VOL. 10, NO. 9, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

GRAMMAR CHECKER FEATURES IN MODERN TAMIL NATURAL LANGUAGE PROCESSING

K. Nirmala and M. K. Kalpana
Department of Computer Science, QMGCW(A),Chennai, India
E-Mail: nimimca@gmail.com

ABSTRACT

Generally, The NLP (Tamil) applications are programming with different kinds of input data. Inputs classified into Text, Image, sound waves etc., Tamil Text based applications are creating under the word formation techniques. These words analysis and generation are activating in these ways, i) Untagging & Tagging and ii) Word-level and Character-level accuracies. This method is processing based on sequence labeling. These are the features for creating preprocessing stages for build the Dravidian applications.

Keywords: Tamil unicode, grammar, word and character analyzer, interlingua machine translation, agglutinative language, rule based machine translation, natural language processing, words formation.

1. INTRODUCTION

The Tamil NLP applications development is giving greatest level of challenges for researchers and programmers. The Tamil language is classified into three, i) Classical/old Tamil, ii) Middle Tamil and iii) Modern/colloquial Tamil. The colloquial Tamil is spoken by the south Indian people. This modern Tamil is not allowing other languages to involving in communication. Word Formations generations are common methods for constructs Grammar based applications. This paper describes about above given information elaborately.

2. TAMIL UNICODE

The Tamil Unicode is having a set of Independent vowels, Dependent Consonants and symbols. The Independent vowels are named as uyir ezhuthukkal in tamil. Tamil Unicode letters are mentioned in below of the content. 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 South India. It is generates with other 9 North Indian letters. Figure-1 explains Tamil Unicode Consortium Chart from Wikipedia.

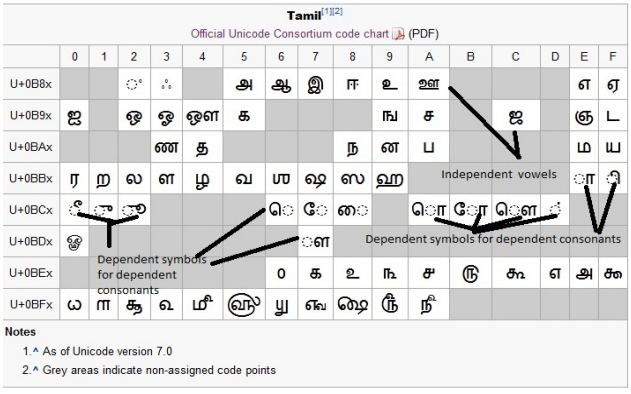

Figure-1. Official unicode consortium chart from Wikipedia.

2.1. Tamil unicode letters

Tamil Unicode letters describes like below in Figure-2.

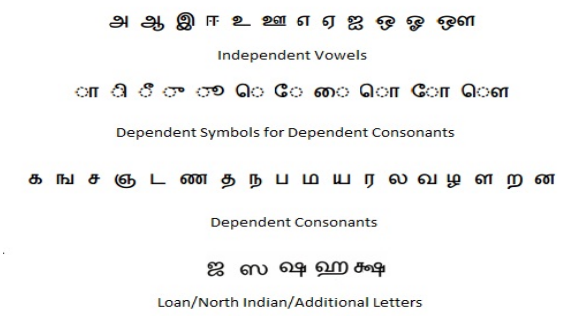

Figure-2. Tamil unicode letters.

3. WORD AND CHARACTER ANALYZER IN MODERN TAMIL

3.1. Word analyzer

In Tamil word analyzer explains about their frequencies for splitting the sentences/paragraphs/files into a separate word and implements some techniques to perform the number of occurrences and percentage of the words occurrences.

Figure-3 elaborates about the performances of Frequency Word Analyzer is clearly explained in three individual columns.

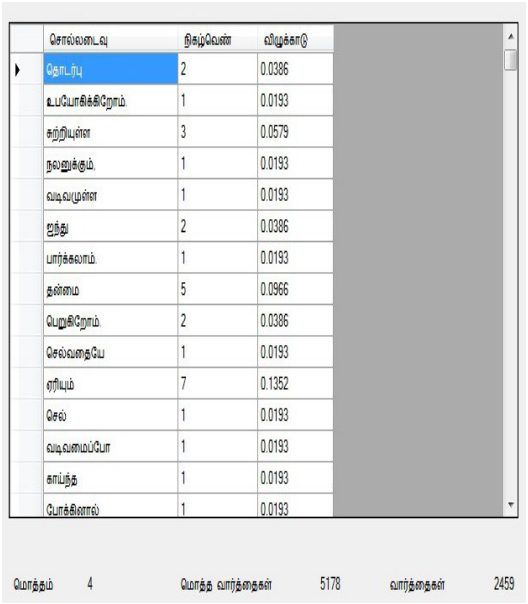

Figure-3. Frequency word analyzer/words splitting.

1) Text data storage

In Figure-3, Tamil Text files are used like data storage, and it can be segment characters, words and sentences. This text data storage provides the information to produce and maintain data storage of Tamil text. This text analysis functions are counting, searching, filtering, arranging and specification.

2) Text data search

It is used to finding the word form, number of terms and length. This is retrieving the information from a particular topic or text.

3) Ability to use of arranging words

The words can be arranged with the help of Tamil characters arranging the words using Tamil alphabetical process. It is arranging the words by their endings.

3.2. Character analyzer

Character analyzer splits words, sentences, paragraphs and files into individual characters and calculates the number of characters. Characters splitting methods using string data type is giving wrong characters segmentations and counts. Char data type giving correct expected results. Using this direct machine learning method, rules generating process is not possible. Figure-4 string based and Figure-5 character based analyzer explains the difficulties between Tamil Unicode letters with direct machine translation.

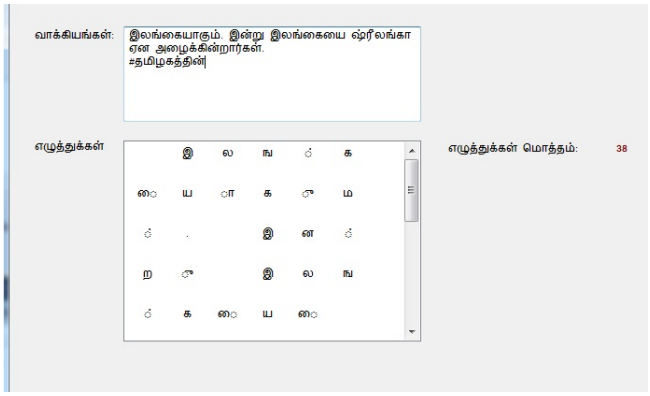

Figure-4. String based direct machine translation.

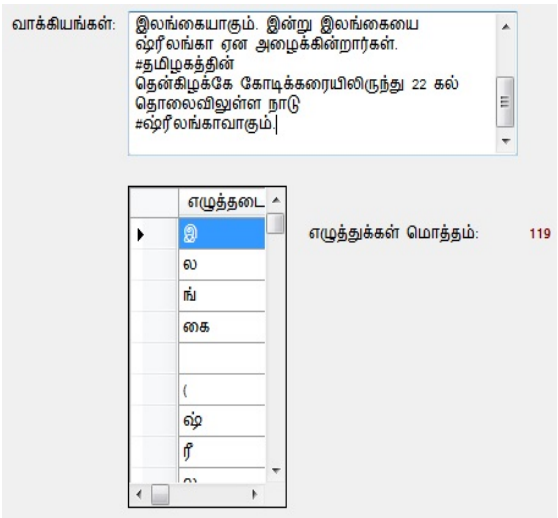

Figure-5. Character based direct machine translation.

4. INTERLINGUA MACHINE TRANSLATION

The source language text is converted into a language independent meaning representation. Interlingua machine translation is otherwise called Transliteration/Indirect/Bi-lingual/cross lingual machine translation. Tamil lingual information transferring to diacritic forms and diacritic forms to Tamil lingual. Using this machine translation method, Tamil Grammatical rules are applying in grammar and linguistics based applications/Tools. The applications are Spell Checker, Information extraction and retrieval, simple machine translation system. Grammar checker, content analysis, Question answering system, automatic sentence analyzer, Speech and Dialogue system, knowledge representation in learning and automatic assessment tool. The Evaluation of Transliteration system can be done manually or automatically by the use of metrics like Transliteration Accuracy. For evaluating the performance of this transliteration system, transliteration accuracy of the transliterated system is checked by determining the total number of Tamil input data generated divided by total number of English generated transliterations. Figure-6(a) defines bi-lingual transliteration. In Indirect machine

translation Tamil Unicode converts into English diacritic or dialogue letters for applying the rules. This is elaborated in Figure-6(b).

Figure-6(a). Bi-lingual transliteration.

Vowels -12		Consonants – 18	
அ	a	க்	k
ஆ	ā	ங்	ṅ
இ	i	ச்	c
ஈ	ī	ஞ்	ñ
உ	u	ட்	ṭ
ஊ	ū	ண்	ṇ
எ	e	த்	th
ஏ	ē	ந்	n
ஐ	ai	ப்	p
ஒ	o	ம்	m
ஓ	ō	ய்	y
ஔ	au	ர்	r
ஃ	ḥ	ல்	l
		வ்	v
		ழ்	z
		ள்	ḷ
		ற்	ṟ
		ன்	ṉ

Figure-6(b). Tamil unicode to diacritics.

5. RULE BASED MACHINE TRANSLATION

A rule based machine translation system consists of collection of rules called grammar rules, lexicon and software programs to process the rules. It is extensible and maintainable. Rule based approach is the first strategy ever developed in the field of machine translation. Rules are written with linguistic knowledge gathered from linguists. Rules play major role in various stages of translation: syntactic processing, semantic interpretation, and contextual processing of language. At present rules involving in previously discussed Interlingua machine translation system.

6. WORD FORMATION RULES

Any new word created by Word Formation Rules (WFR) must be a member of a major words type. The formation of words regulates the kind of the result of the rule. Tamil grammar type changes are probable things, after the operation of WFR. The collections of information and results of different kinds of WFR are in the descent of simple words in Tamil.

6.1. Sandhi rules generator

Sandhi grammar rules handles with all types of activities like insertion, cancellation, replacement etc., when uni, bi or n number words occur together is named as Tamil adjoining letters/Sandhi. Generally, the last letter of the first morpheme or word and the beginning of the following letter are taken into account in Sandhi process. The finishing and the beginning of words formations are inserted can be either a vowel or a consonant or consonant group. There are many combinations of such characters. These Modern Tamil Sandhi rules should be explicitly specified for morphological analysis in a rule based system. These rules can be trained automatically by the system from the training samples and subsequently be applied for new inputs.

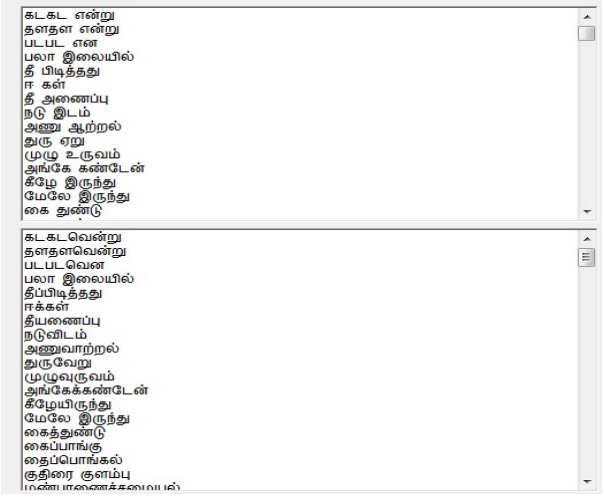

Figure-7. Sandhi rules generator.

i) 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 '.

6.2. 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'

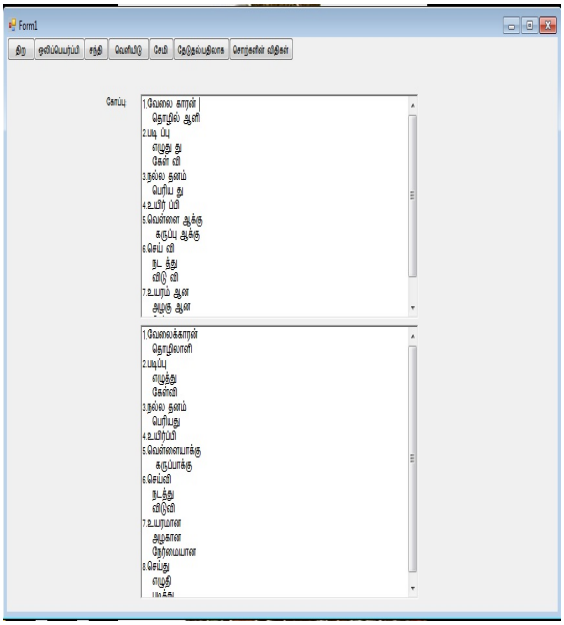

Figure-8. Demonstrate about above mentioned word formation rules in GUI.

6.3. Morphological analyzer

Tamil Word Analyzer/Morphological Analyzer identifies root, suffixes of a word and naming an attributes using Modern Tamil grammatical terms.

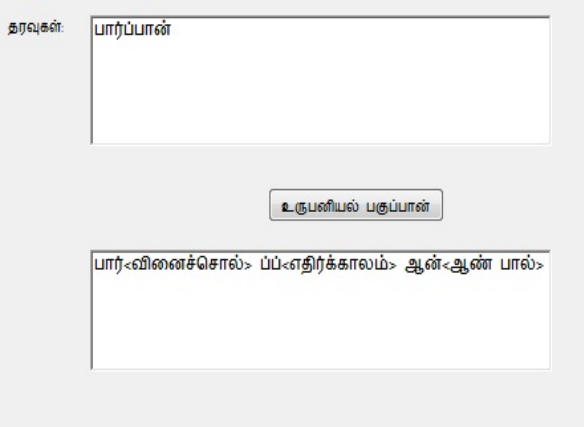

Figure-9. Morphological analyzer GUI.

6.4. Morphological generator

Word formation/Morphological Generator is combining the words using Sandhi rules; root with suffixes and with some grammatical suffixes in Tamil computational methods. Letters and joining the words is used for creating the words. Figure-8 and Figure-9 defines about the initial stages of morphological Generator.

7. CONCLUSIONS

This untagged or tagged character based grammatical checkers is solving the grammatical errors with high accuracy compare than untagged or tagged word based grammatical checkers. A GUI is used to improve end-users usability. Preprocessing, Word Segmentation and word rules labeling describes the process of morphological analyzer.

8. FUTURE ENHANCEMENT

Using, Character Based Sandhi Rules, Morphological Analyzer and Morphological Generator, Researchers and Programmers can able to create a NLP application. These are the techniques bringing next to the stage of Parts of Speech for making the sentences generation or sentences error corrections.

REFERENCES

[1] Dhanabalan T., Ranjani Parthasarathi and T. V. Geetha. 2003. "Tamil Spell Checker." Sixth Tamil Internet, Conference, Chennai, Tamilnadu, India.

[2] Anand Kumar M. *et al.* 2010. "A sequence labeling approach to morphological analyzer for tamil language." IJCSE) International Journal on Computer Science and Engineering. Vol. 2, no. 06, pp. 1944-195.

[3] FAkilan, R., and E. R. Naganathan. 2014. "Morphological Analyzer for Classical Tamil texts: A rule-based approach".

[4] Tapaswi, Mrs Namrata, Suresh Jain, and Mrs Vaishali Chourey. "Morphological-based Spellchecker for Sanskrit Sentences.".

[5] Dhanalakshmi, V., and S. Rajendran. 2010. "Natural Language processing Tools for Tamil grammar Learning and Teaching." International Journal of Computer Applications.

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

[7] “Natural Language Processing and Information Retrieval (Oxford Higher Education)" by U. S. Tiwary (Author), Tanveer Siddiqui (Author).

[8] Habash, Nizar, Owen Rambow, and Ryan Roth. 2009. "Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization." Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

[9] Banerjee, Pratyush, *et al.* 2012. "Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data." Proceedings of the 16th Annual Meeting of the European Association for Machine Translation, Trento, Italy.

[10] Jabbari, Fattaneh, S. Bakhshaei SM Mohammadzadeh Ziabary, and S. Khadivi. 2012. "Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus." Association for MachineTranslation in the Americas (AMTA 2012).

[11] Bigi, Brigitte. 2011. "A multilingual text normalization approach." Proceedings of 5th Language && Technology Conference-The 2nd LRL Workshop.

[12] Iacoboni, Giorgio, and Andrea Di Cagno. 2012/2013. "NATURAL LANGUAGE PROCESSING Course Project in Tokenization, Normalization, and Lemmatization".

[13] Al-Anesi, Bushra Abdullah, and Khalid Omar Thabit. 2012. "An Arabic NLP System for Information Management within Organizations." IJIPM: International Journal of Information Processing and Management, Vol. 3, no. 4, pp. 10-18.

[14] http://en.wikipedia.org/wiki/Morphologylinguistics.