A NEW APPROACH TO PARTS OF SPEECH TAGGING IN MALAYALAM

D. Muhammad Noorul Mubarak[1], Sareesh Madhu[2], S A Shanavas[2]

[1]Department of Computer Science, University of Kerala, India
[2]Department of Linguistics, University of Kerala, India

ABSTRACT

Parts-of-speech tagging is the process of labeling each word in a sentence. A tag mentions the word's usage in the sentence. Usually, these tags indicate syntactic classification like noun or verb, and sometimes include additional information, with case markers (number, gender etc) and tense markers. A large number of current language processing systems use a parts-of-speech tagger for pre-processing.

There are mainly two approaches usually followed in Parts of Speech Tagging. Those are Rule based Approach and Stochastic Approach. Rule based Approach use predefined handwritten rules. This is the oldest approach and it use lexicon or dictionary for reference. Stochastic Approach use probabilistic and statistical information to assign tag to words. It use large corpus, so that Time complexity and Space complexity is high whereas Rule base approach has less complexity for both Time and Space. Stochastic Approach is the widely used one nowadays because of its accuracy.

Malayalam is a Dravidian family of languages, inflectional with suffixes with the root word forms. The currently used Algorithms are efficient Machine Learning Algorithms but these are not built for Malayalam. So it affects the accuracy of the result of Malayalam POS Tagging.

My proposed Approach use Dictionary entries along with adjacent tag information. This algorithm use Multithreaded Technology. Here tagging done with the probability of the occurrence of the sentence structure along with the dictionary entry.

KEYWORDS

NLP, POS tagger, Rule based approach, Stochastic approach, Multithreading, Dictionary entry, Malayalam.

1. INTRODUCTION

Parts-of-speech tagging is simply a grammatical tagging. In natural language, there are a small number parts of speech those are noun, adjective, adverb, verb, Conjunction, preposition etc.). Words are grouped into different classes mentioned above. This type of grouping is called Parts of Speech Tagging. Apart from the natural language, automated models have more numbers of POS tags. Parts-of-speech tagging is the process of assigning a label to each and every word in a corpus. There are two approaches used for automated Parts of Speech Tagging namely Rule based and Stochastic. Rule based Approach follows predefined rule set whereas Stochastic Approach follows Statistical measures. For this, Stochastic Approach use Machine Learning Algorithms like Hidden Markov Model, Support Vector Machine etc.

Malayalam is a Dravidian family of languages, inflectional with suffixes with stem word. The above mentioned Algorithms are efficient Machine Learning Algorithms but these are not built for Malayalam. So it affects the accuracy of the result of Malayalam POS Tagging.

DOI:10.5121/ijcsit.2015.7509

This is an attempt to develop a new Algorithm that is specially designed for Malayalam POS Tagging. This follows hybrid Approach, i.e. it use rule set of Malayalam with lexical dictionary and Statistical measures based on the tagging structure.

2. TAGGING APPROACHES

There are two approaches in POS tagging: rule-based approach and stochastic approach. Rule-based taggers use a large database of words with predefined disambiguation rule. For example, that an ambiguous word is a noun or a verb. Stochastic taggers use training corpus along with statistical algorithms to check for the ambiguity

RULE-BASED PARTS-OF-SPEECH TAGGING

Rule based approach is one of the oldest approach in tagging. It uses predefined rules. The earliest algorithms for automatically assigning parts-of-speech were based on a two-stage architecture. The first stage uses a dictionary to label word. The second stage use large lists of predefined disambiguation rules to labelling word accurately .Rule based taggers use morphological information along with predefined rule to assigns tags to unknown or ambiguous words.

Advantages of Rule Based Taggers:-

a. use simple rules.
b. less storage.

Drawbacks of Rule Based Taggers:-

a. Generally less accurate as compared to stochastic taggers.

STOCHASTIC PARTS-OF-SPEECH TAGGING

Stochastic Approach use probability and statistical information for assigning tag to words. This approach use statistical algorithms rather than grammar rule.

Advantages of Stochastic Parts of Speech Taggers:-

a. Generally more accurate as compared to rule based taggers.

Drawbacks of Stochastic Parts of Speech Taggers:-

a. Relatively complex.
b. Require vast amounts of stored information.

Stochastic taggers are popularly used as compared to rule based taggers because of their higher degree of accuracy. On the other hand, this high amount of accuracy is achieved using some relatively complex procedures and data structures.

3. OBJECTIVE

To develop a high accurate and low complexity (both time and space) Parts of Speech Tagger for Malayalam. A Graphical User Interface is developed for entering input text in Malayalam and also it shows tagged output in Malayalam. The input texts tokenize and analyzing morphologically based on some rules and context, then tagged based on the lexical database.

4. METHODOLOGIES

Parts of Speech Tagging developed here based on a new approach, we can termed it as hybrid approach There are two types of POS tagging approaches-: Rule based and Stochastic based. First one use predefined rule set and second one use probability measures. Compromising with both approaches' merits and demerits, we develop POS tagging algorithm that use lexicon dictionary and structure of the sentence for evaluation or labeling the words.
The GUI designed using Netbeans(Java).

5. LITERATURE REVIEW

For language processing applications like Parser and Chunker, tagger with highest possible accuracy is required. Usually Statistical approach gives better accuracy compared to Rule based Approach. Parts-of-speech tagging is also a very practical application, with uses in many areas, including machine translation, parsing, information retrieval and lexicography. Initially people engineered rule for tagging, sometimes with the aid of a corpus. Later, with the aid of large corpus, Markov-model based stochastic taggers that were trained automatically gives highly accurate tagging result.

Recently, a number of approaches to POS tagging based on statistical and machine learning techniques are applied, including among many others like Hidden Markov Models, and Support Vector Machines. Many natural language tasks require the accurate assignment of Parts-Of-Speech (POS) tags to previously unseen text. Due to the accessibility of huge corpus which has been annotated manually with POS label, many taggers use annotated text to learn probability rules and use them to tag without human intervention to unseen text.

6. OVER VIEW OF MALAYALAM LANGUAGE

Malayalam, the language categorized under the family of Dravidian languages spoken in Kerala. This language uses around 37 million people. Liilaatilakam, is generally considered as the earliest dissertation referring to grammatical structures of Malayalam

In the earlies of 19th centuary Malayalam did not have a proper grammar. Malayalabhasaavyaakaranam published in 1851 by Hermann Gundert and the revised version published in 1868 was the first proper grammatical dissertation of Malayalam. Malayaalmayutevyaakaranam(1863) by Rev. George Mathen, Keeralabhaasaavyaakaranam by PachuMootthatu, Keeralapaaniniyam A.R RajaraajaVarma and Vyaakaranamitram(1904) by M. SeshagiriPrabhu followed. Grammatical literature from this point of time was fundamentally paying attention on Keeralapaaniniiyam, which came to have almost the status of an authorized grammar of Malayalam.

A regular grammatical custom illustration on a variety of grammars failed to develop and as a result the framework of Keeralapaaniniiyam continued as the solitary grammatical model in Malayalam. The grammars written in the post- Keeralapaaniniiyam period are basically descriptive treatises on Keeralapaaniniiyam. While a few grammarians have suggested other analyses in some areas, the grammars themselves truly follow the basic structure of RajarajaVarma. For an era of more than 80 years from Keeralapaaniniiyam, no grammarian attempted extend the Keeralapaaniniyam model to produce a more wide-ranging treatment of Malayalam or to evaluate the grammatical structure of Malayalam using alternative models of grammatical description. Keeralapaaniniyam and other traditional grammars have broadly covered the morphology of the language. However, there is little in them about syntax and semantics. To deal with the formation of a modern language like Malayalam using a controlled grammatical model has had severe repercussions in many fields.

Malayalam has 53 letters including 20 long and short vowels and others are termed as consonants. Malayalam sentences

A sentence is a cluster of words that makes absolute logic. There are four types of sentences available in Malayalam based on activities. Based on production they are of three types.

Sentence Classification Based on activities.

Malayalam sentences are basically of four types based on their activities.

• Assertive Sentence -This sentence makes a statement.
• Interrogative sentence -This ask a question.
• Imperative sentence -This show a command, appeal or a desire.
• Exclamatory Sentence -To express strong feeling, happiness, sorrow or wonder these kinds of sentences are used.

Sentence Classification based on production

Sentences in Malayalam can be of three types based on the production

a. Simple
b. complex
c. Compound

• Simple sentences- This type contains only one main clause.
• Complex sentences-These sentences contain one primary clause and any number of secondary clauses.
• Compound sentences -Compound sentences have any number of main clauses

7. TAGSET FOR MALAYALAM

This tag set has been developed based on the IIIT Hyderabad tag set for Indian languages.
It includes 28 tags.

Table 1.Tagset for Malayalam

Tag Name	**Type**	**Tag Name**	**Type**
NN	Noun	RDP	Reduplication
NST	Noun denoting spatial and temporal expressions.	CC	Conjuncts(coordinatingand subordinating)
NNP	Proper Nouns	UNK	Unknown words
NNS	Plural Nouns	PRO	Proverbs
PRP	Pronoun	IDM	Idioms
DEM	Demonstrative	NNPC	Compound Proper Nouns
VM	Verb Main	NLC	Noun Locative

VAUX	Auxiliary Verb	DOT	
JJ	Adjective	QF	Quantifiers
RB	Adverb	QC	Cardinal
PSP	Postposition	QO	Ordinal
ECH	Echo words	INTF	Intensifier
WQ	Question Words	NEG	Negation
SYM	Special Symbol	QM	Question Mark

Here I am drop the Tag NPC(Noun Compound) from IIITH Tag set and Add a TAG as NLC(Noun Locative). The tag NLC is necessary for most of the POS tagging Applications.
In my findings There is no need for tagging compound noun in Malayalam.

Consider the example:
തിമിരശസ്ത്രക്രിയ

In Malayalam this is a single word, there is no need to treat as compound noun.

Consider another example:
കാലിത്തൊഴുത്ത്

As the above example this is also a single word. Both these words can be dividing to two morphemes but in Malayalam Dictionary, these two words are exists. As my Tagging follows lexical dictionary, there is no need for NPC(Noun Compound) tag.

8. PROPOSED ALGORITHM FOR PARTS OF SPEECH TAGGING

Usually, for Parts of Speech Tagging, some statistical algorithms like Support Vector Machine, Hidden Markov Model, etc. are used. These algorithms are powerful and efficient but it doesn't build for Parts of Speech Tagging in Malayalam Specifically. So that it affects the accuracy of tagging process. In Dravidian languages, particularly for Malayalam language, inflected noun and verb forms are common. Nouns may have inflected with plural marker and case marker. Verbs might inflect with tense markers and are adjectivalized and adverbialized. So, many times we need to depend on syntactic function or context to decide whether the word is a noun or adjective or adverb or post position. This leads to the complexity in Malayalam POS tagging.

A noun may be categorized differently as common noun, proper noun or compound noun. Likewise, verb may be finite, infinite or gerund. Other parts of speech were also divided into their own subcategories.

For example, Malayalam word 'paadilla'(പാടില്ല) in the following sentences gives different parts of speech.

അവൻ പാടില്ല
Avanpaadilla
പുകവലി പാടില്ല
Pukavalipaadilla

In the first sentence, the word 'paadilla' is a verb whereas in the second sentence it is a noun. This is not rare in natural languages. A large amount of words are ambiguous. Also, the parts of speech are many more POS tags rather than noun, verb and post position etc.

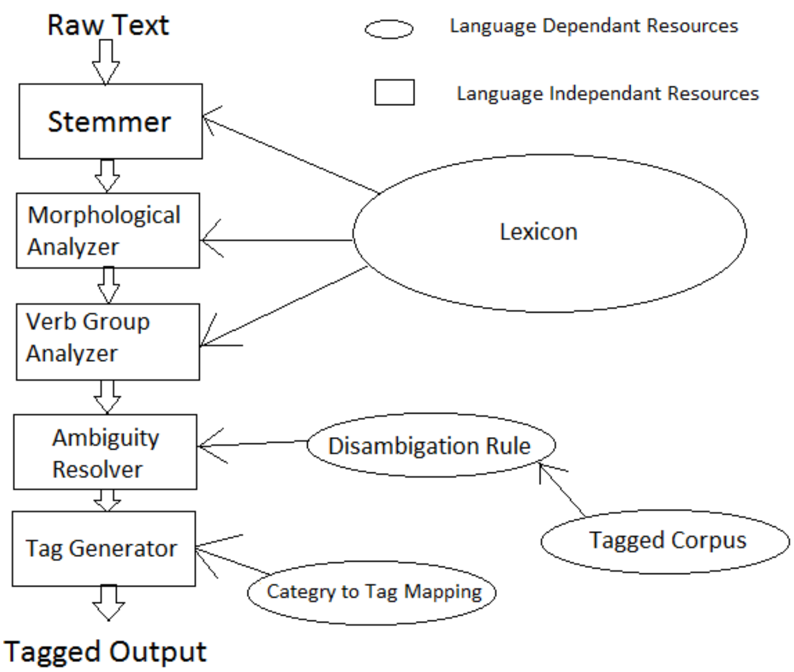
So, here we develop a simple but powerful algorithm for Parts of Speech Tagging for Malayalam. Here the Algorithm first performs tagging based on the lexical entries and then use statistical data if necessary, i.e..suppose a word sometimes act as noun and some other time verb.

Proposed Algorithm for POS tagging

```
1 TAKE INPUT TEXT.
2 TOKENIZE THE INPUT TEXT (PRE-EDITING).
3 MANUAL TAGGING.
4 (First thread.) IF WORD EXIST IN DICTIONARY THEN
         4.1 IF MULTIPLE TAG THEN
                 STORE IT IN A BUFFER
         4.2 ELSE
                 PUT THE APPROPRIATE TAG FROM DICTIONARY.
        REPEAT THESE STEPS UNTIL THE END.
5.  (Second Thread) ELSE
        1.1      DO STEMMING
               1.1.1     IF CASE MARKERS EXIST
               1.1.1.1  IF SUFFIX INCLUDES “KAL’ OR ‘MAAR’
                        TAG AS ‘NNS’(PLURAL NOUN)
               1.1.1.2  ELSE IF SUFFIX INCLUDES ‘IL’ OR ’ILE’
                        TAG AS NLC (NOUN LOCATIVE)
               5.1.1.3 ELSE
                        TAG AS ‘NN’ (NOUN)
               1.1.2     IF TENSE MARKERS EXIST
                        TAG AS ‘VM’ (VERB MAIN)
6.(Third Thread)  IF AN UNKNOWN WORD X
         6.1 IF AN UNKNOWN WORD X  IS PRECEDED BY A DETERMINER AND
        FOLLOWED BY A NOUN,
         TAG IT AS AN ADJECTIVE
        6.2. IF (Prev_word is NOUN or PRONOUN) THEN
        CURRENT WORD TAG AS ‘PSP’ (POSTPOSITION)
        6.3. ELSE IF THE CURRENT_WORD’s SUFFIX INCLUDES ‘YUM’ OR ‘OO’
        CURRENT_Word TAG AS ‘CC’ (CONJUNCTION)
        6.4. ELSE IF (Prev_word is VERB) AND (Current_Word is VERB)
        Current_Word TAG AS ‘VAUX (Auxiliary VERB)
7 IF THE NEXT WORD IS ADJECTIVE THEN
        TAG CURRENT_WORD AS INTENSIFIER
8. (Fourth Thread)READ FROM BUFFER
        IF    (THE TAGGING STRUCTURES IN THE DATABASE (CORPUS) EQUALS
CURRENT WORD’S SENTENCE STRUCTURE ) THEN
                PUT APPROPRIATE TAG
        ELSE
                ABANDON FOR MANUAL TAGGING
9.  GET THE TAGGED OUTPUT TEXT.
10. INSERT THOSE NEW WORDS IN LEXICON.
```

The POS tags are identified by doing a lexicon lookup of the root word. This is especially useful for morphologically rich Indian languages like Malayalam which inflect for gender, number, case etc. The figure 1 below shows the various components of the POS tagger. An “ambiguity resolver” is added to improve the accuracy of the POSTagger.

Figure 1.Pictorial representation of the system



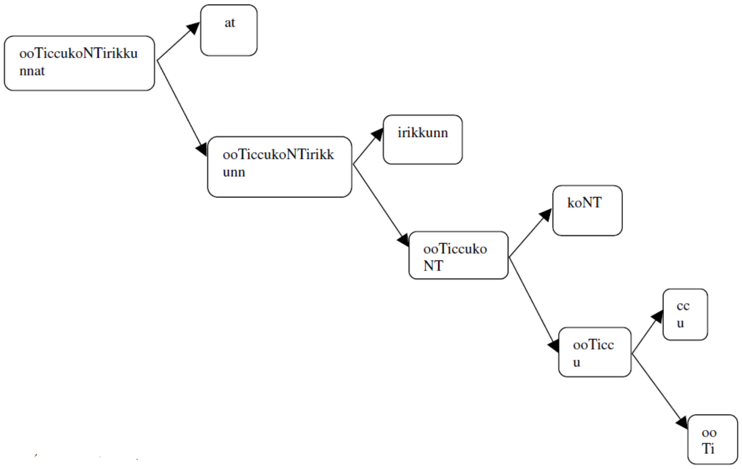
The architecture consist different modules based on their functionalities. The functionalities of each of this module are explained briefly as follow.

Tokenize: Untagged sentences are downloaded from Malayalam newspapers and commercial websites. We changed the input text into a column format suitable to the Algorithm. We used blank space as the column separator. The corpus data was tokenized as the input data to the algorithm must be in form of token.

Manual Tagging: The tokenizing module produces a corpus of untagged tokens. After which, the corpus is tagged manually using proposed IIITH tag set. Initially around 20,000 words are tagged manually.

SUFFIX SEPARATION RULES: The requirement of a pre-processing step in the training phase is exclusively attributable to the unusual characteristics of Malayalam language. The inflected word form in Malayalam can have multiple suffixes appended to its stem. This characteristic of Malayalam language reduces the possibility of a word in the corpus to be present in its stem. For example the word 'ഇന്ത്യ'' appears in the corpus in different forms, for example,ഇന്ത്യയുടെ, ഇന്ത്യക്ക് , ഇന്ത്യയോട് , etc..

Fig 2: Suffix separation Malayalam



Noun Analysis

The general format of input and output of the morphological analyzer of Malayalam is as follows
Word -> stem + suffixes

Linguistic categorization of Nouns, which takes cases markers as Person, Number and Gender information. The categorical information in noun is listed in Table 2 below.

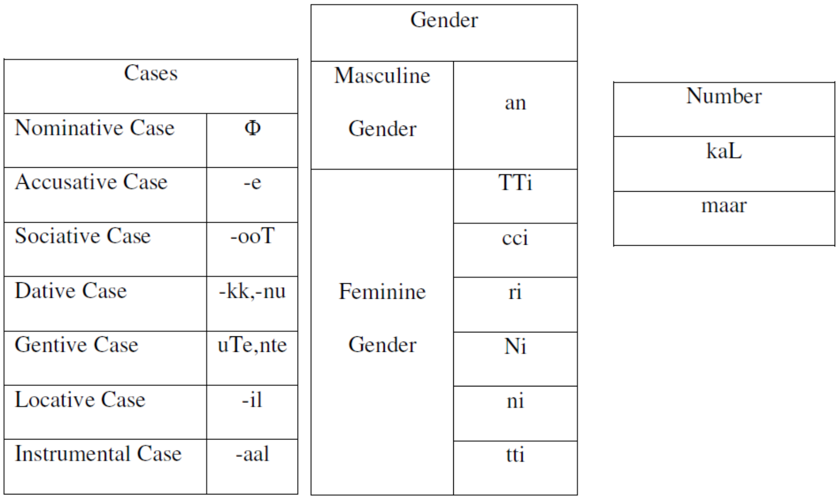
Verb Analysis

Verb in natural language is a grammatical category, which includes tense along with it. Also apprehensive information ie, tense, aspect and modularity (TAM) can be extracted from a verbal form. Many markers are there in the TAM and it is listed in Table 3.

Apart from this two other categories that play an important role in the morphological analysis are
a. Negations
b. Linkmorph.

Malayalam has the following negative markers-: ”aatt”,”aat”,”NTa”

Table 2: Noun Analysis

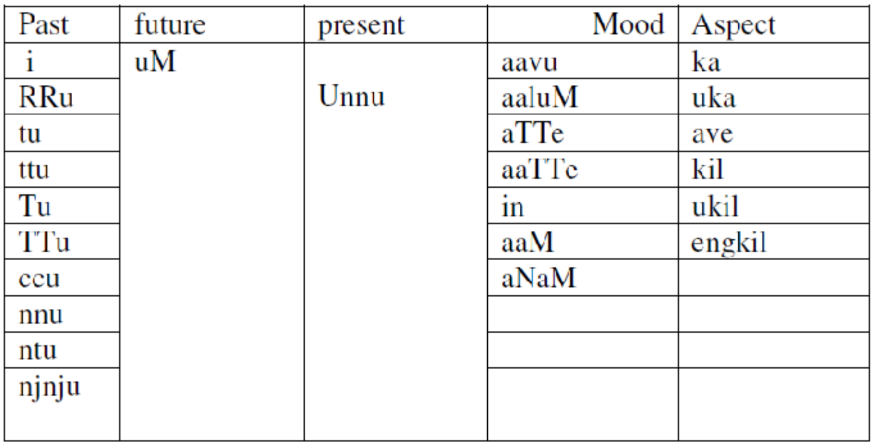


Cases	
Nominative Case	Φ
Accusative Case	-e
Sociative Case	-ooT
Dative Case	-kk,-nu
Gentive Case	uTe,nte
Locative Case	-il
Instrumental Case	-aal

Gender	
Masculine Gender	an
Feminine Gender	TTi
	cci
	ri
	Ni
	ni
	tti

Number
kaL
maar

Table 3: Verb Analysis



Past	future	present	Mood	Aspect
i	uM	Unnu	aavu	ka
RRu			aaluM	uka
tu			aTTe	ave
ttu			aaTTe	kil
Tu			in	ukil
TTu			aaM	engkil
ccu			aNaM	
nnu				
ntu				
njnju				

Ambiguity Resolver

A word in the lexicon contains tags and tag information of the adjacent words. If the word carrying single tag, then no complication, tag that word with the carrying tag. If the word carrying multiple tags, i.e. an ambiguous word, the system looks for the current word's adjacent tag information and those information cross matches with the already stored tag information and tag based on the tag structure. As the system follows multithreading technology, when it comes in the fourth thread adjacent words should tagged. Following figure 4 shows the same graphically.

Table 4: Ambiguity Resolver

Word	Tag	Tag structure

കാലി	NN(Noun)
	VM(Verb)
		NN NLC VM
		NN **VM**

9. CONCLUSION

Parts-of-Speech tagging, the assignment of Parts-of-Speech to the words in a given context of use, is a basic technique in many systems that handle natural languages. Tags play an important role in Natural language applications like text summarization, information retrieval and information extraction etc. In order to alleviate problems for Malayalam language, we proposed a new POS tagger approach. This paper describes a method for supervised training of a Parts-of-Speech tagger using newly developed algorithm for Malayalam. We identified the ambiguities in Malayalam lexical items, and developed a tag set appropriate for Malayalam. Finally, an efficient and accurate POS Tagger model for Malayalam language is built. We hope this will be very useful in natural language application like bilingual machine translation and in many areas.

REFERENCES

[1]. Dr. Ravi Sankar S Nair, A GRAMMAR OF MALAYALAM, Language in India (ISSN 1930-2940), NOV 2012.

[2]. S. DeRose. “Grammatical category disambiguation by statistical optimization”.*Computational Linguistics,* 1988, pp 14:31-39.

[3]. James Allen. Natural Language Understanding. Pearson Education, Second edition, 2002.

[4]. K. S. NarayanaPillai. AdhunikaMalayalaVyakaranam. The State Institute Of Languages,Kerala, Thiruvananthapuram-3, Second edition,2003.

[5]. Vinod P M, Jayan V and Bhadran V K, Implementation of Malayalam Morphological Analyzer Based on Hybrid Approach, in Proceedings of Twenty- Fourth Conference on Computational Linguistics and Speech Processing,2012.

[6]. A.R.Raja Raja Varma.”Kerala Paaniniiyam”.ISDL, 1999.

[7]. [Brill 1992] E. Brill, A simple-rule based part-of-speech tagger, In Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, 1992.