DOI: 10.21917/ijsc.2015.0153

MODERN THAMIZH SANDHI RULES GENERATOR IN NLP

K. Nirmala[1] and M.K. Kalpana[2]
Department of Computer Science, Quaid-E-Millath Goverment College for Women, India
E-mail: [1]nimimca@yahoo.com, [2]kalpanamalathkar.research@gmail.com

Abstract

Thamizh sandhi rules generator deals with addition, deletion, getting changes with existing Information and this adjoining letters/sandhi grammar rules processing with indirect/bi-lingual machine translation. This Modern Thamizh Sandhi Rules Generator is implemented under Unicode based Indic Script. Modern Thamizh sandhi generation is the initial stage of developing the Word Formation rules in Thamizh computational method.

Keywords:

Thamizh Unicode, Diacritical Markings, Bilingual Machine Translation, Sandhi Rules, Computational Generator

1. INTRODUCTION

A rule based machine translation system consists of collection of rules called grammar rules; lexicon and software programs are using 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.

1.1 LITERATURE SURVEY

Tamil is an agglutinative and concatenative language, where morphemes are strung together to form long words. There are free morphemes and bound morphemes. The bound morphemes act as affixes which combine with other morphemes to form inflectional and derivational categories. Affixes can be realized in many ways. Affixes cannot be attached one after another in a free order. Thus in 143 designing language analyzers, one should design and implement the mechanism that performs the phonological alternations and check the validity of ordering for the realization of morphemes [1].

When adding a particular morpheme with another morpheme, the changes occurred depends not only on the characters but also on the type of the morpheme. In the following case, unless we know whether the first member of the given combination is a noun or a verb or something else, it may not be possible for us to predict the resultant form.

நாடு N + ஐ →நாட்டை
நாடு V + ஐ →Not applicable
படி N + ஆல் →படியால் (Instrumental case)
படி V + ஆல் →படித்தால் (Conditional verb)

2. MODERN THAMIZH

Modern thamizh is a colloquial spoken thamizh, it shows with numerous changes compared with classical and middle age thamizh. In recent stage of thamizh consists of European-style punctuation and the use of consonants and vowels groups. Thamizh movement not allows other languages in the way of reading and writing. These Modern Tamil NLP Applications are used for making several kinds of applications in recent trends. This approach is easier than the input of Classical and middle Tamil.

3. THAMIZH UNICODE

Thamizh Unicode Script is having independent vowels, Independent consonants and dependent consonants; Fig.1 explains the details of Thamizh Unicode Characters in computational method.

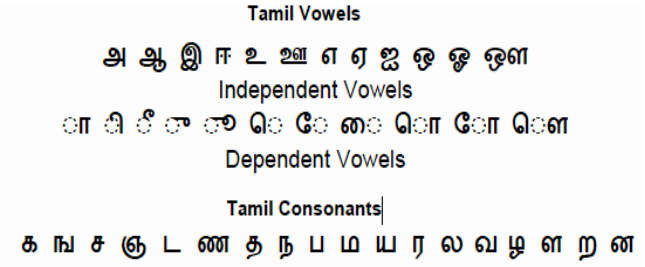

Fig.1.Thamizh characters set of computational method

3.1 PROBLEMS IN DIRECT MACHINE TRANSLATION

Direct machine translation systems provide direct translation, i.e., no intermediate representation is used. A Direct machine translation carries out word-by-word translation. Using direct machine translation method is not possible in Thamizh rule-based analysis/Thamizh NLP applications. The Fig.2 Character analyser using Direct Machine Translation doesn't separated its Unicode characters like an actual Thamizh characters.

English characters/alphabets does not having any dependent vowels in computational method. But, Indic scripts Because of this difference direct machine translation is not possible in Thamizh Natural Language applications development. For this purposes parallel/cross-lingual/bi-lingual transliteration methods are handling in Thamizh computational research areas.

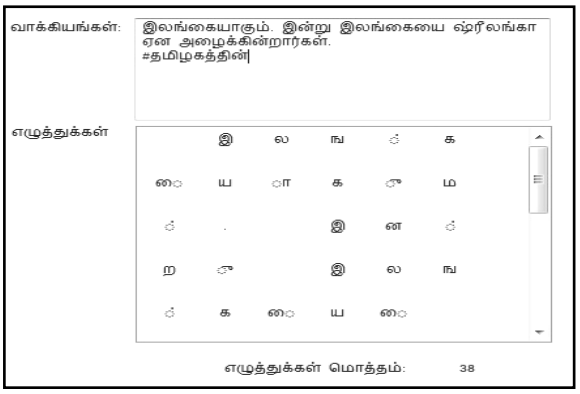

Fig.2. Character analyzer using Direct Machine Translation

3.2 INTERLINGUA OPERATIONS USING IN THAMIZH

The Fig.4, represents the information of bilingual/cross-lingual machine translation system. The source language text is converted into a language independent meaning representation called 'interlingua'. Translation is thus a two-stage process, analysis and synthesis, as shown in Fig.3 Interlingua based translation model.

In 2000, Jurafsky and Martin approach defines, "An interlingua represents all sentences that mean the something in the same way regardless of the source language they happen to be in". The amount of analysis is much more needed compare than Transfer based machine translation.

An Interlingua system has to resolve all ambiguities so that translation to any language can take place from Interlingua representation. Another advantage of Interlingua is that it is a meaning based representation and can be used in applications like information retrieval.

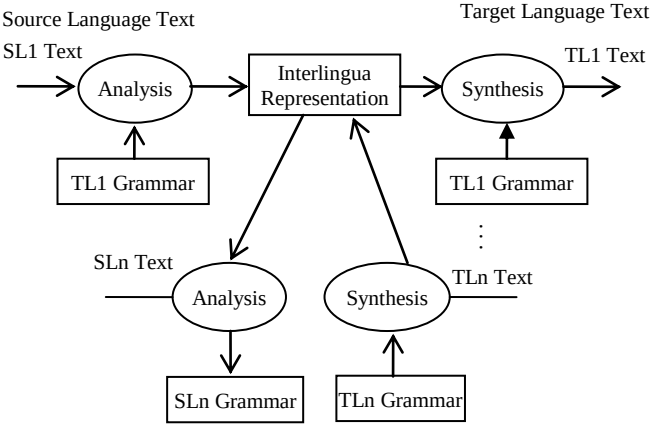

Fig.3. Interlingua based translation model

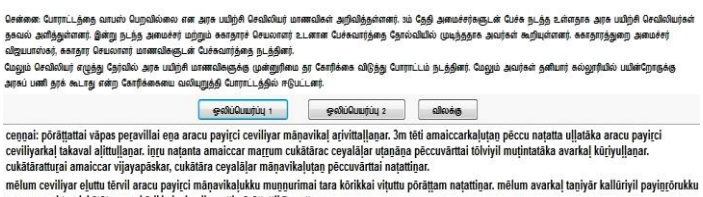

Fig.4. Example of Thamizh to English Transliteration

4. THAMIZH UNICODE TO DIACRITICS

The Thamizh script is a syllabic alphabet script that is used by thamizhians in India, Srilanka, Malaysia and elsewhere. Diacritics not represented Thamizh alphabets. Below listed Fig.5 explains about Thamizh letters (Unicode) and Diacritics formations.

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

Fig.5. Thamizh letters and it Diacritics

4.1 THAMIZH DIACRITICS FORMATION:

The Modern Thamizh complete scripts, consists of the thirty-one letters in their independent form, and an additional 216 combining letters representing a total 247 combinations including Sanskrit using letters 256.

4.2 BI-LINGUAL MACHINE TRANSLATION

The Fig.4 explains about the Transliteration method using in Thamizh computational method. It elaborates the conversation of Thamizh Unicode letters to English diacritical marking texts. From the Fig.4, we come to know about the problems in handling Thamizh Unicode in direct machine translation system. These kinds of issues were avoided in the indirect/bi-lingual/cross-lingual/inter-lingua machine translation.

4.3 USES OF DIACRITICAL MARKING METHODS

Using this Diacritical marking letters in our Thamizh machine translation, we can able to apply rules using several methods. There are, Lexicon, Syntactic, Semantic and Morphological analysis approaches.

5. SANDHI RULES AND GENERATOR

Thamizh grammatical sandhi rules, some text manipulation process was generated in a particular computational method.

Using Indirect Machine Translation Sandhi rules were successfully implemented.

Below the paragraph sandhi rules and generator is shown in Fig.6.

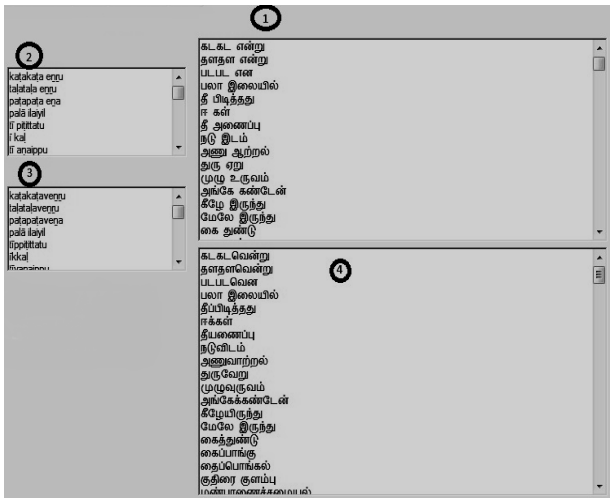

Fig.6. Thamizh Sandhi Rules and Generator

The Fig.6 shows an example of Rules based machine translation. Rules processed under the adding, deleting and alternating the Thamizh diacritical texts.

6. SANDHI RULES USING DIACRITICAL MARKINGS

As per Fig.6: No.1 represents Information retrieved from the Unicode text file. No.2. represents transliterated the Thamizh characters into English diacritical characters. No.3. generated sandhi Thamizh grammatical rules in the English diacritical characters using ranges of the characters. Finally, No.4 defines, transliterated diacritical characters again transliterated into Thamizh Unicode Indic script or results.

Example:

"Ganesh kadaikku cenran".

Kadaikku//Verb

Here, in middle word 'kadaikku' is joined like this,

Example: "kadai+k+ku"

Subject+sandhi+noun suffix

Sandhi is generating under the Thamizh grammar rules. Here 'k' represents '**க்**'.

7. CONCLUSION

The Sandhi rule based machine translation is used to form a Words/Morphological analyzer, Morphological Generator and Unicode diacritic text to speech. This initial grammar generating techniques are used to create different word levels in modern thamizh machine translation.

8. REFERENCES

[1] K. Rajan, V. Ramalingam and M. Ganesan, "Machine Learning for Sandhi Rules in Tamil", *Proceedings of the 11th International Conference INFITT*, pp. 141-146, 2012.

[2] Manji Bhadra, Surjit Kumar Singh, Sachin Kumar, Subash, Muktanand Agrawal, R. Chandrasekhar, Sudhir K. Mishra, Girish Nath Jha, "Sanskrit Analysis System (SAS)", *Sanskrit Computational Linguistics*, Vol. 5406, pp 116-133, 2009.

[3] Pawan Goyal, Vipul Arora and Laxmidhar Behara, "Analysis of Sanskrit Text: Parsing and Semantic Relations", *Sanskrit Computational Linguistics*, pp. 200-218, 2009.

[4] Priyanka Gupta1 and Vishal Goyal, "Implementation of Rule Based Algorithm for Sandhi - Vicheda of Compound Hindi Words", *International Journal of Computer Science Issues*, Vol. 3, pp. 45-49, 2009.

[5] Chimsuk Tawee and Surapong Auwatanamongkol, "An Incremental framework for a Thai-English Machine Translation System using a LFG tree structure as an Interlingual", *International Journal of Computer Science and Engineering*, Vol. 2, No. 2, pp. 280-288, 2010.

[6] Judith Francisca Islam, Mohammad Mamun Mia and Dr. S. M. Monzurur Rahman, "Adapting rule based machine translation from English to Bangla", *Indian Journal of Computer Science and Engineering*, Vol. 2, No. 3, pp. 334-342, 2011.

[7] S. Saraswathi, P. Kanivadhana, M. Anusiya and S. Sathiya, "Bilingual Translation System", *International Journal of Computer Science and Engineering*, Vol. 3 No. 3, 2011.

[8] B. Krithika, V. Ramalingam and K. Rajan, "Performance of machine learning methods for classification tasks", *International Journal of Computer Science and Engineering*, Vol. 5, No. 6, 2013.

[9] T. Kameswara Rao and T. V.Prasad, "Key Issues in Vowel Based Splitting of Telugu Bigrams", *International Journal of Advanced Computer Science and Applications: Special Issue on Natural Language Processing*, pp. 9-16, 2014.

[10] Omar Shirko, Nazlia Omar, Haslina Arshad and Mohammed Albared, "Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach", *Journal of Computer Science*, Vol. 6, No. 3, pp. 350-356, 2010.

[11] Arabic-Malay Machine Translation Using Rule-Based Approach, available at http://ww.itimes.com/citizen-journalism/arabic-malay-machine-translation-using-rule-based-approach.

[12] U. S. Tiwary and Tanveer Siddiqui, "*Natural Language Processing and Information Retrieval*", Oxford University Press India, 2008.

[13] http://en.wikipedia.org/wiki/Tamil_scrip