(i) = \frac{\left(\overline{x_i^{(+)}} - \overline{x_i}\right)^2 + \left(\overline{x_i^{(-)}} - \overline{x_i}\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,j}^{(+)} - \overline{x_i^{(+)}}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \overline{x_i^{(-)}}\right)^2} \quad (3)$$

In Equation (3) $\overline{x_i^{(+)}}, \overline{x_i^{(-)}}, \overline{x_i}, x_{k,i}^{(+)}, x_{k,i}^{(-)}$ represents Positive, negative, whole sample, i^{th} feature vector of k^{th} positive and negative instances respectively. The numerator indicates discrimination between positive and negative sets, and denominator is sum of deviation within each of feature sets [9].

Fisher score is calculated for every feature attribute; here it is computed for twelve attributes. The several possible threshold values (ranges from 0 to 1) are taken for the experiment, and it checks which threshold value give minimum validation error. Consider the threshold value x, then eliminate the features that are below the threshold, similarly the procedure is carried out for all threshold values. The training data is divided into two parts: one subset is new training data that are mapped into a higher dimensional space. The remaining part is used as predictor to predict using HEC procedure. The procedure is repeated several times to determine lowest average validation error and skip the features with low F-score.

5.4 Linear Discriminant Analysis (LDA)

Linear Discriminant analysis is first introduced by Ronald A. Fisher and it is also called as Fisher Linear discriminant. Data mining related applications will have a higher dimension feature which consists of many irrelevant or redundant data that may over fit, and gives less interpretable results. The goal of LDA is to perform dimensionality reduction without loss of any information. The LDA constructs one or more discriminate equations D_i from linear combination of predictor variables X_k such that different groups differ as much as possible as D.

The LDA is expressed in Equation. (4)

$$D_i = b_0 + \sum_{k-1}^{p} b_k x_k \tag{4}$$

Where D represents discriminate score.

The Linear Discriminant Analysis (LDA) finds the features which are best to separate all the classes using discriminate score. Then compute mean μ and followed by covariance for class S. To find a feature that makes maximum projection between the classes is to maximize the function that represents difference between the means and normalized by a measure of within-class variability or scatter. The LDA projection solution obtained by solving a generalized Eigen value problem and the projection vector which gives maximum Eigen value will provide good separability between the classes. Through LDA, selected subsets of features are fed as input to HEC.

5.5 Ensemble Feature Selection (EFS)

This proposed approach is combination all above strategies, where whole training data is given as input to each feature selection approach. As a result of all feature selection approaches, the subset obtained from each approach, gets integrated using union operator to form new feature data. Then the subset determined is fed as training data input to Support Vector machine classifier to retrieve good support vectors. The efficient performance is not obtained, while using full feature set data as an input to

HEC. The support vectors of training data are again given as input to the wavelet neural network.

In WNN, haar wavelet is used as an activation function, of decomposition level two. Then the untagged test corpus data can be directly fed as an input to the wavelet neural network, to obtain tagged corpus result. The structure of ensemble feature selection is shown in Figure 2.

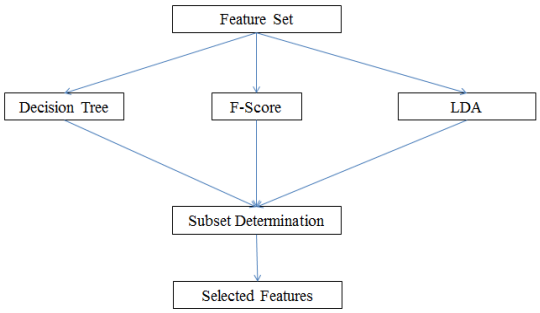

Figure 2: Ensemble Feature Selection

6. Heterogeneous Ensemble Classification (HEC)

The proposed heterogeneous ensemble classifier is built with base classifiers that take different feature subset combinations as input. The SVM classifier is used as preprocess to reduce the training set to subset version that contains more critical details in deciding classification boundary. Five feature selection strategies used in this study were HEC+ no feature selection, DT+HEC, F-Score+HEC, LDA+HEC and ENFS+HEC. The heterogeneous ensemble classifier system built with different learning forms lead to SVM with no feature selection, SVM-WNN with no feature selection, SVM with any one feature selection method, SVM-WNN with any one feature selection method and SVM-WNN with EFS. Wavelet neural network is trained using support vectors of SVM and for testing; data is directly applied to Wavelet Neural Network (WNN) to analyze its efficiency.

Thus the hybrid method classification is performed in two steps (a) Original features is used by SVM and support vectors are extracted. (b) The extracted support vectors and corresponding actual output values are fed as new training set for Wavelet Neural network. The Heterogeneous ensemble classification procedure is presented in Figure 3. The base classifiers used to form heterogeneous ensemble classification is described briefly.

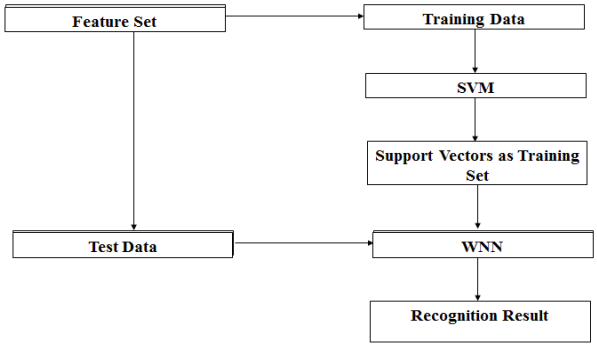

Figure 3: Heterogeneous Ensemble classification procedure

6.1 Support Vector Machine

Support Vector Machine is derived from statistical learning theory by Vapnik et.al in 1992 and it is a supervised learning model used for classification, regression and outlier detection. In Natural language processing SVM classifier is used in text categorization, text recognition, parsing tasks and also it achieved competitive accuracy [10]. Irrespective of feature vector dimensions Support vector machine shows effective generalization performance over other statistical learning conventional algorithm. SVM Tool is a language independent sequential tagger implemented in Tamil POS tagging.

Suppose SVM is example of two classes problem linear classifier assumed to have label y as +1 (positive) and -1 (negative) instance respectively. Then the X denotes a vector with components x_i, where i=1,2…..n, and linear classifier of their dot product is defined as

$W^T X = \sum_{i=1}^{n} w_i x_i$, for all $x_i \in X$. Here $(x_i,y_i)^n_{i=1}$ every x_i vector associated with labelled with y_i . Linear classifier based on linear discriminant function is of the form

$$f(X) = W^T X + b \tag{5}$$

The SVM separates positive and negative examples by constructing the hyper plane, consider if the case bias ‘b’ is set to zero (b=0) this makes all points perpendicular to w and translate the hyper plane away from origin.

$$\{X : f(X) = W^T X + b = 0\} \tag{6}$$

When input examples are linear that constructs linear hyper plane, and linear classifier makes the decision boundary linear. Conversely if input data is present in non-linear manner it will lead to a non-linear classifier [11].

The SVM tool used in this work consists of three components trainer, tagger and evaluator. For SVM learner trains a set of SVM classifiers with different feature combinations using SVM light software in C language. For a two class problem binary SVM is used, if more than two classes binary SVM extended to multiclass SVM [12]. The evaluator component evaluates the performance output for the test set.

In this paper, pair wise sequential labeling SVM classification method with five degree polynomial kernel function is implemented.

6.2 Wavelet Neural Network

Wavelet neural network (WNN) is a feed-forward neural network consists of three layers, where input layer accepts one or more inputs; hidden layer has wavelet neurons or wavelons as activation function, and output layer with linear combination of one or more hidden nodes. The general structure of wavelet neural network is shown in Figure 4. The network output is computed using Equation. (7). In this expression of multidimensional wavelet Ψ is the mother wavelet; x is set of input vectors, m is the number of network inputs, λ is the number of hidden units and w stands for network weight [12].

$$\hat{y}(x) = g_2(x : \hat{w_z}) = w_{\lambda+1}^{[2]} + \sum_{j=1}^{n} w_j^{[2]} \prod_{j=1}^{n} \psi\left(\frac{x_j - w_{(\xi)\gamma}^{[1]}}{w_{(\xi)\gamma}^{[1]}} \right) + \sum_{i=1}^{n} w_t^{[0]} \cdot x_i \tag{7}$$

The selection of mother wavelet is depending on the choice of application. The network parameters were adjusted during the training phase and use of optimized initialization parameters will minimize loss function each time. In this paper, Haar wavelet is used as mother wavelet due to its simplicity than other wavelet functions. The back propagation neural network is most popular in training wavelet network [1], and to design it; in hidden layer, back propagation neural network sigmoid function is replaced with haar wavelet, as an activation function.

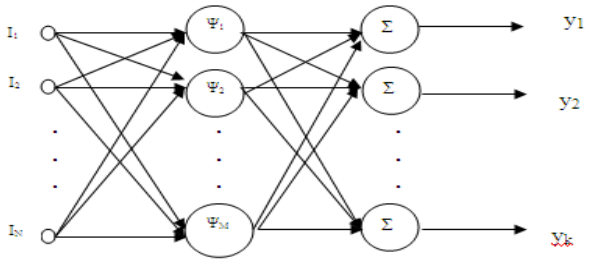

Figure 4: Structure of Wavelet Neural Network

7. Experimental Results and Discussion

Corpus is a huge repository incorporating various types of textual materials such as newspapers, fictions, weeklies, literary writings and so on. Tagged corpus is an important dataset for Natural language processing applications. The tagged corpus of 100K words was built, and these sentences were obtained from Tamil news Corpus of Forum for Information Retrieval (FIRE) dataset 2011. FIRE 2011 data consists of untagged Tamil sentences of newspaper documents which includes various categories are sports, politics, Entertainment, business, front page etc. Only several Tamil newspaper article categories are chosen for experimentation. Nearly 40% of data is tagged manually, and this data are given as training input to tagger component to

obtain tagged corpus; which contains some errors, and these errors are manually edited and corrected to increase the size of the corpus further. Each tag total count distributed in the tagged corpus is listed in Table 3.

In this paper various combinations (Unigram, bigram etc.) of available features are extracted from a tagged corpus, which are high dimensional. The different feature selection approaches are applied on full feature set to find most relevant subset features. To evaluate the performance of different strategies, the selected features obtained from each feature selection method used as training data for HEC. In SVM-WNN classifier, support vectors were fed as training input for Wavelet neural network to train iteratively with updated weights to attain minimum error. The test set consists of 25,690 words that are used to predict accuracy, precision, recall of five strategies and it is reported in Table 4, Table 5, and Table 6 respectively. The POS tagging accuracy are 91.66 %, 93.54 %, 95.78 %, 96.81 %, and 97.84% for Strategy I, II, III, IV, and V respectively. Through experimental analysis proposed EFS method performs more efficient than single feature selection method in the selection of optimal subset. Also proposed HEC classifier gives improved performance than using AMOSA based classifier and individual SVM classifier.

Table 3: Tag Count

S.No	TAG	TAG COUNT	S.No	TAG	TAG COUNT
1	<NN>	139168	14	<QF>	17765
2	<NST>	32893	15	<QC>	20123
3	<NNP>	17765	16	<QO>	453
4	<PRP>	8723	17	<CL>	1345
5	<DEM>	16264	18	<INTF>	2189
6	<VM>	20713	19	<INJ>	734
7	<VAUX>	7231	20	<NEG>	71281
8	<JJ>	6410	21	<UT>	8790
9	<RB>	890	22	<SYM>	3285
10	<PSP>	4897	23	<C>	6156
11	<RP>	145	24	<RDP>	10273
12	<CC>	90	25	<ECH>	1010
13	<VQ>	42167	26	<UNK>	1783

Table 4: Accuracy (%)

	AMOSA based Ensemble Classification	Proposed Ensemble Classification
Full Feature set	80.11	91.66
F-Score	82.32	93.54
LDA	83.46	95.78
DT	85.19	96.81
EFS	**89.46**	**97.84**

Table 5: Precision (%)

	AMOSA based Ensemble Classification	**Proposed Ensemble Classification**
Full Feature set	81.90	91.21
F-Score	83.26	92.15
LDA	83.91	92.77
DT	84.55	93.02
EFS	**91.06**	**95.71**

Table 6: Recall (%)

	AMOSA based Ensemble Classification	**Proposed Ensemble Classification**
Full Feature set	63.61	71.64
F-Score	64.10	74.94
LDA	65.76	75.43
DT	68.29	78.42
EFS	**74.24**	**86.69**

The problem of POS tagging in supervised learning model has attained more than 95% of accuracy in English and European languages [6]. Developing a supervised learning model in Indian languages is still difficult because there is lack in large annotated corpus and other resources. Some of the previous Tamil POS tagging works are developed with their own tag set and dataset.

Dhanalakshmi et al. designed their own tag set for Tamil language consisting of 32 tags and corpus is trained with 25000 sentences using linear programming based SVM, and achieved accuracy of 95.63%. Asif ekbal et.al developed the classifier of ensemble technique, and ensemble decisions are combined based on MOO technique, called as AMOSA. This method is evaluated for Bengali, Hindi languages, and achieved accuracy of 90.45% and 89.88% respectively. The Homogeneous AMOSA based support vector machine ensemble with different feature set strategies when implemented on Tamil language, their percentage of accuracy reduces to 12.6%, which is considered as the baseline performance. The proposed approach of ensemble feature selection with heterogeneous ensemble classifier overcomes baseline performance accuracy and achieves 97.84%. Most of the tagging errors may occur due to ambiguity words and unknown words.

8. Conclusion

In this paper five strategies with a proposed classifier were used for improving the Tamil POS tagger. The full feature set with HEC, F-Score feature selection with HEC, Linear Discriminate Analysis with HEC, and Decision tree with HEC and Ensemble Feature selection with HEC were implemented. These different strategies are

evaluated for Tamil language, and by comparing the performance of these strategies the proposed ensemble feature selection with heterogeneous ensemble classifier performs better than other. The precision and recall may affect due to the presence of ambiguous words in corpus. In future the accuracy will be further increased by making use of word sense disambiguation techniques, and the problem of tagging unknown words can be handled using notion of word patterns [13]. This POS tagging in Tamil language can be extended to tag other language words.

References

[1]. Dhanalakshmi V., Anandkumar., Shivapratap G., Soman KP., and Rajendran S,2009, " Tamil POS tagging using Linear programming," International journal of Recent Trends in Engineering, Vol.1, No.2.

[2]. Hasan F .,Uzzaman N ., and Khan M, 2007, "Advances and Innovations in systems, Computing Science and Software Engineering," Springer. Netherlands.

[3]. Ekbal A., and Saha, S., 2013, " Simulation Annealing based classifier ensemble techniques: Application to part of speech tagging," An International Journal on Multi-sensor, Multi-Source, Information fusion, Vol.14. Issue 3. Pp.288-300.

[4]. Huang M. and Cui Baotong.,2005, "Advances in natural computation," Springer, Berlin Heidelberg.

[5]. Lakshmana Pandian S., and Geetha T.V., 2008,"Morpheme based language model for Tamil Part-of-Speech Tagging," Polibits, No.38.

[6]. Selvam M.; and Natarajan A.M., 2009, "Improvement of rule based Morphological analysis and POS tagging in Tamil Language Via projection and Induction technique," International Journal of computers.Vol.3. Issue 4.

[7]. Lalithadevi S.; and Pattabhi Ramakrishna rao T., 2010, "A hybrid approach for POS tagging for relatively free word order languages," Proceedings of Knowledge sharing event on part-of-speech tagging. LDC-IL. CIIL.Mysore.

[8]. Badulesu L., 2007, "The choice of best attribute selection measure in Decision tree induction," Annals of University of Craiova- Mathematics and computer science series.Vol.34 (1). Pp.88-93.

[9]. Chen Y.; and Lin C., 2006, "Feature Extraction," Springer. Berlin Heidelberg.

[10]. Statistical Data mining tutorials- AutonLab, 2014, Last accessed during July 2014 from http://www.cs.cmu.edu/~awm/tutorials .

[11]. Ben-Hur A., and Weston J., 2010, "Data mining Techniques for the Life sciences," Human Press.USA.

[12]. Alexandridis A.; and Zapranis A., 2013,"Wavelet neural networks: A practical guide," Neural networks.Vol.42. Pp.1-27.

[13]. El-Jihad A., Yousfi A., and Si-Lhoussain A., 2011, “Morpho-Syntactic tagging system based on the pattern words for Arabic texts” The International Arab journal of Information technology. Vol.8. No.4.