Volume : 1 | Issue : 12 | September 2012 ISSN - 2249-555X

Research Paper Computer Science

Font Convertors For Indian Languages-A Survey

*Anjana Kumari **Vishal Goyal

*M.Tech Scholar, Department of Computer science, Punjabi University Patiala, Punjab

** Assistant Professor, India Department of Computer science, Punjabi University Patiala

ABSTRACT

India has diversity in languages. In different regions of India, different types of languages are being spoken. The amount of data stored in Indian language was in ASCII format and the system for identification was manual. And so the conversion of font data into another font was a difficult task. Effort has been made to make the identification of font automatic and conversion of font data into another font. N-gram approach, Empirical approach and TF-IDF approach are the techniques are used for conversion process. The algorithm has been made for identification and conversion process

Keywords :N-gram and TF-IDF.

1. Introduction

The objective of font data identifier and converter is to identify the font of the document and convert into Unicode because Indian languages have a lot of data available in ASCII format which is sufficient for displaying the correct data of the document on every system as it needs the fonts to be available on the system. The scripts in Indian languages have originated from the ancient Brahmi script. The basic units of the writing system are referred to as Aksharas. The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation of a speech sound in an Indian language; (2) Aksharas are syllabic in nature; (3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V. The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants. In

defining the shape of an Akshara, one of the consonant symbols acts as pivotal symbol (referred to as semi-full form). Depending on the context, an Akshara can have a complex shape with there consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol (referred to as half-form). Thus to render an Akshara, a set of semi-full or half-forms have to be rendered, which in turn are rendered using a set of basic shapes referred to as glyphs. Often a semi-full form or half-form is rendered using two or more glyphs, thus there is no one-to-one correspondence between glyphs of a font and semi full or half-forms.

2. Literature Survey

1.1 Dangi Soft (Font Converter)

This converter is developed by Er. Jagdeep Dangi. It consists of two parts. The first part is called Prakhar Devanagari Font Parivartak. It is also called ASCII/ISCII to Unicode Converter. The second part is called UniDev.

Prakhar Devanagari Font Parivartak (ASCII/ISCII to Unicode Converter):-This software can convert various ASCII/ISCII Devanagari texts into Unicode text with 100% accuracy. The advantage of doing that, according to the creator of this software is that Unicode fonts based Devanagari text can be searchable on the Internet. Many govt. sites are using ASCII/ISCII Devanagari fonts, which are not searchable. The software can therefore do wonders for them. This is the first and only software for the purpose of conversion of devanagari (Hindi/Marathi/Sanskrit) text in various ASCII/ISCII (8 bit) fonts (about more than 258 both true type and type-1 fonts like Kruti -dev, Chanakya, Shusha, Shiva, DV-TTYogesh, 4CGandhi, Sanskrit 99, Marathi-Kanak etc) into Unicode (16 bit) text immediately and easily with 100% accuracy.

UniDev: - UniDev can convert Unicode (Mangal) based font to various ASCII/ISCI fonts like Kruti dev, Chanakya etc for (Hindi / Sanskrit / Marathi) Devanagari Script. At present This is the first error free tool in the market for converting Unicode based text to various ASCII/ISCII (8 bit) fonts (at present about 9 both true type and type-1 fonts like Kruti dev, Chanakya, Shiva etc) with 100% accuracy. This tool is very important for those users who are working in some useful software's like Corel draw, Photoshop, PageMaker, Quark Express etc. but these are not capable of supporting the Unicode. This is also useful for those who are working in printing job because ASCII/ISCII fonts have many designs and styles than Unicode Mangal font.

1.2 E-Pandit IME

Mr. Shirish Benjwal Sharma developed this program and is available free of charge. E – Pundit is based on Hindi and Devanagari script and for other languages based Inscript keyboard layout as an IME (Input Method Editor). It support both Unicode and non – Unicode fonts. This tool is used for typing in Hindi fonts like Chanakya, Kruti dev and Walkman-Chanakya etc using Inscript keyboard. This tool especially used by Coral Draw, PageMaker, etc. Non - Unicode programs which do not type Unicode Hindi.

1.3 Google IME

Google's service for Indic languages was previously available as an online text editor, named Google Indic Transliteration. Other language transliteration capabilities were added (beyond just Indic languages) and it was renamed simply Google transliteration. Later on, because of its steady rise in popularity, it was released as Google Transliteration IME for offline use in December 2009. It works on a dictionary-based phonetic transliteration approach, which means that whatever you type in Latin characters, it matches the characters with its dictionary and transliterates them. It also gives suggestions for matching words.

Google transliteration (formerly Google Indic Transliteration) is a transliteration typing service for Indian and others lan-

guages. This tool first appeared in Blogger, Google's popular blogging service. Later on it came into existence as a separate online tool. Keeping in view its popularity it was embedded in Gmail and Orkut. In December 2009 Google released its offline version named Google IME. This tool from Google is based on dictionary based phonetic transliteration approach. In contrast to older Indic typing tools (which type by transliterating under a particular scheme), it transliterates by matching the Latin words with an inbuilt dictionary. So, users do not need to remember the transliteration scheme and because of this, the service is very easy and suitable for first beginners.

1.4 Microsoft Indic Language Input Tool

Microsoft Indic Language Input Tool is a typing tool (Input Method Editor) for Hindi and other Indic languages. It is a virtual keyboard which allows to type Indic language text directly in any application without hassle of copying and pasting. It is available for both, online and offline use. It was released in December 2009.

It works on Dictionary based Phonetic Transliteration approach. It means whatever you type in Latin characters, it matches that with its dictionary and transliterates it, it also gives suggestions for matching words.

1.5 Devanagari Converters

Anand Arokia Raj, Kishore Prahallad [4] discusses the issues related to font encoding identification and font-data conversion in Indian languages.

They uses TF-IDF(Term frequency and Inverse document frequency) weights approach for identification of font encoding by giving weightage to unigrams(current glyph), bigrams (current and next glyph) and trigrams (previous, current and next glyph). These are the terms which are most commonly occurred in the data file and weightage is given accordingly. With the help of this scheme identification for three grams has been done. For conversion of font data , IT3 transliteration scheme has been used that helps in making character map table for each font. Several assimilation rules have been developed to overcome the conversion problems. They tried this for several languages like Hindi, Punjabi, Tamil, Guajarati, Kannada, Malayalam, Oriya, Bengali etc. The results for conversion process are 100% accurate.

3. Conclusion

In future based upon the above literature survey, efforts have been made to do the identification process automatic and convert the font data into Unicode encoding so that the Unicode data can be transferred without any issues related to the system for Hindi language.

4. Results

Based on the experiments performed we concluded that for identification process combination of unigrams, bigrams and trigrams are used that give maximum summation of all the fonts. In font conversion process, with the help of character map table and font assimilation rules, the solution is comes out to be 100%. In future, we can add data corresponding to other languages and use the same system for them.

REFERENCES

[1] http://sites.google.com/site/technicalhi ndi/home/converters | [2] http://www.tamasoft.co.jp/en/general-info/unicode.html | [3]http://www.google.com/ime/transliteration/ | [4] Anand Arokia Raj, Kishore Prahallad (2007)," Identification and Conversion of Font-Data in Indian languages" in International Conference on Universal Digital Library (ICUDL2007) November 2007, Pittsberg, USA |