Natural Language Processing and Same Format of Computational Indic Scripts

(Unicode for Indian languages)

Dr.M.K.Kalpana,
Assistant Professor,
Department of Software Application, A.M.Jain College, Chennai, India

***Abstract*—In India people spokes 22 native languages. In the company of Sandhi is a set of grammatical rules, Sandhi rules assist many languages in India, including specifically known as Malayalam, Assamese, Bengali, Kannada, Pali, Hindi, Marathi, Telugu and Sanskrit. Sandhi is merging the two words with the help of the independent vowels and dependent consonants. Above mentioned specific languages are having same format of letters formation in scripts. Accompanied by, Sandhi rules also following the same format of rule generation in these specific languages.**

***IndexTerms*— Indic Scripts, Unicode, Transliteration, Sandhi, Natural Language Processing.**

I. INTRODUCTION

Technically, sandhi rules set is generate grammar checker for specified sandhi-based error-detection and correction. Researcher explains about the same format of letters, Unicode and sandhi formations in many languages in India, it will be helps the researcher to create a more technical based applications in future. Sandhi is the formation of two letters such as Independent vowels and Dependent consonants in manual language. Unusually, Sandhi formations differed from the computational method. English is accepts both ASCII/ANSI Character and Unicode/String types. Other than that, excluding English most of the natural languages not get accepted by ANSI/ASCII Character types, these languages are getting accepted by Unicode based fonts. Apart from that, Rule-based transliteration and translation methods are helps the researchers and programmer to create a language-based application in many areas of the research.

II. INDIC SCRIPTS

Indic Scripts is explained as a Indian language Unicode/typing fonts. It has same formats of Unicode formation by Unicode consortium. In this method, Indic Scripts was splitted into three categories, 1.Independent Vowels, 2.Independent Consonants and 3. Dependent consonant symbols. Unicode fonts divisions named as UTF-8, 16, 32 and UCS-2. Unicode involving many concepts such as, String, Hexadecimal, Byte, Bit, Binary number and Typeface etc., Under the following few Unicode letters explained about the narration in Hindi language.

◌ः	+	क	=	क:
Dependent consonant Symbol	I	Independent Consonant		Dependent Consonant

अ=अ ➔ Independent Vowel

III. TRANSLITERATION

Transliteration is otherwise called as Indirect Machine Translation or Interlingua or bi-lingual under Rule-based Machine Translation. It is handling for passing the rules on the represented source text in any language. Some of the languages not getting supported by Direct Machine Translation, on that period, the programmer or researcher can used the transliteration methods for their own languages. For converting the native language to English source text researcher or programmer are using many kind of English source text sets types. These types sets named as Unicode, Bamini, Shreelipi, Softview, Diacritics, TAB, TACE, TAM, TSCII etc., Conversation is made up of many tools and applications for passing the rules.

Example of Transliteration:

लक्ष्मण राम का एक छोटा भाई है - lakshman raam ka ek chhota bhaee hai

Meaning of this sentence in English: laxman is a younger brother of rama.

IV. SANDHI

In real-time researcher is made Sandhi with two combinations, one is Independent vowels and Dependent Consonants. Sandhi is classified into two, 1. Internal Sandhi and External Sandhi. Internal sandhi did their string manipulations/expressions with in a word. But, External sandhi placed the manipulation/expression between two words. Computationally, sandhi is first transliterated as English source text, sequentially sandhi grammar rules given to the transliterated contents for generating customized output. Classifications of machine translations are, Direct Machine Translation, Rule-based, Corpus-based and Knowledge-based, In Indic Scripts, Sandhi is using Indirect Machine Translation, it is comes under Rule-based Machine Translation for passing the sandhi grammar rules under Classical Information Retrieval. In existing stage of research, author used Euclidean decision tree/table algorithm for the sandhi grammar checker, including the features of transliteration, string manipulation and Classical Information Retrieval. It is achieved 95.6% of Accuracy for Insertion & Alternation and 99.9% achieved for Deletion and Alternation. Author recognizes that, which languages are having same formats of letter formations those languages are able to score this same 95.6 and 99.9 percentage of accuracy in sandhi rules set.

A. SANDHI IN MANY INDIC SCRIPTS

Languages	Internal Sandhi		Grammar Rules Descriptions	External Sandhi		Grammar Rules Descriptions
	Original Text	Source Text		Original Text	Source Text	
Sanskrit	वच् + अन्ति → वचन्ति	*vac + anti → vacanti*	No sandhi change of any kind occur	भगवन्त् → भगवन्	*bhagavant → bhagavan*	Removing extra consonants
Marathi	सूर्यास्त सूर्य+अस्त	suryosth→ surye+asth	Sandhi change any kind occur	Common rules are following for both Internal and External sandhi		
Hindi	हिम + आलय = हिमालय	Him+aalya→*hi maalya*	Sandhi change any kind occur	Common rules are following for both Internal and External sandhi		
Tamil	எந்தப் பக்கம்	entap pakkam	Sandhi change any kind occur	சிற்சில	circila	Sandhi change any kind occur
Telugu	Common rules are following for both Internal and External sandhi			రాముడు + కోరకు= రాముడికోరకు	Ramadu+koraku→*ramadikoraku*	Sandhi change any kind occur
Malayalam	പര+ഉന്നു→പരയുന്നു	para+unnu→*parayunnu*	Sandhi change any kind occur	Common rules are following for both Internal and External sandhi		
Kannada	ಮರದೇಲೇ→ಮರದ+ಏಲೇ	maradele→maradda+ele	Sandhi change any kind occur	ದೇವವಾಲಯ→ದೇವವ+ಆಲಯ	deevaalaya→deeva+aalaya	Sandhi change any kind occur

Table 1. Internal and External Sandhi in many Indic Scripts

Above mentioned languages are having same format of internal and external sandhi rules for creating a grammar checker using Indirect Machine Translation using classical information retrieval under Rule-based Machine Translation. For an instance, researcher gave a screenshot below in Figure 1 and 2.

V. NATURAL LANGUAGE PROCESSING

NLP stands for Natural Language Processing. It is processing with all over the scripted and script less (Verbal/Non-Verbal) languages. It is functioning with different kinds of data/information. These functionalities or research domains named as Text mining, Digital Image Processing, Cognitive Sciences, Speech Synthesis, Speech recognition and Digital Video Processing. The research domains are includes, the features of Natural Language Processing includes computational methods and linguistics knowledge. This research domain fallen into ANN (Artificial Neural Network). Sandhi-based grammar checker using Indic scripts as their own flow of ANN, it is represented in Fig.2.

Figure:

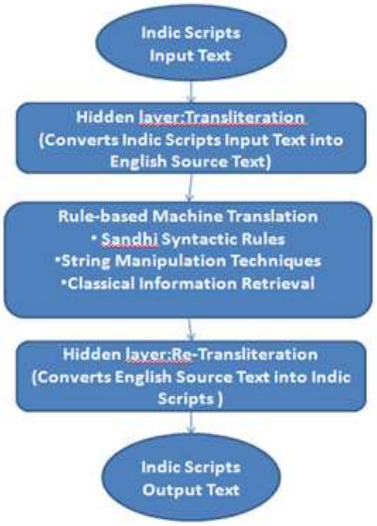

Figure.1. Model of Sandhi checker for Indic Scripts

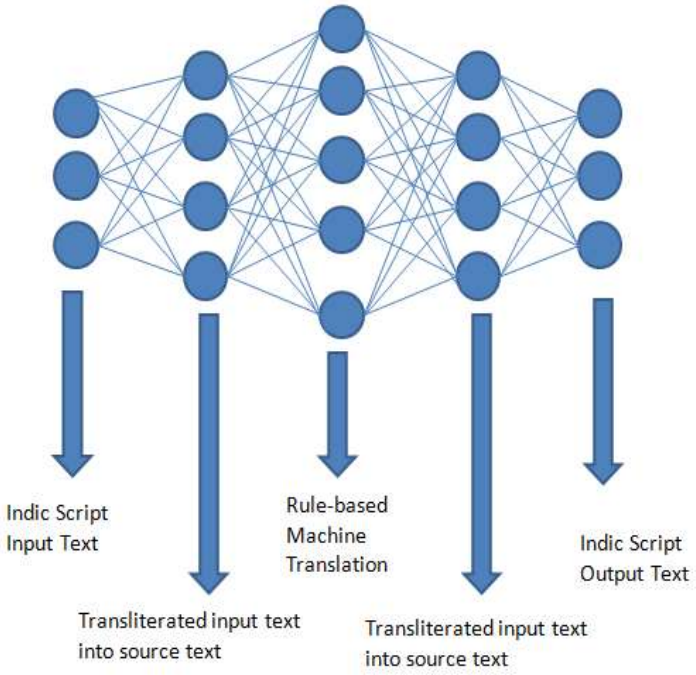

Figure.2.ANN Model of Sandhi Checker for Indic Scripts

VI.CONCLUSION

Under these techniques, the researchers determine the information about the sandhi error detections and corrections on many Indic scripts is having highest accuracy in the terms of transliteration, manipulation and classical information retrieval in the category of Rule-based Machine Translation. Following this predictions, we able to do real-time NLP application for the end users.

REFERENCES

[1] Vijaya, M. S., et al. "English to tamil transliteration using weka." International Journal of Recent Trends in Engineering 1.1 (2009): 498-500.

[2] M.K.Kalpana, Dr.K.Nirmala, Modern Tamil Sandhi Rules Training Sets Results in Weka, International Journal of Applied Engineering Research, Research India Publications, India, Vol.10 No.20 (2015).

[3] Barbieri, Francesco, et al. "SemEval 2018 Task 2: Multilingual Emoji Prediction." Proceedings of the 12th International Workshop on Semantic Evaluation. 2018.

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

[5] Anand Kumar, M., et al. "A sequence labeling approach to morphological analyzer for tamil language."(IJCSE) International Journal on Computer Science and Engineering 2.06 (2010): 1944-195.

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