International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Letter based Processing of Indic Script – Malayalam Case Study

Rejitha K.S.[1]

[1]*Senior Resource Person, LDC-IL, CIIL, Mysore, India*

---***---

Abstract - *Major Indian language scripts are originated from the Brāhmī script. The scripts today are encoded in electronic format using Unicode standards. Unicode has a catalog of millions of characters. In order to process the Indic scripts, which presents such complexity, certain procedures and algorithms have to be developed. Extracting meaningful combinations of letter is a vital process for many NLP applications. This paper gives an overview of the algorithm for text processing which can be adopted to process the Indic scripts in the context of multi-byte character sets using Malayalam Script as a case study.*

Key Words: Indic scripts, Malayalam letters, Writing system, Encoding, Multi-byte character

1. INTRODUCTION

Letters in Indic scripts are multi-byte character sets which take the combinations of vowels, consonants and various signs in different shapes. Adding various signs like Matra, Anuswaram, Visargam and Chandrakala with these combinations of vowels and consonants, multiplies the complexity of finding the boundary of a single letter. Malayalam letter patterns are like CV, CCV, CCCV etc. For example, the single Malayalam letter 'vaː' ('വാ') is a combination of CV, having 4 bytes whereas the single Malayalam letter 'striː' (സ്ത്രീ') is the combination of CCCV, having 12 bytes. Thus in text processing it is vital to identify the boundary of letter in Indic Scripts. Simple ASCII characters and punctuations are also subsets of this complex character set. This complex clustering of multi byte character sets poses certain challenges for computer programmers.

2. SCRIPT EVOLUTION

The development of writing is comparatively a recent phenomenon and the alphabetic writing is traced only around 3000 years ago. Before there were alphabets when some pictures represents some information in a particular way and is called as pictograms. Later these pictures changed into more symbolic form and which is considered as ideograms. The pictograms and ideograms turned into more abstract forms which turned into writing systems, symbols represents the words in a language then they named as logograms i.e., the abstract form of symbolizing the real world entity.

Writing has been developed in different way in different period. It can be classified mainly as Logographic, Syllabic and Alphabetic. In logographic writing system signs are used to represent the morpheme. Chinese is following the logographic writing system but not completely. In syllabic writing system each sign correlate to a syllable. Japanese can be written with a symbol which is the representation of spoken syllable. There are two categories in alphabetic writings and they are Abugida and Abjad. An alphabet is a set of written symbols which represents a single sound. It has graphemes of vowels and consonants. In Abugida consonant - vowel sequence has been written as a single unit. Indic script falls in this category. Abjad lacks the representation of vowels. Arabic language is an example of Abjad.

3. MALAYALAM SCRIPT EVOLUTION

A language is a method of communication and the purpose is to convey our thoughts and understand other's. To transcribe a language one requires a set of alphabets. This set of alphabets is called script. The scripts include letters and numbers. A letter is a graphic representation of the smallest unit of spoken sound. Many languages can share a single script and a language can be written in different scripts. For e.g. Hindi, Marathi and Nepali are written in Devanagari script and Konkani is written in Devanagari, Malayalam and Kannada script. Malayalam is also written in other scripts such as Arabic script then it is called Arabi Malayalam and Syriac Malayalam is Malayalam written by using Syriac script.

The letters and script are related to each other in Malayalam. Malayalam was written by 'Vattezhuthu' which is developed from the Brahmi script from the 9th century onwards. The Pallava kings' popularized Grantha script since the Vattezhuthu had no aspirated consonants and which were not sufficient to record the Sanskrit language. During the reign of Pandya and Vijayanagara kingdoms a version of Grantha script developed in the Malabar region and to form scripts of Dravidian languages like Tulu and Malayalam. The alphabets of modern Malayalam are taken from the 'Grantha Script'. The Malayalam script is also used to write some minority languages like Paniya, Betta Kurumba, and Ravula.

4. MALAYALAM LETTERS

In Malayalam Script, the speech sound is divided into two and they are vowels and consonant. The vowels are produced with the free flow of air in the vocal tract and consonants are articulated with the complete or partial closure of the vocal tract. In every language a certain pattern of sound combination is allowed and which is embedded in each speaker's language knowledge. This basic pattern of segment is called syllable. So a syllable may be a vowel or a combination of consonant/s and a vowel. However a syllable contains a vowel or a vowel like sound. A syllable has two parts; one is onset and other is rime. Onset contains one or more consonants and rime contains vowel or vowel and consonant/s. The vowel in the syllable is called nucleus. The other consonants following the nucleus is called coda. Onset and coda have more than one consonant then that syllable is called consonant cluster. But Malayalam has a special consonant letter called chillaksharam or chillu that exist without the help of vowel, i.e, it is a pure consonant.

All Indic scripts run left to right, although some combining glyphs appear to the left of their base character for display.1 In Malayalam vowels changes their position and shape when attached to a consonant in irregular way. That means the glyph of a letter has different forms. When vowels are attached to consonants then these vowels are orthographically represented in different form known as vowel matra.2
Malayalam has 15 vowels 36 consonants and 5 chillaksharam.

The possible valid Malayalam letters will be

- A vowel
- A vowel + Yogavaha (Anuswaram, Visargam)
- A consonant
- One or more consonant + Matra
- One or more consonant + Matra + Yogavaha
- One or more consonant + Chandrakala which is at end of the word
- Chillaksaram or Chillu which is a pure consonant

5. CHARACTER REPRESENTATION IN BINARY ENCODING

American National Standards Institute (ANSI) came out with the idea of a 7- bit code. In 1963, ANSI announced the American Standard Code for Information Interchange (ASCII) and computer manufacturers from US adopted this code as standard. Later IBM brought its own standard of 8-bits called as EBCDIC which means 'Extended Binary Coded Decimal Interchange Code'. ISCII (Indian Script Code of Information Interchange) was evolved as a standard for Brahmi based Indic Scripts by 1991. ISCII is a single encoding representation for all the Indian Scripts with the concept of Multi-byte expression. It was using upper ASCII code points (128-255 Range of 8 byte expression).

Single-byte encoding is efficient for English and other European languages. When Indian languages came into picture then the single byte expression was not capable to represent it effectively. Because ISCII could represent maximum 256 characters at a time which means it could represent Roman alpha-numerals, some special characters and maximum one or two languages. Then ISCII fails to represent all multilingual characters.

A consortium emerged consisting of software companies mainly, IBM, Xerox, Microsoft, Apple, etc. called as Unicode consortium. The name Unicode was suggested understanding the universal, unique and uniform features of the code. In 1991 the first version of Unicode came into existence. Unicode is a single code which handles all the scripts in the world. It provides space for dead scripts also. It acts as repertoire for all written languages of the world. All Indian language has an average of 50 to 60 alphabets including consonants and vowels and these alphabets can generate thousands of combinations. ISCII and Unicode 5.0 treat a *chillu* as a glyph variant of a normal consonant letter. Later in Unicode 5 they are treated as independent characters.

In initial stages of application development in India, text rendering was a major issue as syllabic writing system in Indic scripts adds to the complexity. Earlier applications gave more emphasis on text entry and display rather than computation. Therefore the standardizations developed are mainly concerned with aspects of writing system rather than linguistic requirements.

In Indic scripts complexities of writing systems includes a large number of written shapes, but linguist content can be specified using a small set of codes for vowels and consonants. The Designers of ISCII and Unicode compromised with smaller set of code but they also incorporated codes conveying rendering information. These codes follow the Devanagari writing system which is not adequate for writing systems of the south. The sorting order of the writing system is also not maintained according to the specific language script. Developers have to take additional care in handling the order in their applications.

In order to process text for Indic Scripts, certain procedures or algorithms have to be followed. This procedure is described below using Malayalam script as a case study. Malayalam text extract from corpus encoded in Unicode may have following expressions.

5.1 Unicode Malayalam Block

In Unicode encoding, Malayalam Block is of the range from code point 3328 to 3455 (Hexadecimal: D00 to D7F). It consists of following types of characters.

Vowels (V): അ, ആ, ഇ, ഈ, ഉ, ഊ, ഋ, എ, ഏ, ഐ, ഒ, ഓ, ഔ
Consonants (C) :ക, ഖ, ഗ, ഘ, ങ, ച, ഛ, ജ, ഝ, ഞ, ട, ഠ, ഡ, ഢ, ണ, ത, ഥ, ദ, ധ, ന, പ, ഫ, ബ, ഭ, മ, യ, ര, ല, വ, ശ, ഷ, സ, ഹ, ള, ഴ, റ
Chillaksharam: ൻ, ൽ, ർ, ൾ, ൺ
Yogavaha (Y): ◌ം, ◌ഃ
Matras (M): ◌ാ, ◌ി, ◌ീ, ◌ു, ◌ൂ, ◌ൃ, ◌െ, ◌േ, ◌ൈ, ◌ൊ, ◌ോ, ◌ൗ
Chandrakala (CK): ്
Numerals (NUM): ൦, ൧, ൨, ൩, ൪, ൫, ൬, ൭, ൮, ൯

Characters found in Malayalam texts outside Unicode Malayalam block are Punctuations (. , ; :), foreign characters (A, a, IV) and symbols (@, $, #) etc.

6. ALGORITHM FOR PROCESSING MULTI-BYTE EXPRESSIONS

The initial step is to read the first character of the given Malayalam text.
Case 1: If the character is a V, check the next character
    Case 1.1: If the character is Y then concatenate previous V+Y as a letter.
    Case 1.2: If the character is not Y, then consider V as a letter.

Case 2: If the character is a C, check the next character
    Case 2.1: If the character is any start character then concatenate all the states from start to previous state as a letter.
    Case 2.2: If the character is a CK then check for next character
        Case 2.2.1: If the character is C then Go to Case 2
        Case 2.2.2: If the character is other than C then concatenate all the states from start to previous state as a letter.
    Case 2.3: If the character is M, then check the next character
        Case 2.3.1: If the character is any start character then concatenate all the states from start to previous state as a letter.
        Case 2.3.2: If the character is Y then concatenate all the states from start to current state as a letter.
    Case 2.4: If the character is Y then concatenate all the states from start to current state as a letter.

Case 3: If the character is *Chillu*, then check the next character
    Case 3.1: If the character is C then go to case 2
    Case 3.2: If the character is not C, then consider *Chillu* as a letter.

Case 4: If the character is other, Ignore and goes to the next character.

Abbreviations in the diagram are V = Vowel; C = Consonant; Chillu = Chillaksharm; Y= Yogavaha; CK= Chandrakala; M = Matra; Other = punctuation, foreign character, number

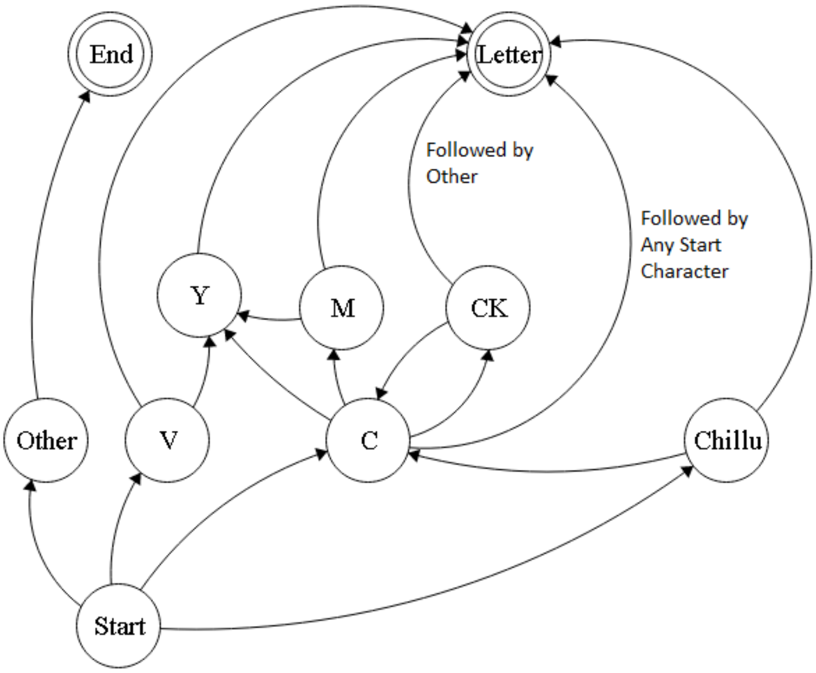

Fig -1: Multi-byte expression of Malayalam letters

7. CONCLUSION

Letters are the basic unit of word hence letter based language models are more effective. This procedure for text processing is imperative for Indic scripts for recognizing the basic unit of the language script. Identifying the boundary of a letter is robust for some of the text processing NLP applications like N-Grams, Spell Checkers, Morphological Analyzers, Sandhi Splitters etc.

REFERENCES

[1] Richard Ishida, An Introduction to Indic Script, World Wide Web Consortium, 2002.
[2] R.K. Joshi, Keyur Shroff, S. P. Mudur, A Phonemic Code Based Scheme for Effective Processing of Indian Languages, 23rd Internationalization and Unicode Conference, 2003.
[3] Yule George, The study of Language, Cambridge University Press, 1996.
[4] https://archive.org/details/gov.in.is.13194.1991/page/n9
[5] https://unicode.org/charts/PDF/U0D00.pdf
[6] http://www.alanwood.net/unicode/