Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis

M L Dhore[1] *S K Dixit*[2] *R M Dhore*[3]
(1)LINGUISTIC RESEARCH GROUP, VIT, Pune, Maharashtra, India
(2) LINGUISTIC RESEARCH GROUP, WIT, Solapur, Maharashtra, India
(3) LINGUISTIC RESEARCH GROUP, PVG COET, Pune, Maharashtra, India

manikrao.dhore@vit.edu, dixitsk@wit.edu, ruchidhore@pvg.edu

ABSTRACT

During last two decades, most of the named entity (NE) machine transliteration work in India has been carried out by using English as a source language and Indian languages as the target languages using grapheme model with statistical probability approaches and classification tools. It is evident that less amount of work has been carried out for Indian languages to English machine transliteration.

This paper focuses on the specific problem of machine transliteration of Hindi to English and Marathi to English which are previously less studied language pairs using a phonetic based direct approach without training any bilingual database. Our study shows that in depth knowledge of word formation in Devanagari script based languages can provide better results as compared to statistical approaches. Proposed phonetic based model transliterates Indian-origin named entities into English using full consonant approach and uses hybrid (rule based and metric based) stress analysis approach for schwa deletion.

KEYWORDS: Machine Transliteration, Named Entity, Full Consonant, Metrical Approach

Proceedings of COLING 2012: Demonstration Papers, pages 111–118,
COLING 2012, Mumbai, December 2012.

1. Introduction

Hindi is the official national language of the India and spoken by around 500 million Indians. Hindi is the world's fourth most commonly used language after Chinese, English and Spanish. Marathi is one of the widely spoken languages in India especially in the state of Maharashtra. Hindi and Marathi languages are derived from the Sanskrit and use the "Devanagari" script for writing. It is challenging to transliterate out of vocabulary words like names and technical terms occurring in the user input across languages with different alphabets and sound inventories. Transliteration is the conversion of a word from one language to another without losing its phonological characteristics (Padariya 2008). Hindi and Marathi to English NE transliteration is quite difficult due to many factors such as difference in writing script, difference in number of alphabets, capitalization of leading characters, phonetic characteristics, character length, number of valid transliterations and availability of the parallel corpus (Saha 2008).

2. Architecture of Transliteration System

The architecture of Hindi/Marathi to English transliteration system and its functional modules are shown in figure 1.

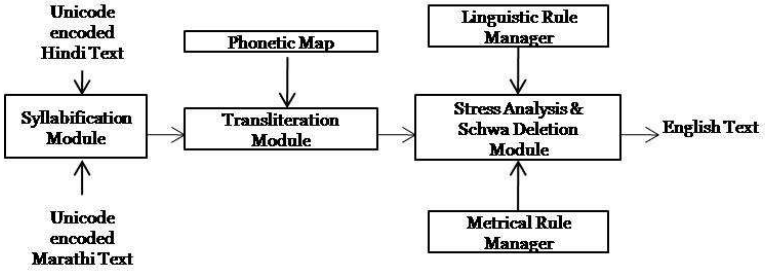

Figure 1. Architecture of Transliteration System

3. Phonetic Map for Hindi/Marathi to English

The possibility of different scripts between source and target languages is the problem that transliteration systems need to tackle. Hindi and Marathi use the Devanagari script whereas English uses the Roman script. Devanagari script used for Hindi and Marathi have 12 pure vowels, two additional loan vowels taken from the Sanskrit and one loan vowel from English. English has only five pure vowels but, the vowel sound is also associated with the consonants w and y (Koul 2008). There are 34 pure consonants, 5 traditional conjuncts, 7 loan consonants and 2 traditional signs in Devanagari script and each consonant have 14 variations through integration of 14 vowels while in Roman script there are only 21 consonants (Walambe 1990, Mudur 1999). Table 1 shows phonetic map for Devanagari to Roman transliteration along with their phonological mapping using full consonant approach. It is fully based on the National Library of

Kolkata and ITRANS of IIT Madras, India (Unicode 2007). The consonant /ळ/ is used only in Marathi language.

Vowel	Matra	Vowel	Matra	Pure consonants				
अ→a		ऋ→RU	ृ	क→ka	ख→kha	ग→ga	घ→gha	ङ→nga
आ→ A	ा	ए→E	े	च→cha	छ→chha	ज→ja	झ→jha	ञ→ya
इ→i	ि	ऐ→ai	ै	ट→Ta	ठ→Tha	ड→Da	ढ→dha	ण→Na
ई→ee	ी	ओ→oo	ो	त→ta	थ→tha	द→da	ध→Dha	न→na
उ→u	ु	औ→au	ौ	प→pa	फ→pha	ब→ba	भ→bha	म→ma
ऊ→U	ू	अं→am	ं/ ँ	य→ya	र→ra	ल→la	व→va	श→sha
ऋ→Ru	ृ	अः→aH	ः	ष→Sha	स→sa	ळ→La	ह→ha	
Conjucts,Symbols, Loan Letters →		क्ष→ksha	ज्ञ→dnya	श्र→shra	द्य→dya	श्री	ॐ→om	क़→kxa
		ढ़→Dhxa	ख़→khxa	ग़→gxa	ज़→jxa	फ़→phxa	ड़→Dxa	

Table 1: Phonetic Map for Transliteration

4. Syllabification

Unicode encoded Hindi or Marathi named entity text input is given to the syllabification module which separates the input named entity into syllabic units. A syllabic unit is equivalent to one *'akshara' in* Devanagari where *'akshara'* is the minimal articulatory unit of speech in Hindi and Marathi. Few examples are

कैलाशनाथ → कै | ला | श | ना | थ and विजयराघवगढ़ → वि | ज | य | रा | घ | व | ग | ढ़

5. Transliteration Module

This module converts each syllabic unit in Devanagari into English by using phonetic map. Phonetic map is implemented by using the translation memory and mapping is done by writing the manual rules. It is to note that the first vowel /**अ**/ in Hindi/Marathi is mapped to English letter 'a' (short vowel) while the second vowel /**आ**/ is mapped to 'ā' (long vowel as per IPA) in English. The alphabet 'a' in English is a short vowel equivalent to /**अ**/ which is also a short vowel in Devanagari while /**आ**/ in Devanagari is a long vowel and mapped capital 'ā' or 'A' in our phonetic scheme. Unicode and ISCII character encoding standards for Indic scripts are based on full form of consonants (Singh 2006, BIS 1991). Few examples are

Input	Transliteration	Observations
अनंता	anantA	Correct Transliteration
माणिक	mAnika	Last 'a' to be removed
कैलाशनाथ	kailashanatha	Schwa followed by 'sh' & last 'a' to be removed
विजयराघवगढ़	vijayarAghavagaDha	Schwa followed by 'y', 'v' & last 'a' to be removed

Table 2. Output of Transliteration Module

Schwa - The schwa is the vowel sound in many lightly pronounced unaccented syllables in words of more than one syllable. It is represented by /ə/ symbol (Naim 2009). The schwas remained in the transliterated output need to be deleted (Pandey 1990). The

schwa identification and deletion is done by applying the manual rules and stress analysis. Instead of using either approach, it is better to combine both the approaches due to phonotactic constraints of Hindi and Marathi languages.

6. Rules for Schwa Retention and Deletion

From empirical observations the following six rules are applicable to all NEs (Choudhury 2004).

Rule 1: The schwa which occurs in isolation at the start of named entity never gets deleted. Example: amar (अमर, əmər) → [**ə**] : [mər]

Rule 2: The schwa preceding a full vowel never gets deleted to maintain lexical distinctions. Example: pawai (पवई, pəw**ə**ī) → [pə] : [w**ə**] : [i]

Rule 3: The schwa which occurs in the first syllable never gets deleted if it is a consonant phoneme without any matra.

Rule 4: The schwa which occurs at the end of word always gets deleted. Example: gopal (गोपाल, gopAl**ə**) → [go]:[pa] :[l**ə**]

Rule 5: If the word ends in a consonant cluster or the rafar diacritic, then the schwa is retained. Example: atharva (अथर्व, əthərv**ə**) → [ə]:[thərv**ə**]

Rule 6: The schwa of the syllable immediately followed by a conjugate syllable should always be retained. Example: brahaspati (ब्रह्स्पती, brah**ə**spəti → [bra]:[h**ə**]:[spa]:[ti]

7. Schwa Deletion Using Stress Analysis

Generally, the location of word stress in Hindi and Marathi is predictable on the basis of syllable stress. Stress is related both to the vowel length and the occurrence of postvocalic consonant. According to Hindi and Marathi phonology literature there are three classes of vowels used for stress analysis but it is possible to obtain the stress analysis using only two classes as shown below.

Class- I: Short vowels /a or ə/, /u/ and /i/ denoted by L (Light)

Class-II: Long vowels /ā/,/e/,/ī/,/o/,/ū/,/ai/, /au/ denoted by H (Heavy)

Algorithm 1: Schwa Deletion Using Stress Analysis

If last position has schwa **then**

If last syllabic unit does not contains consonant cluster or the rafar diacritic **then**

delete word-final schwa and resyllabify named entity

endif

endif

foreach syllabic unit in English **do** . assign metrical class (L or H) **end foreach**

foreach syllabic unit and next syllabic unit **do** create metrical feet if any **end foreach**

foreach foot **do** find unstressed foot

if the contexts of rule 1 to 4 from manual rules does not occur **then**

delete schwa(s) in unstressed foot and resyllabify foot

endif

end foreach

The process of combining syllables is carried out from right to left as schwa always appears to the right position (Naim 2009). With syllable stress and metrical feet, it is possible to find which syllable receives primary lexical stress. The stress information is used to decide whether the schwa is an unstressed syllable and to delete all such unstressed schwas. Few examples are shown in table 3.

Named Entity	Metrical Assignments	Schwa Detection	Transliteration in English	Outcome
हेमवतीनंदन	HLLHHLL	hemə̲vatinandan	hemvatinandan	Correct
त्रिलोकनारायण	HHLHHLL	triloka̲nArAyan	triloknArAyan	Correct
जगदंबाप्रसाद	LLHHHHL	jaga̲dambAprasad	jagdambaprasAd	Correct

Table 3. Schwa Deletion from Unstressed Syllables

8. Demonstration

We have developed real time application for a co-operative banking which allows user to enter data in Marathi or Hindi language and transliterate it into English. Figure 2 shows the snapshot from our experimentation for full name transliteration in English.

Figure 2. Transliteration Using Phonology and Stress Analysis

As there are more number of alphabets in Devanagari as compared to English, one alphabet in Devanagari can be mapped to multiple alphabets in English depending on

the origin of word. Transliteration module maps the highest priority alphabet to generate the first candidate named entity. Table 4 is used to generate multiple transliteration candidates (Chinnakotla 2010).

Hindi	क	ख	ग	ड	व	ई	श	क्ष	औ
English	k,c,q	k,kh	g, gh	d,dh	w,o,b,bh	i,e,ee,ey	sh,s	ksh,x	au,ou

Table 4. Multiple Candidate Generation

Figure 3 shows the snapshot from our experimentation for personal profile of the bank account holder. It allows user to enter information in Marathi language and shows the corresponding transliterated record in English.

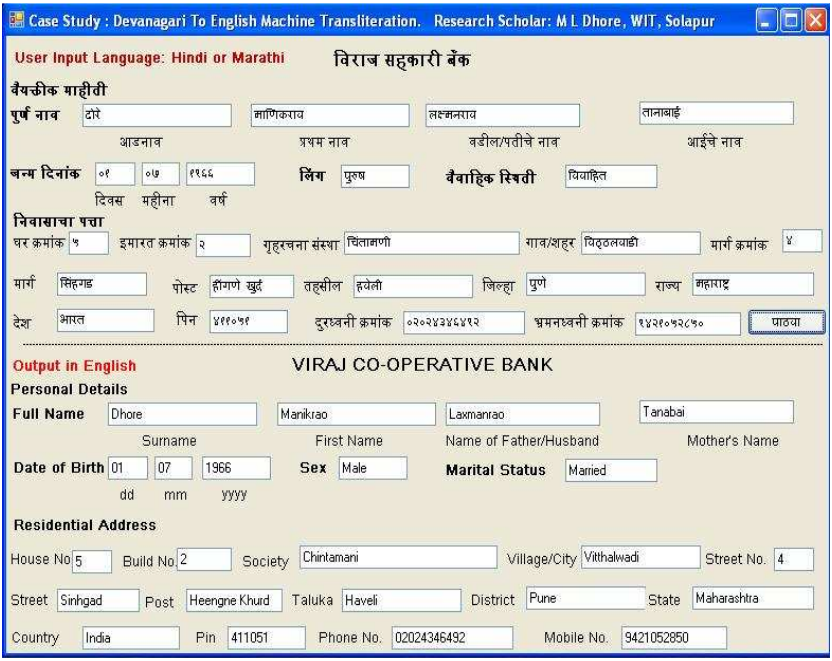

Figure 3. Account Holders Personal Information Form

Conclusion

The use of machine transliteration in cross language applications is imminent and promising. Hindi and Marathi can be supported in all e-governance projects of the Government of India. As there is no need of bi-lingual and multilingual databases to be trained, this application can easily be ported on mobile devices. This application can help masses of India who do not know English.

References

BIS (1991), Indian standard code for information interchange (ISCII), *Bureau Of Indian Standards*, New Delhi.

Chinnakotla Manoj K., Damani Om P., and Satoskar Avijit (2010), Transliteration for Resource-Scarce Languages, *ACM Trans. Asian Lang. Inform*,Article 14, pp 1-30.

Choudhury Monojit and Basu Anupam (2004), A rule based schwa deletion algorithm for Hindi, Indian Institute of Technology, Kharagpur,West Bengal, India.

Koul Omkar N. (2008), Modern Hindi Grammar, Dunwoody Press

Mudur S. P., Nayak N., Shanbhag S., and Joshi R. K. (1999), An architecture for the shaping of indic texts, *Computers and Graphics*, vol. 23, pp. 7–24.

Naim R Tyson and Ila Nagar (2009), Prosodic rules for schwa-deletion in Hindi Text-to-Speech synthesis, *International Journal of Speech Technology*, pp. 15–25

Padariya Nilesh, Chinnakotla Manoj, Nagesh Ajay, Damani Om P.(2008), Evaluation of Hindi to English, Marathi to English and English to Hindi, *IIT Mumbai CLIR at FIRE*.

Pandey Pramod Kumar (1990), Hindi schwa deletion, Department of Linguistics. South Gujarat University, Surat , lndia, *Lingua* 82, pp. 277-31

Saha Sujan Kumar, Ghosh P. S, Sarkar Sudeshna and Mitra Pabitra (2008), Named entity recognition in Hindi using maximum entropy and transliteration.

Singh Anil Kumar (2006), A computational phonetic model for Indian language scripts, Language Technologies Research Centre International Institute of Information Technology Hyderabad, India.

Unicode Standard 5.0 (2007) – Electronic edition, 1991–2007 Unicode, Inc. *Unicode Consortiums*, http://www.unicode.org.

Walambe M. R. (1990), Marathi Shuddalekhan, *Nitin Prakashan*, Pune

Walambe M. R. (1990), Marathi Vyakran, *Nitin Prakashan*, Pune