ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

High Order Conditional Random Field Based Part of Speech Taggar and Ambiguity Resolver for Malayalam -a Highly Agglutinative Language

Bindu.M.S*
Dept. of Computer Science
M.G University College of Engineering
Muttom, Thodupuzha, India
bindu_aisoo@rediffmail.com

Sumam Mary Idicula
Dept.of Computer Science
CUSAT
Cochi, India
sumam@cusat.ac.in

Abstract: Parts of speech tagging also called grammatical tagging assign lexical class markers to each and every word in a document. It is an essential and important preprocessing step in many NLP systems. Tagged corpora play a significant role in Machine Translation, Information Retrieval, and Data Mining. POS tagging in Malayalam is a difficult task as it is an agglutinative language and 80-85% of words in Malayalam text documents are compound words. Decomposition of these words into its constituents is extremely necessary for finalizing the POS tag of these words. Sometimes more than one morphological analysis and hence more than one POS may occur for a single word. A correct resolution of this kind of ambiguity for each occurrence of the word is crucial in many NLP applications. Currently available tag sets in other languages are only giving importance to the morphological and syntactical properties of the language while the tag set designed by us considers the semantic features of the language. For testing this system, documents from well known Malayalam news papers and magazines are selected. Up to 2352 sentences are tested which includes simple, complex and compound type sentences. Word level tagging accuracy of 95% and sentence level accuracy of 91% are obtained.

Keywords: POS Tag set, finite state transducer, compound word splitter, Extended CRF, Malayalam compound word

I. INTRODUCTION

Natural Language Processing is a subfield of Artificial Intelligence and is concerned with the interaction between the computer and human. It is also related to computational linguistics [1]. To process a language, knowledge of the language at various levels-morphological, syntactical, semantical- is essential. NLP systems must be able to analyze, understand and generate languages humans use naturally. There are various applications to NLP such as Machine Translation, Information Retrieval, and Question Answering Systems [2].

Since 100 B.C humans are aware that language consists of several distinct parts called POS. Those POS play a crucial role in many fields of linguistics. POS is based on both its definition and its context or relationship with adjacent and related words in a phrase, sentence or paragraph. Parts of speech tagging is harder because some words can represent more than one POS at different times [3]. The significance of parts of speech for language processing is the large amount of information they give about a word and its neighbor. POS tagging can be used in text to speech, information retrieval, shallow parsing, and information extraction linguistic research.

POS Tagging also called grammatical tagging is a principal issue in natural language processing. The purpose of this task is to assign part-of-speech or other lexical class markers to each and every word in a document. In English there are eight part of speech such as noun, verb etc. But POS tags vary according to the language and application [4].

Malayalam belongs to the Dravidian family of languages and is one of the 4 major languages of this family. It is one of the 22 scheduled languages of India with official language status in the state of Kerala. It is spoken by 35.9 million people [5]. Malayalam is a morphologically rich agglutinative language and relatively of free order. Also Malayalam has a productive morphology that allows the creation of complex words which are often highly ambiguous. Hence POS tagging involves solving the complexity and ambiguity of the words. Many words that occur in Malayalam texts are compound words which are not listed in any catalog or lexicon. A large percentage of words also show ambiguity regarding lexical category. In many other languages, POS taggers have used tag set derived from Penn Tree Bank or Brown corpus. But this is not enough for Malayalam POS tagger developed for the purpose of information Retrieval. We needed a tag set which also considers semantics in its development. Tagging is a difficult task in Malayalam due to the complexity of sentences, complexity of the words and inappropriate word order of the sentences.

Conditional Random Field (CRF) is a probabilistic framework for labeling or segmenting data. It is a form of undirected graphical model in which each edge represent conditional dependencies between random variables at the nodes. Each random variable Y_i is conditioned on an input sequence X. The conditional dependency of the random variable on X is normally represented by some feature functions [6] [7]. This feature function varies according to the application. CRF is commonly used for the labeling of natural language text or biological sequences. They were first used for the task of shallow parsing by Lafterly et al (2001) where CRF were mainly applied for Noun Phrase (NP) chunking. In CRF, with respect to figure 5. Y is dependant only on X while high order CRF or Extended CRF represent a model in which each Y_i is dependants on X as well as on n number of previous variables Y_{i-n}, …., Y_{i-1}

Regular Expression is the standard notation for characterising text sequences. Finite State Automata (FSA) is a mathematical device used for implementing texts represented by regular expression. A variation of FSA called a Finite State Transducer (FST) is a machine that reads a string and outputs another string.Formally an FST is represented by a 6-tuple [8]. FST's applications are in speech recognition,phrase chunking,POS tagging etc

II. RELATED WORK

A lot of work has been done in parts of speech tagging of western languages. These taggers mostly are implemented using stochastic or rule based methods. They vary in accuracy and also in their implementation. Papers [9][10] are examples of rule based approaches. Genetic algorithms for POS is described in [11].This system is able to integrate statistical and rule based approaches into one system.

In Indian Languages natural language processing tools are very less as compared to English and other European languages. A rule based POS tagger for Hindi language is developed at IIT, Bombay which has been used for the word net project. Different taggers using Support Vector Machines (SVM), Hidden Markov Model (HMM) and Maximum Entropy are available for Tamil Language. A Punjabi spell checker has been developed using Rule cum Dictionary based method.

When we developed this system no known work was available for Malayalam. Currently some work is going on at Amrita, Coimbatore and at Centre for Development of Advanced Computing (C-DAC), Trivandrum.

III. SURVEY OF MALAYALAM MORPHOLOGY AND GRAMMAR

Morphology is the study of the way words are built up from smaller meaning bearing units called morphemes. Generally morphemes are classified as stems and affixes. Stem is the main morpheme of the word, supplying the main meaning while the affixes add additional meaning of various kinds. Affixes are further divided into prefixes, suffixes; infixes and circumfixes.There are two broad ways to form words from morphemes: inflection and derivation [12].

A very productive word formation process in Malayalam is compounding which combines simple words to build more complex words. Unlike English, Malayalam does not contain spaces or other word boundaries between the constituents of the compound word. Compound words selected from various corpuses are analyzed and identified 10 possible component types which are taken as basic atomic words. They are mainly nouns, verb, adjective, adverb, qualifier, prefix and suffix. These atoms are classified based on their functionalities and there are simple rules for forming legal words from them. A single word in Malayalam may consist of an arbitary number of prefixes, stems (nouns, verbs, pronouns etc.) and arbitary number of suffixes.

A. Noun Types:

Noun is classified into material noun and abstract noun. The subclasses of material nouns are proper nouns, common nouns, pronouns and collective nouns. Abstract noun is further classified into quality nouns and verbal nouns. Pronouns are available in twelve different forms.

B. Verb Types:

Verbs are divided into four categories based on their meaning, behavior, feature and importance. The first type is divided into transitive and intransitive. Another classification based on the behavior is simple verbs and causatives. Third type is classified into strong and weak verbs. Last division is according to its importance and is named as finite and infinite verbs.

C. Qualifiers:

Three types of qualifiers are there in Malayalam- qualifiers of nouns (adjective), qualifiers of verbs (adverb) and qualifiers of qualifiers.

D. Dhyodhakam:

These words are classified into prepositions, conjunctions and interjections.

E. Affixes:

Prefixes are used to obtain a subdam from a root word with changed meanings. Sometimes new subdam might have an entirely different or opposite meaning. There are three types of prefixes [10].First type- with opposite meaning, Second type – same meaning but with emphasis and third type –same meaning. Postfix is mainly used for completing or changing the meaning of verbs. They are of four types. Words in Malayalam have a strong inflectional component. For verbs these inflections carry information on the tense, mood, aspect etc. For nouns and pronouns inflections distinguish the categories of gender, number and case. These inflections are called Suffixes.

IV. POS TAG SET FOR MALAYALAM

A tag set with 52 tags is developed by manually tagging different documents from various news papers and various fields. This tag set contains all POS necessary for Information Retrieval task. The method used for finding the POS and few examples are given below.

A. Noun:

Nouns without suffixes and nouns with gender or plural suffixes are considered as 'NOUN'. The Penn tag set –a standard tag set used for English- makes distinction between noun singular ,noun plural ,common nouns ,proper nouns etc. In this system, nouns which are acting as agents are given the POS, NOUNS.

B. Noun with case suffix:

If a noun occurs in a sentence with case suffix, POS tag of that noun is determined by the associated case suffix. Here the case Suffix changes the role of the word in the sentence. Case indicates a relation between noun and a verb [13]. Relation indicated by each case suffix is different. Hence we have assigned separate POS to each unique case which is shown in table 1.

Table 1 Example of Case Relations

Case Relation	POS Tag
Nominative	NOUN
Accusative	ACC
Sociative	SOC
Dative	DAT
Instrumental	INST
Possessive	GEN
Locative	LOC

C. Verbs:

Finite verbs are tagged with VERB tag. Infinite verbs are divided into adjectival participle and adverbial participle. Adjectival participle relies on nouns and adverbial participle on verbs. All the three forms function distinctly.

Table 2: POS Tags

Tag	Description	Tag	Description
NOUN	Noun	Adv	Adverbial participle
ACC	Accusative	Adv	Adverb
DAT	Dative	Adj	Adjective
GEN	Genitive	Adv	Adverbial clause of Time
PSP1	Postposition	Adv	Adverbial clause of
VERB	verb	Vau	Auxiliary Verb

D. Preposition:

This is a word used along with a noun or a pronoun to show how it is related to something else. Each preposition is assigned a different POS as they serve different roles.'PRP1' TO 'PRP14' are the tags assigned to different prepositions. Few POS tags are shown in the table 2.

E. Adjectives :

An adjective is a word that adds to the meaning of the noun. It is working as a qualifier of the noun. There are different kinds of adjectives. Adjective indicates quality, number etc. The POS tag varies according to the type of adjective. Example : AdjQt,AdjN, AdjQny

F. Adverb:

An adverb adds something to the meaning of a verb, adjective or another adverb. Adverb indicates time, frequency, place, manner, condition quantity cause, reason etc. POS tags corresponding to these are 'AdvT','AdvC', 'AdvR'

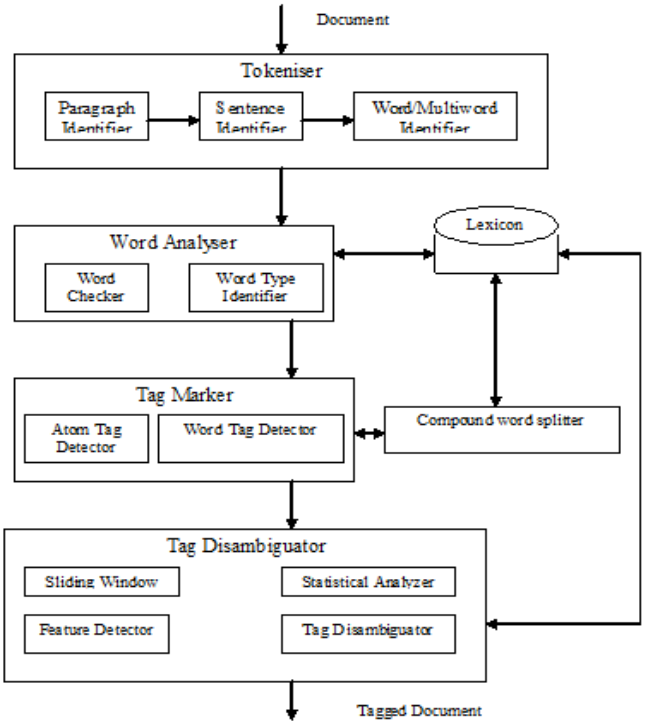

Figure. 1 Block Diagram of POS Tagger

V. PART OF SPEECH TAGGER FOR MALAYALAM

POS tagging is identifying and labeling each word in a sentence with corresponding POS. For English and many other languages they have specific structure for a sentence. Therefore POS assignment can be done using the grammar rules. Malayalam is a free order language; hence words can appear in any order in a sentence. Also 90% of Malayalam words are compound words. But within a phrase, words are in a related order. To determine the POS of a word in Malayalam language, both word level and contextual information is essential.

Fig.1 shows the block diagram of the POS tagger which is developed. Working of this tagger is as follows. The input document is divided into tokens. Then each token is sent to the word analyzer for the detailed analysis. First it checks the token to decide whether it is a compound word or not. If it is a simple word the local information is collected from the lexicon. Else the token is sent to the compound word splitter to find out the morphological details and the constituents of it. The tagger assigns to each token, all possible POS tags with the help of local information provided by the word analyzer. Then to resolve the ambiguity, Extended CRF model is used with contextual information and eliminates all but one tag.

A. Tokenizer:

Input to the Tokenizer block in Fig 1 is a document in Malayalam. During the tokenization process each sentence of the document is taken and split into words or co-occurrence patterns. Multi part words are one of the main issues arise while tagging. We consider these words as single words to keep the semantic information intact for the purpose of efficient information extraction.

B. Word Analyser:

Word Analyzer checks each token to see whether it is present in the lexicon or not. Lexicon has all the root words along with its POS information. If it is present in the lexicon then it is a simple word, then the word is labeled with POS details retrieved from the lexicon. Else the token is a compound word and it is labeled with <CW> tag.

C. Compound Word Splitter:

If the output of the word analyser is <CW> then the corresponding token M is a compound word and it is to be decomposed into its constituents M_1 to Mi.To find each constituent, the longest match method is adopted. When one component Mi is separated the remaining portion is sent to modification algorithm. The component M is searched in the lexicon if it is not found transformation algorithm is called to obtain various forms of M and again searching is carried out. If not found process is repeated with next smaller string M [14].

a. Methodology: Finite State Transducer (FST):

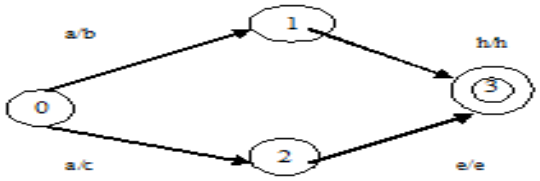

Figure 2 Finite State Transducer

Formally, a finite transducer T is a 6-tuple (Q, Σ, Γ, I, F, δ) such that:

a) Q is a finite set, the set of states;
b) Σ is a finite set, called the input alphabet;
c) Γ is a finite set, called the output alphabet;
d) I is a subset of Q, the set of initial states;
e) F is a subset of Q, the set of final states; and δ is the transition relation.

Representation of the transducer in Fig.2 is

T=({0,1,2,3},{a,b,c,h,e},0,{3},{0,a,b,1},{0,a,c,2},{1,h,h,3},{2,e,e,3})

FST is a machine which accepts a string and translates it into another string. FST can also be used for generating and checking sequences. A compound word is a string of Malayalam characters. To split this string into substrings an FST can be used.

Compound word splitter is an FST with the following definition.

In the 6-tuple, set of states Q = {A,B,C,D,E,F,G,H}
Initial state I= {A}
Final states F= {C,D,E,F,G,H}
Input alphabet Σ = {compound words}
Output alphabet Γ = { valid simple Malayalam words}
Transition function δ= {NOUN, VERB, ADJECTIVE….,SUFFIX}

FST for compound word splitter is given in FIG.3

This system operates in optimal time since the time to assign the tag to sentence corresponds to the time required to follow a single path in a deterministic finite state machine.

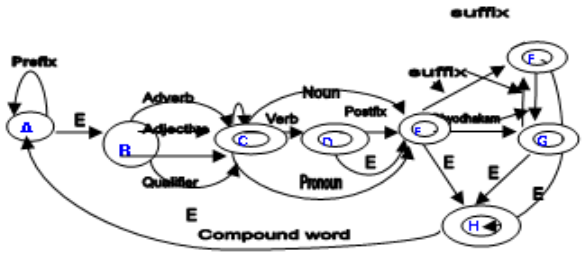

Figure.3 FST for Compound Word Splitter

aasUthraNavaibhavaTHOTukUTiya (planning capacity)

This Malayalam word is a combination of 5 atoms of different categories. The compound word splitter splits this word into five components. Fig.3 explains the working of compound word splitter.

aasUthraNam+vaibhavam+Out+kUTi+ a
(Verb+noun+suffix+verb+dhyodhakam)

Traversal Path- A-B-C-D-E-H-A-B-C-E-F-H-A-B-C-D-E-G-H

D. Tag Marker:

Using the information provided by the word analyzer tag Marker marks the token with the most appropriate POS tag. It is designed with FST. Tokens marked with <CW> tags are sent to the compound word splitter one by one and then receives the constituents of the words along with their POS information. Next the Tag Marker assigns suitable tags to each compound word based on the constituents.

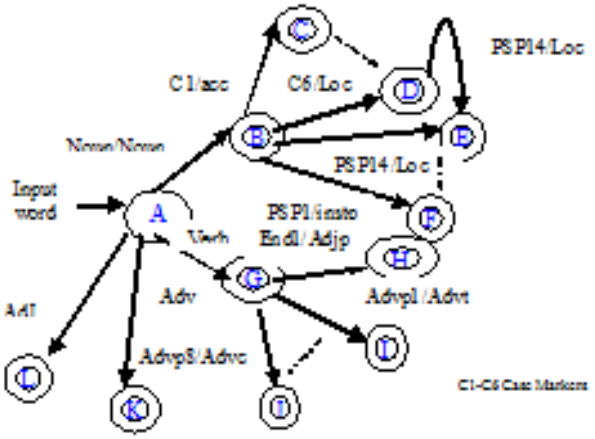

Figure.4 FST for POS Labeling

Sometimes there will be different valid decompositions possible for a compound word and in such cases word will be marked with multiple tags. FST is used for the implementation of the Tag Marker. For this FST following elements form the tuple.

Initial state – A
Final states-{B,C,…Z}
Input Alphabet- {NOUN, VERB…, DHYODHAKAM}
Output alphabet- {Any POS from the tag set}
Transition function-{any valid POS}

Fig4 explains the working of the tag Marker. For the word nagaraThilekku(To city) Compound word splitter produces the following output.

nagaram<NOUN> iL<C6> ekku<PSP14>

Then the FST for tagger (Figure.4) takes <NOUN> <C6> <PSP14> as the input string and by traversing the path A-B-D-E produces the output 'LOC'.ie The POS for the Malayalam word nagaraThilekku is 'LOC'.

E. Tag Disambiguator:

Tag Marker block assigns each input token with a POS tag or multiple tags. Tokens with multiple tags are sent to the Disambiguator to solve the tag ambiguity which removes all tags except one. Output of tag Disambiguator is a string of all tokens along with their POS tags.

a. Methodology:

Tag Disambiguator is implemented using high order CRF or extended CRF. It is an undirected graphical model in which each vertex represents a random variable whose probability distribution is to be inferred and each edge represents a dependency between two random variables. CRF's avoid the label bias problem, a weakness exhibited by MEMM. The primary advantage of CRF's over HMM is their conditional nature [6].

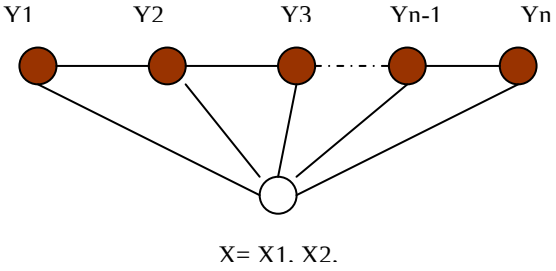

Figure: 5 Graphical structure of chain-structured CRFs

Let X={X1 …XN} and Y= {Y1…YN} be two sets of random fields. For the given input sequence X, Y represents a hidden state variable and CRF's define conditional probability distributions P (Y|X) over the input sequence. Sometimes the conditional dependency of each Yi on X will be defined through a fixed set of feature functions (potential functions) of the form f (i, Yi-1, Yi, X). The model assigns each feature a numerical weight and combines them to determine the probability of a certain value for Yi. CRF's can contain any number of feature functions and the feature function can inspect the entire input sequence X at any point during inference. CRF's are extended into high order models by making each Yi dependant on a fixed number of previous variables Yi-o ... Yi-1.

POS tagging can be modeled as a sequence labeling task where X=X1X2X3...Xn represents an input sequence of words and Y= Y1Y2Y3...Yn represents corresponding POS label sequence. The general label sequence Y has the highest probability of occurrence for the word sequence X among all possible label sequences that is Y = argmax {Pr (Y|X)}. These labels are determined by the feature functions.

Main features for POS tagging have been identified based on the word combination and word context. The features also include prefix and suffix for all words [15].Following are the features used for POS tagging in Malayalam.

a) Constituents of current word: These determines the POS tag of the word as noun, verb etc.

b) Context word features: Preceding (pw) and following words (nw) of the current word. We have taken pw1, pw2, pw3, nw1, nw2, nw3 as the feature.

c) POS information: POS of previous words and in ambiguity resolution, POS of the following words are helpful.

d) Contains digits or symbols. If the word contains digits they are marked with 'NUMBER' POS (cardinal number(CN) or ordinal number(ON))

e) Lexicon feature: It contains Malayalam root words and their basic POS information such as noun, verb, adjective, adverb etc.

f) Inflection lists: After analyzing various classes of words inflection lists of nouns, verbs and participles are prepared to improve the performance of the POS tagger.

For the Disambiguator, input sentence represents X and corresponding POS 'S represent Y. According to the principles of CRF, each POS Yi is dependant on corresponding word of the sentence X. But in Malayalam language Yi is dependant on Xi and other features mentioned above.

VI. RESULTS AND DISCUSSION

The performance of the POS tagger is carried out using the standardized formulation techniques precision and recall where precision is defined as the ratio of correct number of token tag pair sequence in the output to Total number of token tag pair appear in the output and recall is the ratio of correct number of token tag pair sequence to the Number of correct token tag pair that is possible [16].

Documents from Malayalam dailies Malayala Manorama, Mathrubhumi, Karshaka Sree, and few Text Books pertaining to five different fields are selected as test corpus. These documents are texts with simple, compound and complex sentences and 95% words were compound words. Totally 2352 sentences are tested and obtained average precision of about 92% and recall of 95%. Some words are not correctly tagged since all features are not included in the system.

A. Our approach is more accurate and efficient for a language like Malayalam since it considers language level and word level features.

B. Malayalam language has no specific structure for a sentence. Hence it is very difficult to assign POS tag for a word; many times a word will have multiple tags. This problem is solved in our system by taking into account the contextual information.

C. A POS tag set is developed which is unique in its nature as it reflects morphological, syntactical, in addition semantic features of the language.

VII. CONCLUSION

Natural language processing is dealing with speech and language processing which aids in communication between human and computer. For effective communication computer

must learn the human language at the basic levels of POS, syntax, semantics etc.

For Malayalam language no known work or tag set was available. Also tag sets available in other languages are insufficient for our purpose since they have not considered semantics in their implementation. Malayalam is a compounding language without specific word order. Compounding feature and tag disambiguation is given special importance during the development of this system.

We have presented an approach which is suited for using it as a method for preparing text information for Named Entity Recognition and thereby for Information Retrieval scenarios. We have achieved good POS tagging results for Malayalam, a fairly wide spoken language which had very little prior computational linguistic work.

Part of Speech Tagging is a difficult task in a compounding language like Malayalam. This issue is solved by implementing a compound word splitter tool employing Finite state models. Our main achievement is the design and implementation of a tagger and language specific POS tag set for a highly agglutinative, morphologically rich language. This tagger gives good results as it considers the local and global features of each word for its design.

VIII.REFERENCES

[1] Stefan Schwarzler Joachim Schenk,Frank Wallhoff and Gunther Ruske,"Natural Language Understanding by Combining Statistical methods and Extended Control Free Grammars",Proceedings of 30th DAGM Symposium on Pattern Recognition,Springer-Verlag Berlin,Heidelberg,2008.

[2] Noriko Nagata,"An Effective Application of NLP in 2nd Language Instruction",CALICO Journal ,vol.13,No.1 pp 47-67.

[3] Navanath Sabaria, Dhrubajyothi Das,Utpal Sharma,Jugal Kalita,"Part of Speech Taggar for Assamese Text", Proceedings of ACL-IJCNLP 2009 Conference pp 33-36.

[4] Isabelle Tellier ,Iris Eshkol Samer Taalab and Jean –Philippe Prost,"POS tagging for Oral text with CRF and Category Decomposition", 11th International Conference on Intelligent Text Processing and Computational Linguistics Romania-2010.

[5] A.R ajarajavarma,"Keralapanineeyam", National Book Stall, Kottayam, 2000.

[6] Hanna.M.Wallach,"Conditional Random Fields",University of Pennsylvania CIS Technical Report MS-CIS-04-21.

[7] Chirag Patel and Karthik Gali,"POS tagging for Gujarathi using CRF",Proceedings of the IJCNLP-08 Workshop on NLP for lLess Privileged Languages pp 117-122.

[8] Daniel Jurafsky and James H Martin,"Speech and Language Processing",AI Pearson Education Series in AI.,First Indian Print 2002.

[9] Eric Brill,"A Simple Rule Based POS Taggar", Proceedings of the third Conference on Applied Computational Linguistics, Torento, Italy, 1992.

[10] Ihsan Rabbi,Mohammed Abid Khan and Rahman Ali,"Rule Based POS Tagging for poshto languages", Proceedings of the Conference on Langage and Technology

[11] K.T Lua, "POS Tagging of Chinese Sentences using Genetic Algorithms", Conference on Chinese Computing 1996 4-7 June National University of Singapore pp 45-49.

[12] Prof. K.S.Narayana Pillai, "Adhunika Malayala Vyakaranam", Kerala Bhasha Institute, Thiruvananthapuram.

[13] Dr.C.K.Chandrasekharan Nair,"Adisthana Vyakaranam", Kerala Bhasha Institute, Thiruvananthapuram, 1997

[14] Bindu.M.S, Sumam Mary Idicula,"Analysis of Malayalam compound words and Implementation of a compound word splitter tool using Finite State Models", International Conference on Modeling and Simulation India 1-3 December 2009.

[15] Asit Ekbal ,Rejwanul Haque,Sivaji Bandyopadhyay,"Maximum Entropy Based Bengali Part of Speech Tagging ", Advances in Natural Language Processing and Applications, Research in Computer science 2008,pp 67-78.

[16] Ghassan Kannan,Riyad-al-Shalabi and Majdi Sawalha,"Improving Arabic Information Retrieval System Using Part of Speech Tagging",Information Technology Journal 4(1) 32-37,2005