Parts of Speech Taggers for Dravidian Languages

Anjali M K[1], BabuAnto P[2]
[1#] Research Scholar, Department of Information Technology,Kannur University, Kerala, India
[2#] Associate Professor, , Department of Information Technology,Kannur University, Kerala, India

***Abstract*-The process of assigning one of the parts-of- speech(POS) to the given word in a text is called Parts-of-speech tagging. POS tagging is a very important pre-processing task for language processing activities. This paper made a detailed study about the taggers available on morphologically rich Dravidian languages which includes Malayalam, Kannada, Tamil and Telugu. It also briefs various approaches used for POS tagging.**

Keywords- POS Tagger; Dravidian Language; Rule based; Stochastic; Hybrid; Malayalam; Telugu; Kannada; Tamil.

I. INTRODUCTION

Natural Language Processing (NLP) is concerned with the development of computational models of aspects of human language processing[1]. Natural Languages are ambiguous, Part-of-speech tagging is one of the disambiguation techniques at Lexical level of NLP. The significance of part-of-speech for language processing is the large amount of information they give about a word and its neighbors.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 one of the basic necessary tools. The accuracy of many NLP applications depend on the accuracy of POS tagger.

Dravidian languages consists of a family of about 70 languages spoken primarily in South Asia.The Dravidian languages are divided into South, South-Central, Central, and North groups; these groups are further organized into 24 subgroups. The four major literary languagesTelugu, Tamil,Malayalam, and Kannada—are recognized by the constitution of India. They are also the official languages of the states of AndhraPradesh, Tamil Nadu, Kerala and Karnatakarespectively[2]. These are four of the 22 official languages and 14 regional languages of India. In 2004 , The government declared Tamil as Classical language of India. In 2008, Kannada and Telugu got Classical status and in 2013, Malayalam was also given status of Classical language[3]. The rest of the paper is organised as follows. The next two sections give a brief description of the methods and statistical techniques used for POS tagging. The Sections IV,V, VI and VII presents the attempts made at Dravidian languages Malayalam, Kannada, Tamil and Telugu respectively.

II. PARTS-OF-SPEECH TAGGING METHODS

Part-of- speech tagging methods fall under the three general categories[1][4].

- Rule-based (linguistic)
- Stochastic (data-driven)
- Hybrid

Rulebased tagger- The rule based taggers use a set of handwritten rules based on morphological and contextual information. Most rule-based taggers have a two stage architecture. The first stage is a dictionary look-up procedure, which returns a set of potential tags (part-of-speech) and appropriate syntactic features for each word. The second stage uses a set of hand-coded rules to discard contextually illegitimate tags to get a single part of speech for each word. An example of rule based tagger is TAGGIT, which was used for the initial tagging of the brown corpus. Another rule-based tagger is ENGTWOL .

Stochastic Tagger- Stochastic taggers have data-driven approaches in which frequency based information is automatically derived from corpus and used to tag words. Stochastic taggers disambiguate words based on the probability that a word occurs with a particular tag. An early example of stochastic tagger was CLAWS (constituent likelihood automatic word-tagging system).

Hybrid Tagger- Hybrid taggers combine features of both the rule based and stochastic approaches. Like rule-based systems, they use rules to specify tags. Like stochastic systems, they use machine-learning to induce rules from a tagged training corpus automatically. The transformation based tagger or Brill tagger is an example of the hybrid approach.

III. STOCHASTIC APPROACHES

In Stochastic approaches, for a given word sequence, pick the most likely tag for each word, based on the probability of certain tag occurrences. This required a large sized training corpus for calculating the frequency and machine learning algorithms required which automatically learn to make accurate predictions based on past observations. Supervised taggers require pre-tagged corpora which serves as a basis for calculating word or tag frequencies, the tag sequence probabilities etc. Whereas the Unsupervised taggers do not rely on a pre-tagged corpus, but use computational methods that automatically induce word groupings and calculate the needed probability information. The approaches used in Dravidian languages are described below:

Hidden Markov Model(HMM)- An HMM is a doubly stochastic process with an underlyingstochastic process that is not observable (it is hidden),but can only be observed through another set ofstochastic processes that produce the sequence of observedsymbols. The elements and the mechanism of the type of HMMs are there a finite number of states say n in the model, at each clock time t, a new state is entered, based upon a transition probability distributionwhich depends on the previous state. After each transition is made, an observation output symbol is produced according to a probability distribution which depends on the current state[5].For a given word sequence, HMM taggers choose the tag sequence that maximizes the following formula: P(word|tag) * P(tag|previous n tags)[4]. For a model such as HMM that contains hidden variables the task of determining which sequence of variables is the underlying source of some sequence of observations is called the decoding task, the Viterbi algorithm is the most common decoding algorithm used for HMMs for part-of-speech tagging.

Support Vector Machines(SVM)- SVM is a learning machinethat classifies an input vector x using the decision function:f(x) = sign((x · w) + b)
SVMs are hyperplaneclassifiers and work by determiningwhich side of the hyperplane x lies. In the above formula, thehyperplane isperpendicular to w and at a distance b/||w|| from
the origin.An SVM is a parameterized functionwhose functional form is defined beforetraining. Training an SVM requires a labeledtraining set, because the SVM will fitthe function from a set of examples. Thetraining set consists of a set of *N* examples.Each example consists of an input vector,$\mathbf{x}_i$, and a label, y_i, which describes whetherthe input vector is in a predefined category.There are *N* free parameters in an SVM
trained with *N* examples.Given a training set x_i and its classification values $y_i \in \{-1, 1\}$, the training problem for SVMs is a minimization problemwith the constraints
$x_i \cdot w + b \geq +1$ for $y_i = +1$ (positive examples)
$x_i \cdot w + b \leq -1$ for $y_i = -1$ (negative examples)
which can be combined as:
$y_i(x_i \cdot w + b) \geq 0 \quad \forall i$ (combined constraints)
Intuitively, minimizing f with respect to this constraintfinds the location of the hyperplane directly between the positiveand negative training sets[6][7].

Maximum Entropy Models- Maximum entropy modeling is a framework for integrating information from many heterogeneous information sources for classification. The data for a classification problem is described as a number of features, each feature corresponds to a constraint on the model. Then compute the maximum entropy model, the model with the maximum entropy of all the models that satisfy the constraints. Choosing the maximum entropy model is motivated by the desire to preserve as much uncertainty as possible[8].Many problems in natural language processing (NLP) can be re-formulated as statistical classification problems, in which the task is to estimate the probability of "class" a occurring with "context" b, or p(a,b).The problem is to find a method for using the sparse evidence about the a's and b's to reliably estimate a probability method p(a,b). The principle of maximum entropy states that the correct distribution p(a,b) is that which maximizes entropy or uncertainty, subject to the constraints, which represents evidence ie the facts known to the experimenter.
If Adenotes the set of possible classes, and B denotes the set of possible contexts, p should maximize the entropy

$$H(p) = -\sum_{x \in \varepsilon} p(x) log p(x)$$

Where x = (a,b), aϵ A, bϵ B, and $\mathcal{E}$ = A x B,and should remain consistent with the evidence, the representation of the evidence then determines the form of p[9]. Maximum Entropy principle states to select a model from a set C of allowed probability distributions , choose the model $p_* \in C$ with maximum entropy H(p):

$$p_* = \underset{p \in C}{argmax}\, H(p)$$

there is always a unique model p_* with maximum entropy in any constraint setC[10].

Conditional Random Fields(CRF)- Conditional random fields, a frameworkfor building probabilistic models to segment and label sequence data. Consider, X is a random variable over data sequencesto be labeled, and Y is a random variable overcorresponding label sequences. All components Yi of Yare assumed to range over a finite label alphabet $\mathcal{Y}$. Forexample,X might range over natural language sentences andY range over part-of-speech taggings of those sentences,with $\mathcal{Y}$the set of possible part-of-speech tags. The randomvariables X and Y are jointly distributed, but in a discriminativeframework we construct a conditional modelp(Y|X) from pairedobservation and label sequences, anddo not explicitly model the marginal p(X).Let G=(V,E) be a graph such that $Y=(Y_v)_{v \in V}$,so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|X,Y_w,w \neq v)= p(Y_v|X, Y_w,w\sim v)$, where w~v means that w and v are neighbors in G. Thus, a CRF is a random field globally conditioned on the observation X[11]. CRFoffer several advantages over HMM and Maximum Entropy approach.

IV. MALAYALAM

Among the four cultivated languages of the Dravidian family, Malayalam comes last in the development of grammar and literature. Malayalam was lagging behind in other languages in the field of automation also. Works based on HMM, and SVM were developed. The following are the taggers developed:

Manju K etal proposed a stochastic Hidden Markov Model (HMM) based part of speech tagger [12]. 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 POS Tagger developed gave an accuracy of about 90%.

Rajeev R R etal proposed a Malayalam POS tagger based on Trigrams 'n' Tagger (TnT) with the Hidden Markov Model following the Viterbi algorithm[13]. With HMM and Viterbi algorithm, a search algorithm is used for various lexical calculations. The algorithm assumes that both the observed and the hidden word must be in a sequence. In the application of TnT, first the model parameters are created from a tagged training corpus and then the parameters are applied to the new text and actual tagging is performed.The tagset used are developed by IIIT. The authors claimed an accuracy of about 90.5%.

Antony P J et alof AMRITA university Coimbatore proposed a new tagger which is based on machine learning approach in which training , testing and evaluation are performed with Support Vector Machine (SVM) algorithm[14]. 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 system. The performance of the SVM based tagger achieves 94% accuracy and showed an improved result than HMM based tagger.

CIIL Mysore developed Automatic POS Tagger for Indian Languages using hybrid approach[15]. The precision at present is 86.2% (LDC-IL Tagset 84.2%,BIS Tagset 88.2%) but it is expected to go higher after more rounds of fine tuning.

V. KANNADA

In Kannada various POS taggers were developed based on SVM, HMM, CRF, Maximum Entropy and Rule based approaches. They are:

Antony et al proposed a part-of-speech tagger for Kannada language that can be used for analyzing and annotating Kannada texts[16]. Proposed a Tagset consist of 30 tags and part-of-speech Tagger is based on machine learning approach using Support Vector Machine (SVM). A corpus of texts, extracted from Kannada news papers and books, is manually morphologically analyzed and tagged using thedeveloped tagset. The system tested on56,000 words and the accuracy was 86%.

Shambavi et al proposed a Maximum Entropy based POS Tagger[17]. For training data they manually tagged 51267 words from EMILLE corpus. The tagset included 25 tags. The best suited feature set for the language was finalised after rigorous experiments. For testing data size of 2892 word forms was downloaded from Kannada websites. The reported accuracy was81.6%.

Shambavi et al proposed another two taggers based on machine learning algorithms which applied second order Hidden Markov Model (HMM) and Conditional Random Fields (CRF)[18]. For training and Testing data is taken from EMILLE corpus. Training data includes 51,269 words and test data consists of around 2932 tokens. The HMM tagger reported an accuracy of 79.9% and for CRF 84.58%.
Pallavi et al proposed a machine learning algorithm using CRFs[19]. Uses AUKBC tagset consisting of 45 tags. POS tagger is being developed using Javaprogramming and CRF++ (Yet Another CRF++ toolkit).To train and test the system, 1000words werecollected from on-line Kannada newspapers, and thenmanually processed and tagged. The manually tagged words are used to trainthe model with window size of 3. The authors claimed an accuracy of 99.49%.

Bhuvaneshwari proposed a rule based tagger based on a hierarchical tagset[20]. The system needed resources like tagset, dictionary, morphological system, named entity recognizer etc.The system takes an input aword or it may be file. If the word is found inthe dictionary the dictionary tag is assigned to the word. Otherwise it is passed to morphological system. The morphmodule checks for its inflections or derivational features and analyzes the word and assigns the tag using the hierarchicaltag set.Morphological system isdeveloped using well defined sandhi rules and using finite state transducer (FST) transition file shows the order ofsuffixation. The system gives more than 90% results for nouns and around 85% for verbs.

VI. TAMIL

In Dravidian Languages, the works in POS tagging were reported first in Tamil language, and more works were reported based on rule based approach. Works based on HMM, SVM and Multilingual POS Tagging also reported. The developed taggers are:

VasuRanganathan proposed a tagger named tagtamil that was built by implementing theprinciples of the theory of Lexical Phonology and Morphology and is tested with a number ofnatural language processing tasks[21]. The tagger is written inProlog is built with a knowledge baseconsisting of the rules of morphology of Tamil in a systematic manner in that the processingof input words takes place with suitable consultations of the knowledge base in successivestages. When a chunk of text is fed to this system, it processes individual sentences and produces asequence of lists containing information about every word. This system iscapable of recognizing and generating considerable number of Tamil word forms includingfinite and non-finite form of verbs such as aspectual forms, modal forms, tense forms besidesthe noun forms such as participial nouns, verbal nouns, case forms and so on.

Arulmozhi et al proposed a rule based POS tagger whichtries to find the POS of the rootword using the inflection of the word without usingany root word dictionary[22]. After splitting the sentences into words find out the suffixes. apply the lexical rules and assign the category. Apply the context sensitive rules on theunknown words and on thewrongly

tagged words. If again there is an unknown word tag it as noun. The tagset consists of 12 tags. The system gives a precision of 93.22%.

M Ganesan et al proposed a taggerbased on the morphologicalanalysis of these languages[23]. There are 34 tags at word level and 132 tags atmorpheme level. The system first identifies the valid morphin the word one by one and label them at morphemelevel then the entire word is tagged for its grammaticalcategory at word level. This system has three majorcomponents: Machine ReadableDictionaries(MRD) for Stem, Suffix and a set ofmorphophonemic rules.The system reads a word from the corpus andtries to match with those entries marked ID as statusin the stem – MRD. If fails, it tries to segment the last suffixand to match with suffixes, This procedure is continued till a stem isreached. If the system does not find a match in the last element itself, it tries to use the morphophonemicrules to revert the sandhi operation.

Arulmozhi et al proposed an HMM based tagger using Viterbi algorithm[24]. That is the tag forthe current word depends up on the previous word and its tag. Here in the statesequence the tags are considered as states and the transition from one state toanother state has a transition probability. The basic tag set including the inflection is 53. So, the tagset increases to 350, asthe combinations become high. The training corpus is tagged with thecombination of basic tags and tags for inflection of the word. The evaluationgave encouraging result.

S Lakshmana et al created a morpheme based language model for Tamil POS tagging[25].For categorizing the part of speech, this language model was based on the information of the stemtype, last morpheme, and previous to the last morpheme part ofthe word. The tagset consists of 35 tags. They identified 79 morpheme components, which can be combinedto form about 2,000 possible combinations of integratedsuffixes. The tagged corpus contain 4,70,910 words. The accuracy obtained is 95.92%.

M Selvam et al created a rule based POS tagger which makes an improvement on rule based POS tagging via Projection and Induction techniques[26]. Using alignment andPOS projection from English to Tamil, POS tagged sentences were Generated.Using POS projection and induction from English to Tamil, Improvement of rule basedPOS tagging were done.The rule based tagger gives an accuracy of 85.56%. Root words were inducedfrom English to Tamil through alignment, lemmatization andinduction processes. Using Categorical information , projection and alignment techniques POS tagged sentences in Tamil were obtained for theBible corpus. With this 7% improvement was achieved.

Dhanalakshmi et al proposed an SVM based POS tagger based on Linear Programming[27]. For preparing the annotated corpus they have developed the tagsetconsisting of 32 tags. The corpus was divided into a training set (15,000 sentences) and a test set (10,000 sentences). They considered a centered window of five tokens, fromwhich basic and n-gram patterns are evaluated to form binary features. Two previous tags are used as POS features. The suffix and prefix information are also considered. The obtained accuracy is 95.63%.

In an another attempt Dhanalakshmi et al proposed a POS tagger which is based on the same tagset but a corpus of size two hundred and twenty five thousand words was used[28]. The corpus is divided into training set (1, 65,000 words) and test set (60,000 words). The corpus is trained with the machine learning based SVMTool by tuning the parameters and feature patterns based on Tamil language. The accuracy was 95.64%.

MadhuRamanathan et al proposed a POS tagger which is based on a multilingual parallel corpora for Tamil, consists of otherlanguages namely Hindi, English andFrench[29]. Apply techniques such as HMM, SVM and CRF for tagging. The corpus used is Universal Human Rights Declarationcorpus (UDHR) and the tag set used is consisting of 12 tags. The dataset is splitted as 80% for training set and 20% for test set. They found that the addition of languages does not always produce an increase in accuracy of POS tagging and the use parallel corpus also leads to a dropin accuracy in many of the preprocessing stages.

VII. TELUGU

In telugu, works were reported based on rule based approach which uses Morphological Analyzer and rules for disambiguation , and works on Maximum Entropy and SVM were also reported. They are:

Sreeganesh implemented a rule based POS tagger[30]. In the first stage take a text or corpus,send the text to Telugu Morphological Analyzer, take the output of morphological analyser, , In this stage add all the possibilities of different POS Tags to theOriginal Text and524Morpho-syntactic Rules willdisambiguate the ambiguity in POS.

Three Telugu POS taggers were developed by RamaSree et al, based on Rule-based, Brill and Maximum Entropy based approaches[31]. An annotated corpus of 12000 words is used to train the Brilland Maximum Entropy taggers.

i) Telugu Rule-based POS tagger consists of a series of modules such as Sentence tokenizer, Telugu morphological analyzer ,Morph to POS translator, POS disambiguator, and Annotator . Tokenizer segregate the input text into a series of sentences and each sentence into words. Morphological analyzer give all possible analyses of each word. Morph to POS translator converts all the morphological analyses into their corresponding POS tags using some pattern rules. POS disambiguator reduces the above POS ambiguity for each word by the application of unigram and bigram rules.Annotator produces the tagged text. The accuracy of the system was 98%.

ii)Brill's tagger-There are three main phases in implementing Brill Tagger Training, Verification and Testing. The accuracy

was 92%.

iii)Maximum Entropy tagger-The tagger was implemented usingthe Maximum Entropy Modeling toolkit [MxEnTk] freelyavailable on the net. The accuracy of the system was 87%.

Another work was by SindhiyaBinulal et al who applied SVMToolfor POStagging[32].Pos tagging can be seen asmulticlass classification problem. Thetagsetconsists of 10 tags. The training corpusconsists of 25000 words. The corpus is divided into training set (20,000 words)and test set (5,000 words). The obtained accuracy is around 95%,for unknown words accuracy was 86.25%.

Recent work has been by Srinivasu who present a morphological based automatictagging for Telugu without requiring any machine learningalgorithm or training data[33]. The pre-processing moduleperforms several useful tasks but the most important taskis to identify words correctly. The lexicon assigns tags towords that appear without any morphologicalinflection. Morphology handles all the derived andinflected words, including many forms of sandhi. Thebridge module combines the tags given by the dictionaryand the additional information given by the morph,making suitable changes to reflect the correct structureand meaning where required.

CONCLUSION

This paper made a detailed study about POS Taggers and Taggers developed on Dravidian Languages – Malayalam, Tamil, Kannada and Telugu. Rule based, Stochastic and Hybrid approaches were used to build the taggers. Seen that rule based approaches suit for Dravidian Languages because these are morphologically rich. Deep linguistic study is needed to infer rules and rule based method is language dependent. Stochastic methods require large corpora and language independent and in Dravidian languages HMM, SVM, CRF and Maximum Entropy were used. Found that SVM works better than HMM, Maximum Entropy and CRF also outperforms HMM. In the literature works were reported first in Tamil and diverse methods were implemented in Kannada. Rule based and stochastic methods were reported in Telugu and in Malayalam works on HMM and SVM were reported, and SVM performed better than HMM. The size of the corpus play a major role in the accuracy of the POS taggers, and large annotated corpora are required for implementing these techniques. The unavailability of it creates barriers in developing full fledged POS taggers. Researchers create small corporafor their own.There are large annotated corpora in English like British National Corpus, Brown Corpus etc. Corpora are available in some Indian Languages also. Standardization of tag set is needed because the researchers creates their own tagset and the comparison of the efficiency of the POS taggers become difficult.There is a lot to do in Dravidian languages in case of part-of-speech tagging.

REFERENCES

[1]TanveerSiddiqui , U S Tiwary, *Natural Language Processing and Information Retrieva*l, Oxford University Press.

[2]http://www.britannica.com/EBchecked/topic/171083/Dravidian-languages

[3] http://www.gktoday.in/classical-languages-of-india/

[4]Jurafsky, D., Martin, J. H., *Speech and Language Processing-An Introduction to Natural LanguageProcessing, Computational Linguistics andSpeech Recognition*, Prentice Hall, Upper Saddle River, NewJersey,2000

[5]L. R. .Rabiner, B. H. Juang, *An Introduction to Hidden Markov Models*,IEEE ASSP MAGAZINE JANUARY 1986

[6]Chris Mueller, *Support Vector Machines*, www.osl.iu.edu/~chemuell/projects/presentations/svm.pdf

[7] Marti A Hearst,*Support Vector Machines*, IEEE Intelligent Systems Magazine, 1998.

[8] maxent.sourceforge.net/about.html

[9] AdwaitRatnaparkhi, *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*, IRCS Report 97.

[10] Adam L Berger, Stephen A Della Pietra, Vincent J Della Pietra, *A Maximum Entropy Approach to Natural Language Processing*,Journal Computational Linguistics, Volume 22, Issue 1, March 1996.

[11] John Lafferty, Andrew McCallum, Fernando Pereira, "Conditional Random Fields: Probabilistic Modelsfor Segmenting and Labeling Sequence Data,"Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pages282-289.

[12] Manju K., Soumya S., Sumam Mary Idicula, *Developmentof a POS Tagger for Malayalam - An Experience*,artcom, pp.709-713, 2009 International Conference on Advances inRecent Technologies in Communication and Computing,2009

[13]Rajeev R R, Jisha P jayan, Elizabeth Sherly, *Parts of speech Tagger forMalayalam*, IJCSIT International Journal of Computer Science and Information Technology, vol 2, No.2, December 2009, pp 209-213

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

[15]http://www.ldcil.org/workInProgress.aspx

[16] P J Antony, K P Soman, *Kernel based part of speech tagger for Kannada*, International Conference on Machine Learning and Cybernetics (ICMLC), 2010 (Volume:4)

[17] Shambhavi.B. R,Ramakanth Kumar P Revanth G, *A Maximum EntropyApproach to Kannada Part Of Speech Tagging*, International Journal of Computer Applications (0975 – 8887), Volume 41– No.13, March 2012

[18]Shambhavi B R,Ramakanth Kumar P, *Kannada Part-Of-Speech Taggingwith Probabilistic Classifiers*, International Journal of Computer Applications (0975 – 888) Volume 48– No.17, June 2012

[19]Pallavi, Anitha S Pillai, *Parts Of Speech (POS) Tagger for Kannada UsingConditional Random Fields (CRFs)*, National Conference on IndianLanguage Computing, 2014

[20] Bhuvaneshwari C. Melinamath, *Hierarchical Annotator System For Kannada Language*, Impact: International Journal of Research in Engineering & Technology Vol. 2, Issue 5, May 2014, 97-110

[21] VasuRanganathan, *Development of Morphological Tagger for Tamil*, Tamil Internet Conference 2001

[22] Arulmozhi. P, Sobha. L, Kumara Shanmugam. B, *Parts of Speech Tagger for Tamil*, Symposium on Indian Morphology, Phonology& Language Engineering19 – 21 March, 2004

[23]M Ganesan, S. Raja, *Morpheme and Parts-of-Speech tagging ofTamil Corpus*, Symposium on Indian Morphology, Phonology& Language Engineering19 – 21 March, 2004

[24] ArulmozhiPalanisamy,SobhaLalitha Devi, *HMM based POS Tagger fora Relatively Free Word Order Language*,Research in Computing Science,2006, pp. 37-48

[25] S. LakshmanaPandian, T. V. Geetha, *Morpheme based Language Model for Tamil Part-of-Speech Tagging*,Research journal on Computer science and computer engineering with applications Issue 38 (July-December 2008)pp 19-25

[26]M. Selvam, A.M. Natarajan,*Improvement Of Rule Based Morphological*

Analysis AndPos Tagging In Tamil Language Via Projection And Induction Techniques, International Journal Of Computers Issue 4, Volume 3, 2009

[27] Dhanalakshmi V, Anand Kumar, Shivapratap G, Soman KP,Rajendran S, *Tamil POS Tagging using Linear Programming*,International Journal ofRecent Trends in Engineering, Vol. 1, No. 2, May 2009
[28]Dhanalakshmi V, Anandkumar M, Rajendran S, Soman K P, *POS Taggerand Chunker for Tamil Language*, Tamil Internet Conference 2009
[29] MadhuRamanathan, Vijay Chidambaram, AshishPatro, *An Attempt at Multilingual POS Tagging for Tamil*, pages.cs.wisc.edu/~madhurm/CS769_final_report.pdf
[30]T. Sree Ganesh,*Telugu Parts Of Speech Tagging In WSD*, Language In India, Volume 6 : 8 August 2006
[31]RamaSree, R.J., KusumaKumari P., *Combining Pos Taggers For Improved Accuracy To Create Telugu Annotated Texts For Information Retrieval*,www.Ulib.Org/Conference/2007/Ramasree.pdf
[32]G.SindhiyaBinulal, P. AnandGoud, K.P.Soman, *A SVM based Approachto Telugu Parts Of SpeechTagging using SVMTool*, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009
[33] SrinivasuBadugu, *Morphology Based POS Tagging on Telugu*,IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, No 1, January 2014.