A Hybrid Parts Of Speech Tagger for Malayalam Language

Anisha Aziz T
Department of Computer Science and Engg.
VAST, Thalakkottukara
Thrissur,Kerala
anishaashijin05139185@gmail.com

Sunitha C
Associate Professor
Department of Computer Science and Engg.
VAST, Thalakkottukara
Thrissur, Kerala
sunitha@vidyaacademy.ac.in

Abstract— **Parts of speech tagging is an important research topic in Natural Language Processing research are. Since it is one among the first steps of any natural language processing (NLP) techniques such as machine translation, if any error happens for tagging the same will repeat in the whole NLP process. So far works had been done on POS tagging based on SVM, MBLP, HMM, Ngram. All of these methods were not fixing the problem of ambiguity. So for fixing ambiguity, we put forward a new Hybrid tagger for Malayalam. The combination of traditional rules and n-gram may produce better result compared to other methodologies. And also the ambiguity will be reduced by enriching the bigram dictionary. A bigram dictionary of co-occurring words are built with their tags. About 100000 more words are there in bigram dictionary. A corpus for Malayalam must be built which may be supposed to access by the model. It contains about 100000 words which are Malayalam words as well as the words originated from English. Since it's a hybrid tagger, we can take advantage of both traditional rules as well as bigrams. Also the heart of the research is the rule set, which contains 267 manually created rules. Rules can be applied with help of a morph analyzer. Rules are also used for tagging if bigram and corpus can't be referred for tagging. The proposed method when tested on 150 words, only 11 words were not identified, and obtained 90.5% accuracy. For the unidentified words, it can be caused by either the root word may not be in corpus or bigram, or the absence of rule. So adding the word, bigram or rule , we can improve the result and enhance the work. Addition is simple task. The size of bigram dictionary, corpus, and rule set and accuracy of morph analyzer influences the performance of the system.**

Keywords— N-gram, Rule based tagger, Natural Language Processing, Parts Of Speech tagging, Bigram dictionary, Corpus;

I. INTRODUCTION

Natural Language Processing is a research area related to linguistics, philosophy, artificial intelligence, and psychology. The need for this discipline is designing systems which is capable of understanding, processing and interpreting the computational mechanisms related to natural languages. Researches in natural language processing has been motivated by two main necessities which are getting computers for better understanding of the structure, semantics and functions of human language and to design natural language interfaces and thus to facilitate human-computer communication.

Classifying the words in a text into word classes is called tagging. The purpose of POS tagging is to assign part of speech tags to words which describe their syntactic category. A part-of-speech tagger is a software that uses various sources of knowledge like dictionary, corpus, frequencies, ngram to assign possibly unique POS tag to words. Automatic tagging is an important step in defining the linguistic structure of a large text corpus. It is an important component in high level analysis of text corpus. Its output can be used in many NLP applications, such as: speech recognition, spelling correction, speech synthesis, machine translation, query answering, , information extraction and searching large text databases.

Part-of-speech tagging , is the process of classifying the words in a text, based on both its semantics, as well as its context —i.e., by considering the contextual information as well such as relationship of the current word or token with adjacent and related words in a phrase, sentence. Normally we may think that , word categories are having tendencies toward semantic coherence (nouns describe names of people, places or things, and adjectives describe object properties), and semantic coherence is not at all used as a definition criterion for parts of speech, but this is not the case. Basically for the language English, parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and interjection, articles etc[2].

Example:

Word: Book	Tag: Noun
Word: Check	Tag: Verb

Some words can be associated with more than one tag. For example, run can be noun or verb which depends on the context. In a sentence like "I want to run out of here", the word or token "here is verb. While in sentence "Sachin has gained 100 runs", the token run is fallen into the category noun.

Ideally a typical tagger should have the following features, like simplicity, efficiency, accuracy, tunability and reusability

978-1-4799-8792-4/15/$31.00 ©2015 IEEE

and also unambiguity. In reality taggers either exactly identify the tag for the tokens of given text or make the possible best guess on the basis of available knowledge source like contextual information. As the natural language is complex and dynamically diverse it is sometimes very difficult for the POS taggers to make accurate decisions on tags. So errors that occur occasionally in tagging is not considered as a major roadblock to research.

When we think about NLP tasks with language Malayalam, we have to study the structure of Malayalam sentence and text. We need to know Malayalam in deep and profound. It's considered as language Malayalam was born from the language Tamil . However, Malayalam finds it's rich diversity of words and complex alphabets from the traditions of language Sanskrit, the Indo-Aryan language. This dynamic diversities, inflections and agglutinations has been achieved by no other language in India.

Due to this diverse characteristics and complexity the tagging for Malayalam text is considered to be very tedious and error prone. Here comes the greatest challenge for POS tagging in Malayalam. Ambiguity is hard to resolve in Malayalam. The richness and dynamic diversities make this language very severe to process the text. Hence it is sometimes difficult for the system to make accurate decisions when ambiguity encountered. So occasional errors are not considered as major roadblock to POS tagging.

II. Related Works

For Malayalam Language, POS tagger works have done based on MBLP,HMM[10],SVM, Ngram [9] etc. All of them are hard statistical methods. No works have done based on rules or transformations. And also they are supervised methods, ie they use some dictionary or corpus to work with.

A POS tagger, [3] proposes a Parts of Speech tagger for agglutinative language Malayalam which uses a stochastic approach .The corpus includes word frequencies as well as bigram statistics which are being used by this tagger for the process. Here the purpose of morphological analyzer is to generate a tagged corpus since an annotated corpus is unavailable in Malayalam. The process follows mainly three steps. If the training corpus is not available for the given input text , in first step, it uses the morphological analyzer to generate the tagged corpus which acts as the training corpus for tagging. In the second step, using the unigram and bigram probability the statistical analyzer module compiles the statistical data of the training corpus. Following this, the main component, the tagger module, decides the parts of speech of the tokens of the Test set. Even though the experiments had been performed on a very small sized corpus, the results shows that the statistical approach works fairly with a highly agglutinative language like Malayalam.

Another POS tagger is there which is based on Support Vector Machine method[4][8].In this supervised machine learning POS tagging approach, there is a requirement of a large amount of annotated training corpus for proper tagging. The main focus of this project was in the ambiguities in Malayalam lexical items and its resolution also develop an effective and accurate POS Tagger. They have proposed tag set for Malayalam which includes 29 tags where there are 1 tag for pronoun, 5 tags for nouns, two for number, 3 for punctuations, 7 tags for verbs, and 1 for each complimentizer, determiners, adjective, conjunction, adverb, emphasize, postposition and question word. Proposed algorithm for POS tagging is as follows: given the input text, it is first tokenized, manual Tagging and training of the corpus have to be performed, tagging using SVM, get the tagged output text. In this the tagging direction can be decided[8]. It may depend on our choice. The corpus they designed with 1, 80,000 tagged words. The authors claim that they have achieved performance of proposed system also increased to 94%.

A different approach to tag Malayalam sentences using MBLP approach has been described in [6]. The combination of two powerful and complex techniques: the efficient storage of solved examples, and similarity based reasoning on the basis of these stored examples to solve newly entering ones, give rise to a new idea called Memory based language processing (MBLP). This is implemented with the Tilberg Memory based Language (TiMBL) tagger tool and tested with existing SVM-tagger for Malayalam POS tagging. The system follows these steps for processing. The test set is first preprocessed, ie. the first each words are tokenized. After each token is separated, each token is first given a default tag of Noun, it is because for supervised training both training vector and training set must be in same format. After that test set is used by the TiMBL software. The TiMBL calculates the gain and entropy for each attributes and allocates the tag. The test was performed on a Linux machine with version Ubuntu Linux 12.04.Though this was well studied in Dutch language, in this paper it was extended to Malayalam language, and found to work perfectly well.

III. A Hybrid Parts Of Speech Tagger For Malayalam

Some words can have more than one tag associated with. For example, chair can be noun or verb depending on the context. In a sentence like "I want sit in this chair", the word or token here is verb. While in sentence "Sachin chairs in this conference", the token chair is fallen into the category noun.

When dealing with Malayalam language in NLP applications it's relevant to design a POS tagger which considers the properties of Malayalam language like diversity, inflection, agglutination and reduce the ambiguity. A hybrid tagging [5] methodology can be used with Malayalam POS tagging to reduce the ambiguity . The model will be supposed to use a tagged corpus of Malayalam. The hybrid tagger is a rule-based algorithm which also uses bigrams for automatic tagging of parts-of-speech to the given text. Also it uses a bigram dictionary if any ambiguous word encountered. The rules in the hybrid tagger for Malayalam is of the form

if Tagi +affix , then Tagj ;

Where Tagj is the final tag of the given word.

The procedure of the proposed hybrid tagger is as follows:-

1. The given text should be tokenized. The output of the tokenization step will be stored onto an array WORD.
2. For each token from WORD, if the token is present in the corpus with one , then output the tag.
3. If the stem word in the corpus itself has more than one tag, then search in the bigram dictionary for the token with the 2 adjacent tokens. Then output the tag for that particular bigram.
4. If the token is not present in the corpus then morphological analyzer is used for suffix stripping [7] from the token to get the stem word.
5. After that apply the rules and get the tag for that token.

Here bigram dictionary has been built with more than 10,0000 bigrams with the tags of each bigram. The bigrams are generated based on the articles of online Malayalam blog Jalakam. Also the Corpus is built from the online dictionary Olangal and the words that are originated from English also are included in the corpus. The corpus contains about 100,000 words. Because of the limitation of the size of corpus, the challenge is be to maximize the performance with the same corpus. But it is advantageous than statistical methods because it is free from the false prediction due to probability errors.

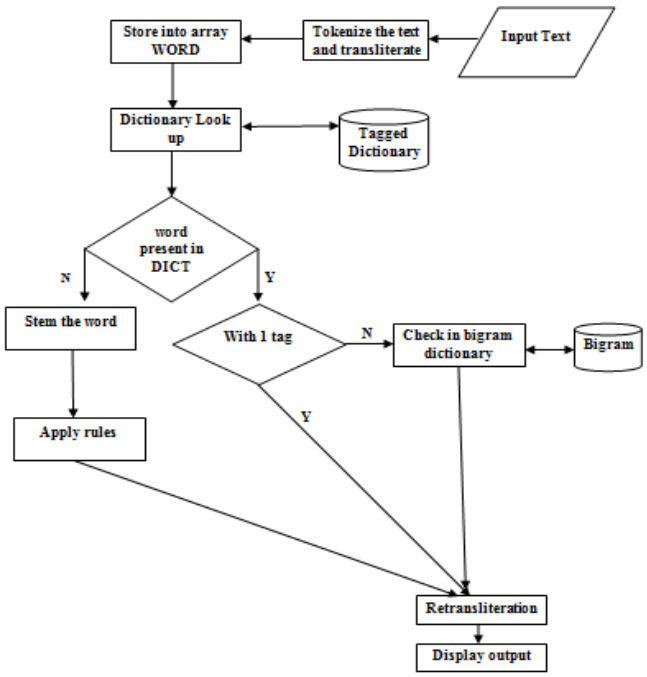

Fig 1: A Hybrid POS tagger For Malayalam

A. Traditional Rules used in the proposed vsystem

Since it's a hybrid tagger, it utilizes both the traditional rules as well as the bigrams. The rules in this system is of the form

Tagi +affix , then Tagj ;

The rules are used in the case of word not present in the corpus. The word not present in the corpus means that will be a complex word. So we need to use the morphological analyzer to strip the suffix from the word. So we get the output of morphological analyzer as

Stem(Tagi)+suffix

We then compare the tag and suffix part, that's

Tagi+suffix

with the traditional rule set. If any match occurs with the "Tagi+suffix" part with the left side of the rule, then output the final tag as the right side of the rule. ie the Tagj. There are 267 rules in the rule set which are created manually as a part of research, with the help of grammatical reference books in Malayalam. Now let's see an example of application of rules
eg: The word

ശക്തിയോടെ(ShakthiyOoTe)

Since it's not present in the corpus, we use the morphological analyzer and we get output as

ശക്തി(Shakthi-Noun)+ഓടെ(OoTe)

Then we check Noun+ഓടെ(OoTe) against the rule set. Since there is a rule

Noun+ഓടെ(OoTe)=Adverb

So we output the result as the tag for

ശക്തിയോടെ(ShakthiyOoTe) is Adverb

Like this we can see some more rules which are manually created as a part of our research. About 267 rules are manually created through our research. The rules are generated by profound study of many grammar reference books. More rules will be encountered while improving the accuracy. The following are some rules that are created on developing the tagger.
1. Noun+ ഏ(e)=Accusative Noun
2.Noun+ ക്(k)=Dative Noun
3. Noun+ ു(u)=Dative Noun
4.Noun+ ഓട്(Oot)=Sociative Noun
5.Verb+ ുണ്ട്(undu)=Verb-Progressive
6.Verb+ അല്ല/ ഇല്ല(alla/illa)=Verb-Negation
7.Verb+ ഏണ്ട(eenta)=Verb-Infinite
8. അല്ല +ഏ(ee)=Question Word
9. Noun+ ുടെ(uTe) =Adjective
10. Number+ ആം (aam) =Determiner
Here the second term after the addition sign is the suffix attached to the first word. So for identifying the suffixes attached, we may have to use morphological analyzer. And the

morphological analyzer used must be suffix stripping based one, so that suffixes can be simply identified and the result of morph analyzer is compared with the rule set. So the performance of the system depends on the accuracy of the morph analyzer also. The morph analyzer used for this system is [7], which is enhanced for verb and adjective for our better performance.

B. Proposed Tagset

The proposed tagset contains 39 tags of Malayalam which are as follows

Number	Tag	Description
1.	N	Noun
2.	ADJ	Adjective
3.	Q-ADJ	Qualitative Adjective
4.	ADV	Adverb
5.	Q-ADV	Qualitative Adverb
6.	PRN	Pronoun
7.	V	Verb
8.	V-AUX	Auxiliary Verb
9.	V-AUX-NEGATION	Auxiliary Verb Negation
10.	V-INF	Non finite Verb
11.	V-IMPER	Verb imperative
12.	V-INTERO	Verb Interrogative
13.	V-CONJ	Verb Conjunction
14.	V-CONJ-NEGATION-	Verb Conjunction negation
15.	COND-V-NEGATION	Verb Condition Negation
16.	V-NEGATION	Verb Negation
17.	ACCUSATIVE NOUN	
18.	LOCATIVE NOUN	
19.	DATIVE NOUN	
20.	CONJ	Conjunction
21.	ENUMERATIVE	
22.	SUBORDINATION	
23.	QUANTIFIER	
24.	REDUPLICATIVE ADV	Reduplicative Adverb
25.	REDUPLICATIVE ADJ	Reduplicative Adjective
26.	DEMONSTRATIVE	
27.	INTERJECTION	
28.	DERIVED NOUN	
29.	POSTPOSITION	
30.	NEGATION	
31.	SP	Special Utterances in Malayalam
32.	AVYAYAM	
33.	QW	Question Word
34.	DETERMINER	
35.	INTENSIFIER	
36.	DISTRIBUTIVE	
37.	NUMBER	
38.	V-PROGRESSIVE	Progressive Verb
39.	V-ITERATIVE	Iterative Verb

TABLE I :Proposed Tag Set

C. Solved example of proposed Hybrid tagger

Now let's see how an input text is being processed and obtain the tags

Input text: അവൻ രാവണനെ വധിച്ചു(Avana Ravanane vadhichchu).

Here the input text is the sentence given above. So in the first step, we may tokenize the sentence. The tokens

'അവൻ(avan)', ' രാവണനെ(Ravanane) ', ' വധിച്ചു(vadhichchu) '

are saved on to an array WORD. For each token from WORD, if it's present in corpus then output the tag. So the first token is

അവൻ(Avan)

It's already present in the corpus, with tag pronoun. So output the result. Go for the next token,

രാവണനെ(Ravanane)

which is not present in the corpus. So we may use the morphological analyzer. The output of morphological analyzer will be as follows.

രാവണനെ (Ravanane-Noun)+ എ(e)

Here the token

രാവണൻ(Ravanan)

which is present in corpus, and the result of morph analyzer is compared with the rule set, and get the tag as ACCUSATIVE NOUN, because there is a rule " N+ e =ACCUSATIVE NOUN"

Get the next token which is

വധിച്ചു(Vadhichchu)

which is not present in corpus, so we use morph analyzer and get the result as

വധിക്കുക(Vadhikkuka-V)+cch+cchu

Now we compare the result of morph analyzer with the rule set. And we get the tag V since there is a rule like " V+cch+cchu=V"

The bigram are used when the token is the stem word itself with more than one tag.

For eg: The word

എന്നാല് (ennal)

Here

എന്നാല്(ennal)

can have meaning "myself" with PRN and "but" with tag CONJ. So the case of ambiguity for the stem word happens. So Since the stem word itself is ambiguous we can't use the morphological analyzer and rules. So we go for bigrams

,

if the token

എന്നാല്(ennal) has ആവുന്ന(aavunna) or
സാധിക്കുന്ന(saadhikunna) or കഴിയുന്ന(kazhiyunna)

as adjacent to it, then we can output a

എന്നാല്(ennal) is PRN

Otherwise it is CONJ. Like this the co occurring bigrams are stored in a bigram dictionary to resolve the ambiguity of the stem word itself. The bigrams are created using the Malayalam blog "JALAKAM" Also some bigrams that introduce ambiguity also added into the dictionary.

IV. RESULT ANALYSYS AND PERFORMANCE EVALUATION

If we have given the input -രാമന് രാവണനെ വധിച്ചു (Raman Ravanane Vadhichu) we get output as
രാമന് -N , രാവണനെ-Accusative noun, വധിച്ചു-V
Another example is -എ‌ിറെ വരണം ആണ്. നീ വരണം.(Ente varanam aanu. nee varanam) we get output as എ‌ിറെ-PRN, വരണം-N, ആണ്-V-AUX, നീ-PRN, വരണം-V
Here are the analysis of existing systems and their advantages as well as disadvantages are listed along with their accuracy. Here the best tagger is HMM based tagger of IIITM-K which is available online. The same set of input has been given to our tagger and HMM tagger. The best result is our tagger's result. Because it identifies 39 word classes , noun suffixes are identifies and the corresponding tag is obtained, also pronouns are correctly classified. Also the ambiguity is resolved in our tagger, which is not in any tagger in Malayalam. So our tagger is efficient than existing taggers. And the performance is obtained to be 90.5%

Method used	**Advantages**	**Disadvantages**	**Accuracy**
SVM	Good performance	-Large computational cost - Large testing time	87%
TnT	Fast training and tagging	-Large amount of ambiguities - Insufficient amount of Data	75%
HMM	Ever best performance in Malayalam	-Pronouns are not correctly identified -doesn't identify the suffixes attached to noun and it's tag -Only 11 major tags are there	90.5%

TABLE II-Comparison of existing POS taggers

V. APPLICATION

POS tagger finds wide application in most of the NLP tasks. The POS tagger can be widely used as a preprocessor for any NLP task like Machine Translation. Text retrieval and indexing uses Parts Of Speech information. Speech synthesis uses Parts Of Speech tags to decide the pronunciation and resolve the sound ambiguity. POS tagger is also used for making tagged corpus. Also POS tagging is first important step in Machine translation. So if the first step is erroneous then the translation may also give false solution. So it's very necessary for POS tagging to be accurate for the machine translation to be accurate in Natural Language applications.

VI. CONCLUSION

Words are classified into different classes called Parts of Speech Tagging, morphological categories, word classes, or lexical tags. Parts Of Speech Tags play an important role in Natural language applications like machine translation, natural language parsing, speech recognition, information extraction and information retrieval. We have tried to resolve the ambiguities in Malayalam lexemes, and developed an appropriate a tag set appropriate for Malayalam. Finally, an effective and accurate Hybrid POS Tagger for Malayalam language is built. A Hybrid Parts Of Speech tagger has been built in order to reduce the grade of ambiguity. A Hybrid of traditional rules and bigrams are used in the proposed approach.

REFFERENCE

[1] Beáta Megyesi, "Brill's Rule-Based Part of Speech Tagger for Hungarian", Computational, Linguistics Spring 1998 ,Department of Linguistics ,Stockholm University,1992

[2] Itrc.iiit.ac.in/nlptools2010/file/documents/POS-tag-list.pdf "IIIT Hyderabad POS tag set", 2006

[3] Manju k, soumya s, sumam mary idicula "Development Of A Pos Tagger For Malayalam-An Experience", ,Department of computer science, Cochin university of science and technology,2009

[4] Antony P.J, Santhanu P Mohan, Soman K.P "SVM based Parts of Speech Tagger for Malayalam", CEN, Amrita University, Coimbatore, India antonypjohn@gmail.com,2010

[5] Cynthia Myint, "A hybrid approach for part-of-speech tagging of Burmese texts", University of Computer Studies, Mandalay, Myanmar,2011

[6] Robert Jesuraj K ,"MBLP approach applied to POS tagging in Malayalam Language" , 3rd Sem, M.Tech in Computational Linguistics, Dept. of Computer Science and Engineering , Government Engineering College, Sreekrishnapuram, Palakkad, India,2013

[7] Nimal J Valath, Narsheedha Beegum , "Malayalam Morphological Analyzer: A Simple Approach" , M.Tech Student, Department of Computer Science, Vidya Academy of Science and Technology, Kerala, India,2013

[8] Jes´us Gim´enez and Llu´ıs M`arquez "SVMTool: A general POS tagger generator based on Support Vector Machines", TALP Research Center, LSI Department, Universitat Polit`ecnica de Catalunya , Jordi Girona Salgado 1–3, E-08034, Barcelona , jgimenez,lluism_@talp.upc.es,2004

[9] Fahim Muhammad Hasan, Naushad UzZaman and Mumit Khan ,"Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla" ,Center for Research on Bangla Language Processing, BRAC University, Bangladesh

[10] Dinesh Kumar, Gurpreet Singh Josan "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010