FORMAL VERB ANALYSIS IN MALAYALAM LANGUAGE

Sumithra M D, Divya S Nair

Abstract—**Natural Language Processing (NLP) is a subfield of Artificial Intelligence and Linguistics. In NLP, the formal analyzer plays an important role in processing different forms of human language for formal writing and reading. A formal analyzer supplies information concerning morphosyntactic properties of the words it analyses or constructs. Formal Analysis is an important component for building computational grammars as well as Machine Translation. The formal analyzer mainly deals with the study of the internal structure of the words based on its grammatical features of any language. Malayalam is morphologically rich and agglutinative language. The proposed system aims to return 'root/stem' of a word along with its grammatical information depending upon its word category using string matching algorithm.**

Keywords- Natural Language Processing, Artificial Intelligence, Morphological analyzer.

I. INTRODUCTION

Natural Language Processing is a subfield of Artificial Intelligence and Linguistics. In NLP, the formal language analyzer plays an important role in processing different forms of human language for formal writing and reading. The formal language analyzer also called morphological analyzermainly deals with the study of the internal structure of the words based on its grammatical features of any language. It will return its root/stem word along with its grammatical information depending upon its word category. Formal language analysis is the process of splitting the surface form into its lemma and grammatical information. For example, the word 'Cats' splits into its lemma 'cat' and grammatical information <noun><plural>.

Malayalam is a morphologically rich and agglutinative language. The design and implementation of the formal language analyzer for Malayalam is a promising research for various applications in NLP. The previous works carried out in the field of Malayalam Morphological Analyzer, is not complete. Eg : Brute Force Method [1]. As Malayalam requires many morphophonemic changes in the word format ion, the above mentioned method is not sufficient. The proposed work use the method called Suffix Stripping which uses String matching algorithm and Sandhi rules to find the root/stem form.

This Malayalam formal language Analyzer would help in automatic spelling and grammar checking, natural language understanding, web searching, machine translation, speech recognition, speech synthesis, part of speech tagging, and parsing applications.

II. OBJECTIVE

Theproposed system,**"Formal Language AnalyzerForMalayalam Language",** specifically aims to return 'root/stem' of aword along with its grammatical information depending upon its word category.

For nouns, it will provide gender, number, caseinformation.

For verbs, it will provide tense, modularity.

Sample Input

അവൾ കാണുകയായിരുന്നു

Obtained Output

Noun : അവൾ
(വിഭക്തി വിഭാഗം : നിർദ്ദേശിക)
Verb : കാണുകയായിരുന്നു
(നിശ്ചയഭൂതകാലം)

III. LITERATURE SURVEY

A. Brute Force Method:

The term brute force [1] is the concept of artificial intelligence research and problem solving concept in mathematics denoted as brute force search. Brute force stemmers employ a lookup table which contains relations between root forms and inflected forms. To stem a word, the table is queried to find a matching inflection. If a matching inflection is found, the associated root form is returned. Brute force approaches are criticized for their general lack of elegance in that no algorithm is applied that would more quickly converge on a solution. In other words, there are more operations performed during the search than should be necessary. The algorithm is only accurate to the extent that the inflected form already exists in the database. Given the number of words in a given language, like English, it is unrealistic to expect that all word forms can be captured and manually recorded by human action alone. Manual training of the algorithm is overly time-intensive and the ratio between the effort and the increase in accuracy is marginal at best. Brute force algorithms are initially very difficult to design given the immense amount of relations that must be initially stored to produce an acceptable level of accuracy.The method is only accurate to the extent that the inflected form already exists in the database.

B. Root Driven Approach

In the root driven approach, the stem of the word should be firstly found in a lexicon before starting the morphological

analysis. The major drawback of this approach is the cost of the searching process required to find the stem. In the work done by Solak and Oflazer [2],the word itself and its subparts, which have been obtained by removing the letters one by one from the end of this word, are looked up in a lexicon to find all the possible stems. The real stem is discovered after the morphological analysis made by using these possible stems. Even though there are different search methods improving the performance, the examining of each subpart is obviously a very time consuming process especially for the languages where the words can appear in very long forms.

C. *Stemming Algorithm*

A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form. Affix removal conflation techniques are referred to as stemming algorithms and can be implemented in a variety of different methods. All remove suffixes and/or prefixes in an attempt to reduce a word to its stem. Suffixes that are concatenated to words are often done so in a certain order, such that a set of order-classes will exist among suffices. An iterative stemming algorithm will remove suffixes one at a time, starting at the end of the word and working towards the beginning. A stemming algorithm, a procedure to reduce all words with the same stem to a common form, is useful in many areas of computational linguistics and information-retrieval work.

IV. PROPOSED SYSTEM

The objective of the proposed system is to identify the root/stem word along with its grammatical information and more importance is given to verb analysis. The current morphological analysis deals only the word is noun or a verb. In our method we identify the verb with its tense and its categories (ie., For Present tense – Simple Present, Present Continuous etc.).

Suffix Stripping [4] which is a method used for analysis makes use of a stem dictionary (for identifying a valid stem), a suffix dictionary containing all possible suffixes that nouns/verbs in the language can have (to identify a valid suffix), morphotactic rules(case, gender, number) and morphophonemic rules or sandhi rules. This method is economical. Even if the item does not exist in the dictionary, the analyzer can identify the suffixes and stem. Once the suffixes are identified, removing the suffixes and applying proper sandhi rules can obtain the stem. Suffix stripping algorithms do not rely on a lookup tables; instead, rules are stored to find the root/stem.

Root	Inflected Forms
ഓട്ടം	ഓടി
	ഓടുന്നു
	ഓടും
കുട്ടി	കുട്ടികൾ
	കുട്ടിക്ക്
	കുട്ടിയെ
	കുട്ടിയോട്
	കുട്ടിയിൽ

ഉപവിഭാഗങ്ങൾ	പ്രത്യയം	ഉദാഹരണം
സാമാന്യഭാവികാലം	ഉം	കാണും, പറയും
നിശ്ചയഭാവികാലം	ഇരിക്കും	കണ്ടിരിക്കും, പറഞ്ഞിരിക്കും
തുടർഭാവികാലം	കൊണ്ടിരിക്കും	കണ്ടുകൊണ്ടിരിക്കും, പറഞ്ഞുകൊണ്ടിരിക്കും
ആസന്നഭാവികാലം	ആൻ+പോവുക യാണ്	കാണാൻപോവുകയാണ്, പറയാൻപോവുകയാണ്
സന്നിഗ്ധഭാവികാലം	ആയിരിക്കും	കാണുമായിരിക്കും, പറയുമായിരിക്കും
വികല്പഭാവികാലം	എക്കും, എക്കാം	കണ്ടേക്കും, പറഞ്ഞേക്കാം

Table 1: Sample Lookup T able

A. *Noun Analysis*

The general format of input and output of the morphological analyzer of Malayalam is as follows:

Word -> stem + suffixes

Nouns are linguistic categories, which takes cases and PNG (Person, Number and Gender) information. The categorical information in noun is listed in Table2 below.

B. *Verb Analysis*

Verb is a grammatical category, which takes tense with it. Also modular information can be extracted from a verbal form. Many markers are present in the Tense and Mood is listed in the suffix dictionary.

വിഭക്തികൾ	
നിർദ്ദേശിക	ϕ
പ്രതിഗ്രാഹിക	എ
സംയോജിക	ഓട്
ഉദ്ദേശിക	ന്,ഉ്
പ്രയോജിക	ആൽ
സംബന്ധിക	റ്റെ,ഉടെ
ആധാരിക	ഇൽ, കൽ

ബഹുവചനം
കൾ
മാർ
അൻ

ലിംഗഭേദം	
പുല്ലിംഗം	അൻ
സ്ത്രീലിംഗം	അൾ, ഇ
നപുംസകലിംഗം	ഉ, തു

Table 2: Categorical Information In Noun

കാലം	
ഭൂത കാലം	തു, ഇ
ഭാവികാലം	ഉം
വർത്തമാന കാലം	ഉന്നു

പ്രകാരം	
നിർദ്ദേശിക	ф
നിയോജകം	അട്ടെ
വിധായകം	അണം
അനുജായകം	ആം

Table 3: Categorical Information In Verb

C. Verb Analysis

1. ഭൂതകാലം (Past Tense)

- സാമാന്യഭൂതകാലം(Simple Past Tense) : The simple past is used to express the idea that an action started and finished at a specific time in the past.
  Eg: I saw a movie yesterday
- നിശ്ചയഭൂതകാലം (Past Definite Tense) : The past definite is used when we have personal knowledge and witness of the action.
  Eg: I have made it.
- തുടർഭൂതകാലം (Past Continuous Tense) : The past continuous is used to indicate that a longer action in the past was interrupted.
  Eg: I was watching TV when she called.
- ആസന്നഭൂതകാലം (Past Immediate Tense) : Immediate past tense refers to a time considered very recent in relation to the moment of utterance.
  Eg: I have seen it.
- പൂർണ്ണഭൂതകാലം (Past Perfect Tense) : It is used to express an action which has occurred in past and action which has occurred in past before another action in past.
  Eg: The student has gone before the teacher came.
- സന്നിഗ്ധഭൂതകാലം (Past Indefinite Tense) : It is used to describe actions but do not state whether the action is complete or on-going.
  Eg: I think that I went out.
- ഹേതുഹേതുമത് ഭൂതകാലം (Past Conditional) :

Eg: When I had a duty off from work, I often went to the beach.

ഉപവിഭാഗങ്ങൾ	പ്രത്യയം	ഉദാഹരണം
സാമാന്യഭൂതകാലം	ഇ, തു	പോയി, കണ്ടു
നിശ്ചയഭൂതകാലം	ആയിരുന്നു	കാണുകയായിരുന്നു, പറയുകയായിരുന്നു
തുടർഭൂതകാലം	കൊണ്ടിരുന്നു	കണ്ടുകൊണ്ടിരുന്നു പറഞ്ഞുകൊണ്ടിരുന്നു
ആസന്നഭൂതകാലം	ഇട്ടുണ്ട്	കണ്ടിട്ടുണ്ട്
പൂർണ്ണഭൂതകാലം	ഇരുന്നു, ഇട്ടുണ്ടായിരുന്നു	വന്നിരുന്നു, പോയിരുന്നു, കണ്ടിട്ടുണ്ടായിരുന്നു
സന്നിഗ്ധഭൂതകാലം	ഇട്ടുണ്ടാണ്ടിരിക്കും	കണ്ടിട്ടുണ്ടായിരിക്കും പറഞ്ഞിട്ടുണ്ടായിരിക്കും
ഹേതുഹേതുമത് ഭൂതകാലം	ഉം+ആയിരുന്നു	കാണുമായിരുന്നു വരുമായിരുന്നു

Table 4: Categorical Information In Past Tense

2. വർത്തമാനകാലം (Present Tense)

- സാമാന്യവർത്തമാനകാലം (Simple Present Tense):
  The simple present tense in English is used to describe an action that is regular, true or normal.
  Eg: The President of the USA lives in the white house.
- നിശ്ചയവർത്തമാനകാലം(PresentDefinite Tense):
  This is formed with a present tense form of "to have" plus the past participle of the verb. This tense indicates either that an action was completed at some point in the past or that the action extends to the present.
  Eg: I have walked two miles already.
- തുടർവർത്തമാനകാലം (Present Continuous Tense):
  It is used to express a continued or ongoing action at present time. It expresses an action which is in progress at the time of speaking.
  Eg: I am playing cricket.
- ആസന്നവർത്തമാനകാലം (Present Immediate Tense):
  The immediate present to indicate action that is actually happening right now
  Eg: I'm on my way to the store

ഉപവിഭാഗങ്ങൾ	പ്രത്യയം	ഉദാഹരണം
സാമാന്യവർത്തമാനകാലം	ഉന്നു	കാണുന്നു, പറയുന്നു
നിശ്ചയവർത്തമാനകാലം	ആകുന്നു,ആണ്	കാണുകയാണ്, പറയുകയാകുന്നു
തുടർവർത്തമാനകാലം	കൊണ്ടിരിക്കുന്നു	കണ്ടുകൊണ്ടിരിക്കുന്നു പറഞ്ഞുകൊണ്ടിരിക്കുന്നു
ആസന്നവർത്തമാനകാലം	ഉന്നുണ്ട്	കാണുന്നുണ്ട്, പറയുന്നുണ്ട്

Table 5: Categorical Information In Present Tense

3. ഭാവികാലം (Future Tense)

- സാമാന്യഭാവികാലം (Simple Future Tense):
  The simple future tense is used for an action that will occur in the future.
  Eg: I will buy a computer tomorrow.
- നിശ്ചയഭാവികാലം (Future Definite Tense):
  Future Definite tense is used to expresssituations that will last for a specified period of time at a definite moment in the future.
  Eg: Before they come, we will have been cleaning the house for 5 hours.
- തുടർഭാവികാലം (Future Continuous Tense):
  The Future Continuous to indicate that a longer action in the future will be interrupted by a shorter action in the future.
  Eg: I will be watching TV when she *arrives* tonight.
- ആസന്നഭാവികാലം (Future Immediate Tense):
  The immediate future tense is used to talk about what is going to happen in the future.
- സന്നിഗ്ധഭാവികാലം (Future Indefinite Tense):
  It is used to express an action which has not occurred/happened yet and will occur/happen after some time in future.
  e.g. Anu will take bath after half an hour.
- വികല്പഭാവികാലം(Future Conditional Tense):
  The Future Conditional describes what you think you will do in a specific situation in the future and is used to talk about imaginary situations in the future
  Eg:If Ihad a day off from work next week, I would go to the beach.

Table 6: Categorical Information In Future Tense

D. Sandhi Rules

Sandhiis classified as Elision, Augmentation, Substitution and Reduplication.

Elision: When two sounds join together one sound willbe lost is called Elision.

കണ്ടു+ഇല്ല=കണ്ടില്ല

Aug mentation: When two words join together onesound comes between these twowords.

താളി+ഓല=താളിയോല

Substitution: When two words join together, one phoneme is substituted by another

ഹിമ+അദ്രി=ഹിമാദ്രി

Reduplication: When two words combined together, the letter gets duplicated. This process iscalled reduplication.

നീല+താമര=നീലത്താമര

E. Algorithm

1. Accept Input.
2. Scan the input from right to left character by character
3. Check the Unicode Decimal value of each scanned character to form a Suffix
4. Check the validity of the formed suffix
   If suffix not valid, scan the next character,
   GOTO step 3
   Else:
   (a) Find the hash value of Suffix.
   (b) Separate Suffix from remaining portion.
5. The remaining input characters are considered as the root word by applying Sandhi rules
6. Display the grammar.
7. End

V. SYSTEM DESIGN

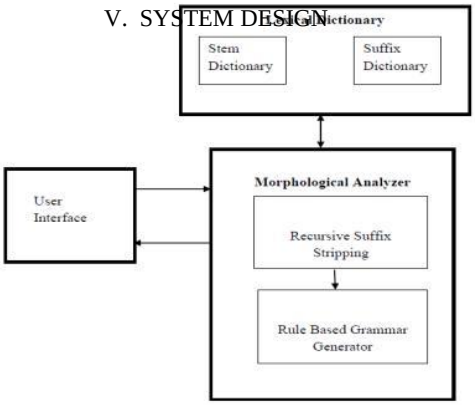

Figure 1: System Architecture

A. Architecture

User Interface:

This module helps the user to interact with the system. User can enter the word and gets the result through this module.

Lexical Dictionary:

This contains two units Stem Dictionary & Suffix Dictionary. Stem dictionary contains the root words and Suffix dictionary contains the suffixes.

Morphological Analyzer:

It is the processing component which identifies the root word from the user input.The stem dictionary is searched to verify whether the input is already a root word.If input is not a root word, further processing is continued under the assumption that user input is a valid Malayalam word.Suffix of input is extracted using Malayalam grammatical rules.The suffix dictionary is searched for the identified suffix, based

on hash values.Finally validity of user input may be verified using stem dictionary.

VI. CONCLUSION

The suffix stripping method proposed here for identifying the grammar from any of the agglutinative word from Malayalam vocabulary gives a fine tuned result .The method also support fast convergence compared with the existing brute force method and root driven method.

ACKNOWLEDGMENT

I am greatly indebted to Dr.K.C Raveendranathan, Principal, LBS Institute Of Technology For Women and Dr Shreelekshmi R, Head of Department, Dept. of Computer Science & Engineering, for providing all the required resources for my thesis work. I would like to sincerely thank my project guide, Mrs Sumithra M D Dept of Computer Science &Engineering for her valuable suggestions and guidance. I would like to express my sincere gratitude to all teachers of teachers of computer science department for their moral and technical support throughout the course of this thesis work

REFERENCES

[1] Dinesh Kumar, Prince Rana, "Stemming of Punjabi Words By Using Brute Force Technique", *International Journal of EngineeringScience and Technology*, Vol. 3 No. 2 Feb 2011

[2] A. Solak& K. Oflazer, Design and Implementationof a Spelling Checker for Turkish, *Literary and LinguisticComputing*, *8*(3), 1993, 113-130.

[3] Vinod P M, Jayan V, Bhadran V K, CDAC Thiruvananathapuram, "Implementation of Malayalam Morphological Analyzer Based on Hybrid Approach", *Proceedings of the 24thConference onComputational Linguistics & Speech Processing* , 2012.

[4] Rajeev R R, Dr Elizabeth Sherly, "Morphological Analyzer for Malayalam Language: A Suffix Stripping approach", *Proceedings ofthe 20th Kerala Science Congress*, Thiruvananthapuram, 2008.

[5] Jisha P Jayan, Rajeev R R, Dr Elizabeth Sherly, "Parts of Speech Tagger for Malayalam", *International Journal of Computer Science& Information Technology*, Volume 2, No.2, December 2009.pp.209-213

[6] Ravi Sankar S Nair, Ph.D, "A Grammar of Malayalam",Language in India, *www.languageinindia.com*, ISSN 1930-2940, 12 : 11 November 2012