ISSN: 2278-0181
RACMS-2014 Conference Proceedings

Modern Tamil Word Formation Rules in NLP

Dr. K. Nirmala.
MCA.,Ph.D
Research Supervisor and Associate Professor,
Department of Computer Science,
Quaid-E-Millath Government College for Women (Autonomous)
Chennai, India

M. K. Kalpana
Department of Computer Science,
Quaid-E-Millath Government College for Women (Autonomous)
Chennai, India

***Abstract*—In Modern Tamil NLP(Natural Language Processing) new words created by Word Formation Rules(WFR) under eight categories, There are, Noun to Noun, Verb to Noun, Adjective to Noun, Noun to Verb, Adjective to verb, Verb to verb, Noun to Adjective, Verb to adverb. This WFR is the enhanced feature compare than sandhi rules/word joining and splitting. Word Formation techniques is a pre-processing stage of Tamil Noun and Verb Morphological operations in Tamil grammar and linguistics.**

Keywords*— Tamil Unicode, Rule based Machine Translation, Sandhi Generator, Word Formation Rules and Tamil Morphology.

I. INTRODUCTION

The Word Formation Rules technology is applicable in language technology including, Text processing, Speech Processing, Machine learning and understanding techniques. This is process is having possibilities in all Natural Languages communications under the human being. Using set of datasets and rules this process is generating in a system. Computational technicians and researchers handling these methods in machine level and programming level techniques.

II. TAMIL UNICODE

Tamil is a Unicode block containing characters for the Tamil. Tamil is having 247 letters for a modern Tamil communication. Modern Tamil is a colloquial spoken language in TamilNadu in India. It is generates with other 9 North Indian letters. Mostly all Indian languages functioning with independent vowels and dependent consonants. Dependent consonants not processing individually in language technology. It is used to generate different kinds of tamil letters in computational method like a manual way. Tamil Unicode/letters also functioning as an other Indian Unicode characters like other Indic script. Tamil having more letters compare than other Indian languages. So, positions of the character are having similar differences compare than other Indic scripts. In below, Tamil Unicode consortium code chart given for an example from Wikipedia.

Tamil[1][2]
Official Unicode Consortium code chart (PDF)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+0B8x			ஂ	ஃ		அ	ஆ	இ	ஈ	உ	ஊ				எ	ஏ
U+0B9x	ஐ		ஒ	ஓ	ஔ	க				ங	ச		ஜ		ஞ	ட
U+0BAx				ண	த				ந	ன	ப				ம	ய
U+0BBx	ர	ற	ல	ள	ழ	வ	ஶ	ஷ	ஸ	ஹ					ா	ி
U+0BCx	ீ	ு	ூ				ெ	ே	ை		ொ	ோ	ௌ	்		
U+0BDx	ௐ							ௗ								
U+0BEx							௦	௧	௨	௩	௪	௫	௬	௭	௮	௯
U+0BFx	௰	௱	௲	௳	௴	௵	௶	௷	௸	௹	௺					

Notes
1.^ As of Unicode version 7.0
2.^ Grey areas indicate non-assigned code points

Fig.1.Tamil Unicode Consortium Chart from Wikipedia

Tamil letters/Unicodes in computational method.

ன + ி ≠ ன + ் + இ

Tamil Grammar rules manually,

ன்+இ=னி

A. *Rule based machine translation*

All Tanveer Siddiqui, U. S. Tiwary says, "Rule-based Machine Translation systems parse the source text and produce an intermediate representation, which may be a parse tree or some abstract representation. The target language text is generated from the intermediate representation. These systems rely on specification of rules of morphology, syntax, lexical selection and transfer, semantic analysis and generation, and are hence called rule-based systems. It is classified into two:

i. Transfer-based machine translation

This transformation requires an understanding of the differences between the source and target language. In order to get the structure of the input, some form of parse is needed. Thus, a transfer-based machine translation system has the following three components:

1. Analysis-To produce source language structure
2. Transfer-To transfer the source language representation
3. Generation-To generate target language text using target level structure.

ii. Interlingua machine translation

In interlingua-based machine translation approach, the source language text is converted into a language independent meaning representation called “interlingua”.

III. SANDHI GENERATOR

[10]Sandhi deals with all kinds of changes like insertion, cancellation, replacement etc., when two or more morphemes or words occur together is called „Sandhi‟. In general the end of the first morpheme or word and the beginning of the following one are taken into account in Sandhi. The end and the beginning of morphemes that are added can be either a vowel or a consonant or consonant cluster. There are many combinations of such characters. These Sandhi rules should be explicitly specified for morphological analysis in a rule based system. These rules can be learned automatically by the system from the training samples and subsequently be applied for new inputs.

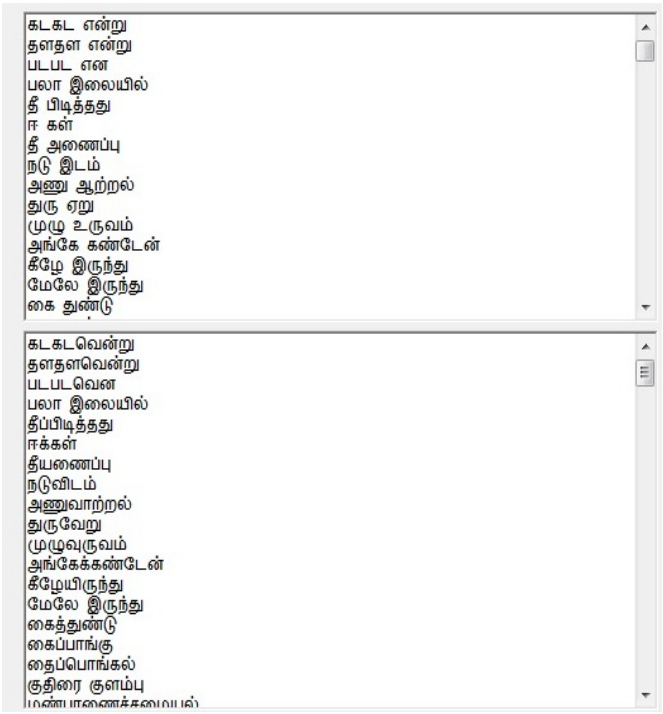

Fig.1. Sandhi Rules Generator

A. *Sandhi Rules using Diacritical markings*

Diacritics is a sign, such as an accent or cedilla, which when written above or below a letter indicates a difference in pronunciation from the same letter when unmarked or differently marked.

Example:
“Ganesh kadaikku cenran”.
Here, in middle word „kadaikku‟ is joined like this,
Example:”kadai+k+ku”
Subject+sandhi+plural
Sandhi is generating under the Tamil grammar rules. Here “k” represents க் .

B. *Methodology for Word Formation Rules(WFR)*

- Step1:Add Tamil input datas.
- Step2:Transliteration using bi-lingual machine translation method.
- Step3: Rule Based machine translation system involves Word Formation Rules(WFR)
- Step 4: Re- Transliteration using bi-lingual machine translation method.
  (or)
- Tamil output datas implemented.equation.

1) Eight WRF Rules in Tamil:

Any new word created by Word Formation Rules (WFR) must be a member of a major lexical category. The WFR determines the category of the output of the rule. In Tamil, the grammatical category may change or may not change after the operation of WFR. The following is the list of inputs and outputs of different kinds of WFR's in the derivation of simple words in Tamil.

- 1) Noun → Noun[[vElai]N + kAran]suf]N 'servant'
- 2) Verb → Noun[[padi]V + ppu]suf]N 'education'
- 3) Adjective → Noun[[walla]adj + thanam]suf]N 'good quality'
- 4) Noun → Verb[[uyir]N + ppi]suf]V 'to give life'
- 5) Adjective → Verb[[veLLai]adj + aakku]suf]V 'to make (something) white'
- 6) Verb → Verb[[cey]V + vi]suf]V 'cause to do'
- 7) Noun → Adjective[[uyaram]N + Ana]suf]adj 'high'
- 8) Verb → Adverb[[cey]V + tu]suf]adv 'having done'

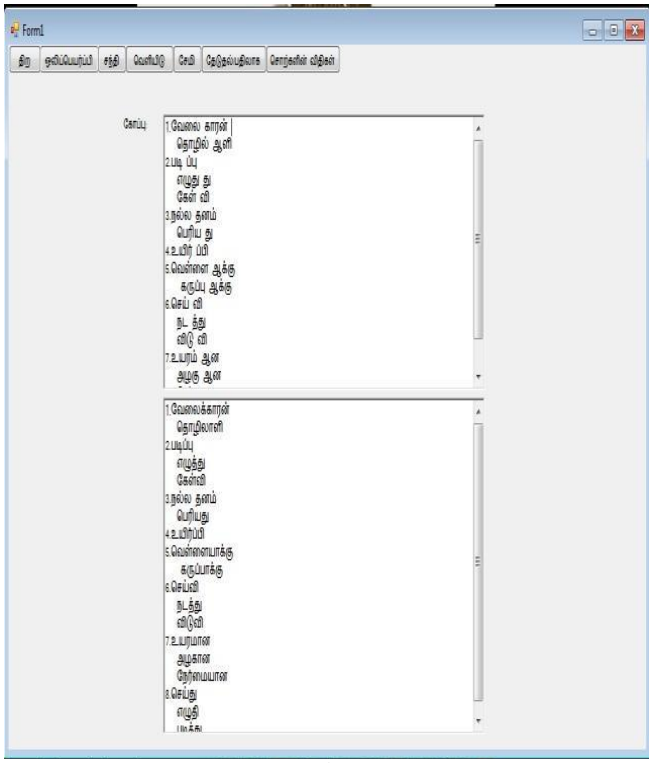

Fig. Generator of Word Formation Rules

C. Tamil Morphology:

The morphological analyzer identifies root and suffixes of a word. Generally, rule based approaches are used for morphological analysis which are based on a set of rules and dictionary that contains root words and morphemes. In rule based approach, a particular word is given as an input to the morphological analyzer and if the corresponding morphemes or root word is missing in the dictionary, then the rule based system fails. Here, each rule depends on the previous rule.

Example:
Root word: maram
Suffixes: ai,,āṉa, āka, atu

Root word + Suffixes = Morphological Analyzer
Morphological analyzer: maramai
maraththai
maram āka
maram atu.

D. Related Works:

There has been a large amount of interesting work in the arena of Transliteration from the past few decades.

i. Dhanalakshmi V,Anand Kumar M, Soman K.P,CEN, Amrita Vishwa, says "Word Formation rules is helpful in the situation of making a Tamil Grammar Tools.

ii. K.Rajan, Dr.V.Ramalingam, Dr.M.Ganesan, says "Sandhi Rule generator is helpful in the situation of making Insertion, Deletion and Alternation in Tamil Language/Linguistics".

iii. R.Akilan,Prof E.R.Naganathan, Dr.G.Palanirajan, says "Rule-based approach is applied for Plural marker. These rule-based approaches for Plural markers produce the result with accuracy. In future, using this approach we can develop a rule-based approach for the analyzing not only of Plural markers but also of other markers and grammatical variations."

IV. CONCLUSIONS

This Tamil Word Formation Rules is useful for making the Tamil grammar pattern. Patterns were created by their own utilization. Patterns are finite and infinite, using this word formation rules we able to produce tamil grammar tool, Spell checker, Search Engines, Information extraction and retrieval, Machine translation system, Content analysis and Question and answering system.

REFERENCES

[1] Dhanalakshmi V,Anand Kumar M, Soman K.P,CEN, Amrita Vishwa, "Natural Language Processing Tools for Tamil Grammar Learning and Teaching," International Journal of Computer Applications (0975 – 8887), Volume 8– No.14, October 2010.

[2] Antony P J and Dr. Soman K P, "Computational Morphology and Natural Language Parsing for Indian Languages: A Literature Survey," International Journal of Scientific & Engineering Research ISSN 2229-5518 IJSER © 2012, Volume 3, Issue 3, March-2012 .

[3] Madhu Ramanathan, Vijay Chidambaram, Ashish Patro, "An Attempt at Multilingual POS Tagging for Tamil," Phys. Rev.,

[4] R.Akilan,E.R.Naganathan, "POS TAGGING FOR CLASSICAL TAMIL TEXTS," International Journal of Business Intelligent, volume: 1 No: 01 January – June 2012.

[5] Nimal J Valath,Nasreedha Beegum, "Malayalam Noun and Verb Morphological Analyzer:A Simple Approach," International Journal of Software and Hardware of Engineering,ISSN No:2347-4890, volume: 2 Issue:8 August 2014.

[6] Jisha P.Jayan,Rajeev R R,Dr. S Rajendran, "Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation," International Journal of Computer Applications (0975 – 8887), Volume 13– No.8, January 2011.

[7] Veena Dixit, Satish Dethe, Rushikesh K. Joshi, "Malayalam Design and Implementation of a Morphology-based Spellchecker for Marathi,an Indian Language,"

[8] R.Akilan,Prof E.R.Naganathan,Dr. G.Palanirajan, "Morphological Analyzer for Classical Tamil Texts - A Rule based approach: special Reference to Plural Markers," Infitt Volume 7, 2011.

[9] Machine Learning of Sandhi Rules for Tamil, K.Rajan, Dr.V.Ramalingam, Dr.M.Ganesan, 2012.

[10] "Natural Language Processing and Information Retrieval", Tanveer Siddiqui,U.S.Tiwary..