DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4801

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ENGLISH TO HINDI TRANSLITERATION SYSTEM USING COMBINATION-BASED APPROACH

Baljeet Kaur Dhindsa
Assistant Professor
Department of Computer Science and Applications
Guru Gobind Singh College for Women, Sector 26
Chandigarh, India

Dharam Veer Sharma
Associate Professor
Department of Computer Science
Punjabi University
Patiala, India

Abstract: Transliteration plays a very significant role in machine translation, which has many applications such as cross-lingual information retrieval, communication, question-answering etc. The main objective of this research paper is to provide a method for transliteration of named entities from English to Hindi language. The proposed method consists of two modules, both of which apply phoneme-based approach to transliterate named entities. For transliteration, Module-I utilizes CMU Pronouncing dictionary, which is a collection of 133270 words along with their pronunciation. If the word to be transliterated is not found in CMU Pronouncing dictionary, Module-II is used. Module-II is based on 5-gram model, in which a maximum of five letters (two left, two right and one target letter) are used to generate transliterated target letter. The system has been tested on a database of 2408 North-Indian names. Google Input tool for Windows has been used for comparative study of the proposed transliteration system. The word accuracy of the transliteration system has been found to be 70.22% against 58.73% of Google Input tool.

Keywords: Transliteration; English-to-Hindi Transliteration; Combination-based Transliteration.

1. INTRODUCTION

Transliteration is representation of the word given in source language (SL) into a word in target language (TL), in which the alphabet of the TL only needs to be used without changing the phonemes of the word. Transliteration is a combination of two words: Trans + Littera, where Trans means change and Littera is a Latin word which means letter. Dictionary meaning of transliteration is change letters/words into corresponding characters of another alphabet/language[1]. It can also be defined as pronunciation preserving translation. Transliteration of a word may be required under the following circumstances [1]:

- For named entities like names of people, places, organizations, products etc.
- For language specific words, which are not part of vocabulary of any other language, such as Ikebana: a Japanese technique of flower arrangement, Kuchipudi: an Indian dance form etc.
- For technical terms like computer, printer, television etc which may not have a corresponding word in the TL or it may have a word which is not commonly recognizable by the users.

Thus, transliteration plays a very important role in translation. Machine Transliteration is automatic transliteration done by machine without human intervention.

Machine transliteration is performed on a word in language. A word consists of collection of valid letters also known as graphemes. For example, English language supports 26 graphemes (21 consonants and 5 vowels), Hindi language supports 46 graphemes (35 consonants and 11 vowels) [2]. Graphemes when combined with other graphemes to create a word produce sounds called phonemes. Grapheme is smallest unit of written language and phoneme is smallest unit of spoken language [1]. Natural languages can be divided into two types based on the type of script they use: (i) Languages following phonetic script (ii) Languages following non-phonetic script. In phonetic script, every grapheme produces the same sound, no matter at which position in the word it is used. The grapheme may appear in the beginning, in the middle or in the end of the word. Languages following phonetic script are Hindi and Sanskrit. In non-phonetic script, graphemes may result in producing different sounds not only with change in its position in the word, but also with change in its neighboring letters. English follows non-phonetic script [2]. There are many approaches used by researchers to perform machine transliteration, which may be divided into four types [3]:

(i) Grapheme-based/ Direct Method
(ii) Phoneme-based/ Pivot Method
(iii) Hybrid and Correspondence based Method
(iv) Combination-based Method

Grapheme-based approach is preferred for languages using phonetic script and Phoneme-based is used for languages using non-phonetic scripts. Hybrid approach combines the best features of both the grapheme-based and phoneme-based approaches as most of the languages cannot be categorized as entirely phonetic or non-phonetic. Combination-based approach is preferred when transliteration is required among multiple languages [4]. It can also be used to improve transliteration quality of existing systems or for designing transliteration systems for languages for which parallel corpora of names does not exist [3].

[1] Dictionary.com. http://www.dictionary.com/browse/transliteration.

2. RELATED LITERATURE

Research work of many has contributed directly and indirectly in making the transliteration problem comprehensible. A few researchers whose work has directly influenced the current research are discussed below.

Mathur and Saxena [5] have used hybrid approach in designing English to Hindi transliteration system. Here, rule-based approach is used to extract phonemes from English source word. Statistical approach is then applied on the extracted phonemes for conversion into Hindi language phonemes to produce Hindi word.

Das ,Ekbal,Mandal and Bandyopadhyay [6] have combined three transliteration models to perform English to Hindi transliteration. Initially the English word to be transliterated is mapped against Direct example base consisting of bilingual training examples. In case no match is found, direct orthographic mapping is applied, where for every transliteration unit in English word; a corresponding transliteration unit in Hindi is generated.

Haque, Dandapat, Srivastava, Naskar and Way [7] have combined Phrase-based statistical machine transliteration and source context modeling to design English to Hindi transliteration system. The left and right context character/ transliteration-unit are used to decide the transliteration value of the character/ transliteration-unit.

Hindi and Punjabi are closely related languages. Thus research literature for transliteration to Punjabi language is also found to be helpful. *Bhalla, Joshi and Mathur* [8] give brief description of transliteration system for proper names and places from English to Punjabi using phoneme-based approach. In this system, named entities are divided into syllables using rules. It uses rule-based approach to transliterate syllables to Punjabi. MOSES, statistical MT toolkit is used to calculate probability values of words not categorized as named entities.

3. AREAS OF CONCERN IN ENGLISH TO HINDI TRANSLITERATION

Hindi is a phonetic language, whereas English is non-phonetic language. Main issues that require attention while transliterating text from English to Hindi are:

- Hindi, which is based on Sanskrit alphabet system, has strong phonetic features and is based on linguistics. Furthermore, its alphabet sequence depends on place of articulation of vowels and consonants i.e. velar, palatial, retroflex, dental, labial, nasal etc. Even the consonants in every category are placed according to the pronunciation features [2]. On the other hand, in English alphabet system, which is based on Roman alphabet system, the alphabet sequence does not depend on place of articulation of vowels and consonants. Moreover, there is no sequence, based on linguistic and phonetics, in Roman alphabets.
- Every letter in English language has a basic pronunciation value, which can change when used in a word. In many cases, pronunciation depends on previous and/or next letter(s) used in a word and sometimes it is decided simply by its traditional usage. For example, 'u' produces sound of **'उ'{oo}** by default, but when used in the word 'but', it gives sound of **'अ'{a}** and in the word 'put', it produces sound of **'उ'{oo}**, although difference is of 'b' as first character or 'p' as first character. In Hindi language, pronunciation of vowels and consonants is almost fixed and it does not change with its position in the word nor does it gets affected by its previous or next letters [2].
- In English language a word may contain a silent letter whose phoneme is not pronounced [9]. For example, in the word Psychology (P is silent), Prix (x is silent), Write (W is silent), Knife (K is silent) etc. In Hindi language, there is fixed sound of every letter and there are no silent letters.
- There are a few words in English language which comprise of collection of letters resulting in pronunciation which is different from its actual phonology. For example, Yacht (ch being pronounced as "औ" (Au)), Jnanpith (Jn being pronounced as "ज्ञ"(Gya)).
- While comparing alphabet of two languages, there may be some letters in one which are not supported by the other language. There are 26 letters including consonants and vowels in English language. Hindi language has 35 consonants and 11 vowels. To accommodate Urdu and Arabic language eight more letters have been added to Hindi, like क़, ख़, ज़ फ़ etc [2]. Some letters like ख (KH), छ (CHH), झ (JHH) etc. are part of Hindi alphabet but are not supported by English language.

In view of the above mentioned points, it is difficult to design English to Hindi transliteration system using either phoneme-based technique alone or grapheme-based only.

4. PROPOSED SYSTEM

The transliteration system which has been designed utilizes Combination-based approach. It consists of two modules, both of which are phoneme-based.

(i) Module-I: Pronunciation Dictionary Method.
(ii) Module-II: Rule-based system (5-gram model).

The system can accept multi-word named entities to be transliterated, in which case every word is transliterated individually. Every word that needs to be transliterated first goes to Module-I, which searches the word in Carnegie Mellon University (CMU) pronouncing dictionary [10]. If the search is unsuccessful, it goes to Module-II, which transliterates the word using rule-based method. Figure 1 depicts the working of transliteration system. Section A and section B describes the modules.

A. Module-I: Pronunciation Dictionary Method

In Module-I, the word to be transliterated is searched in the CMU pronouncing dictionary. This dictionary contains 133270 words along with their pronunciation. For the proposed system, the CMU pronouncing dictionary, which is available in text form, has been converted into local database using MS-Access 2007. For designing the transliteration system, words in this dictionary are divided into 13 tables, each having a few thousand entries, arranged alphabetically. This helps in

reducing search time for a word among 133270 words. Depending on the first letter of the word to be transliterated, the search is limited to only a single table containing dictionary entries for that letter. If the word is found in the table, its corresponding field representing its pronunciation is accessed. The dictionary originally contains 69 unique pronunciation symbols (PS), out of which 45 symbols correspond to different sounds produced with vowels or vowel-consonant combination. For the proposed system, careful analysis of the PS in the pronunciation fields of various words has been done and corresponding Hindi consonants and/ or vowels has been assigned for each PS. Twelve PS out of total 69 have been found to be diphthong and result in combination of letters in Hindi language. There are some letters in Hindi, which are not supported by English. Thus eleven additional PSs are added to this dictionary to accommodate these Hindi language letters. This has also resulted in changes in some of the pronunciation fields of the dictionary. For example, pronunciation field value for the word "KHAN" has been changed from "K AA1 N" to "KH AA1 N", because its actual pronunciation is "खान"{Khaan}, but its earlier pronunciation has been resulting in "कान"{Kaan}. All the PSs corresponding to the word to be transliterated are converted to the corresponding alphabet(s) in the TL Hindi. Once the graphemes in Hindi are found, they are combined to form a word.

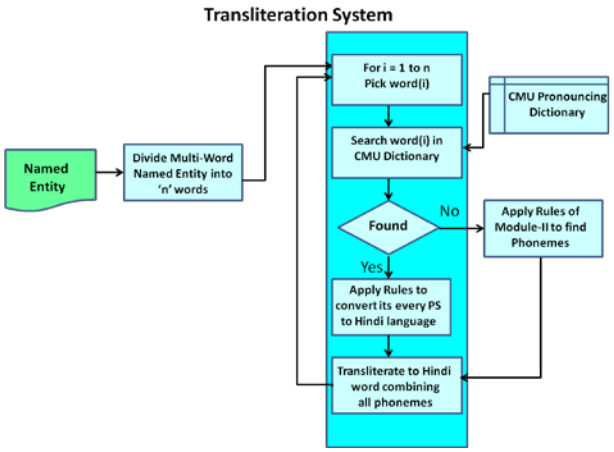

Figure 1. Working Model of Transliteration System

Hindi language also supports half consonants. Every alphabet in Hindi language has a default and hidden 'अ' {a} sound in it, e.g. pronunciation of 'क' in Hindi is 'कअ', 'स' is actually 'सअ', where 'अ' is not explicitly written [2]. Sometimes two successive alphabets are required to be pronounced without 'अ' sound, e.g. "Chennai" {चेन्नई}, "Chandigarh" {चन्डीगढ}, "Kavya" {काव्या}. In case, where an alphabet has neither inherent 'अ' sound nor any other vowel sound, it is written as half consonant like "न्न" in Chennai, first 'न' is half, same is the case in "Chandigarh" and 'व' is half in "Kavya". Table I illustrates the working of Module-I with the help of few examples.

Table I. Examples illustrating working of Module-I

S.No.	*English Word*	*CMU's Pronunciation Field*	*Corresponding Hindi Consonants/ Vowels*	*Transliterated Word in Hindi*
1.	Khan	KH* AA1 N	ख + आ + न= ख ा न	खान
2.	Aagya	AA1 G Y AA1	आ + ग + य + आ = आ ग ् य ा	आग्या
3.	Chain	CH EY1 N	च + ए + न = च े न	चेन
4.	Daniel	D AE1 N Y AH0 L	ड + ऐ + न + य + अ + ल= ड ै न ् य ल	डैन्यल
5.	Google	G UW1 G AH0 L	ग + ऊ + ग + अ + ल = ग ू ग ल	गूगल
6.	Echo	EH1 K OW0	ऍ + क + ओ = ऍ क ो	ऍको
7.	Eco	IY1 K OW0	ई + क + ओ = ई क ो	ईको

B. *Module-II: Rule-Based Method*

This system is used when the English word to be transliterated is not found in CMU pronouncing dictionary. It is a rule-based method, in which rules are created based on the pronunciation of Indian names, especially North-Indian names. The given system is a 5-gram model, in which maximum of five consecutive letters of the English word are considered while pronouncing a named entity. It uses two preceding letters, if exists, and two following letters, if exists, of the target letter to determine the pronunciation of the target letter. This module consists of 72 rules for handling consonants and 51 rules for handling vowels. In order to determine, the pronunciation of the letter at the i^{th} position in English word, following neighboring letters are considered:

$$E_{i-2}E_{i-1}E_iE_{i+1}E_{i+2}$$

Here, E_i is the target letter, E_{i-1} is its immediate left neighbor, E_{i-2} is its left to left neighbor, E_{i+1} is its immediate right neighbor and E_{i+2} is its right to right neighbor. Sometimes all five letters are not required to be considered to make a decision on the pronunciation to be used. It is demonstrated in the following cases:

- **Transliteration of letter 'F'**

Letter 'F' always transliterates to 'फ'. Its immediate neighbors and its position in the word do not affect its transliteration and thus there is no need to check for any of its neighbors. Table II presents examples to demonstrate the transliteration of letter 'F'.

Table II. Transliteration Examples for Letter 'F'

S.No.	*English Word*	*Transliterated Word In Hindi*	*Remarks*
1.	Falak	फलक	'F' appearing at the beginning followed by a vowel

2.	Flowerjit	फलोवरजीत	'F' appearing at the beginning followed by consonant
3.	Gurfateh	गुरफतेह	'F' appearing in between
4.	Altaf	अलताफ	'F' appearing at the end

- **Transliteration of letter 'B'**

Letter 'B' requires checking of its immediate right neighbor only to decide its pronunciation. For example, letter 'B', if followed by 'H' i.e. 'BH' gives sound of '**भ**', but if it is followed by 'B' again it is pronounced as '**ब्ब**', otherwise it is always pronounced as '**ब**', no matter at which place it appears in the named entity. Table III presents examples to demonstrate the transliteration of letter 'B'.

Table III. Transliteration Examples for Letter 'B'

S.No.	*English Word*	*Transliterated Word In Hindi*	*Remarks*
1.	Abhey	अभे	'B' followed by 'h' appearing in the middle.
2.	Bhabhi	भाभी	'B' followed by 'h' appearing at the beginning and at the end.
3.	Bachan	बचन	'B' followed by character other than 'H' or 'B'.
4.	Babban	बब्बन	'B' followed by another 'b' appearing in the middle.
5.	Ekbir	एकबीर	'B' appearing in the middle and not followed by 'H' or 'B'.
6.	Panjab	पंजाब	'B' appearing at the end.

- **Transliteration of letter 'R'**

Letter 'R' can result in consonant '**र**' as in "**राम**"(Ram), '**त्र**' which is a combination of '**त्**' and '**र**' as in "**त्रिदेव**" (Tridev) or in vowels '**ृ**' (RI) "**कृष्णा**"(Krishna), "**द्राविड**" (Dravid) where 'R' is placed as a vowel at the foot of the previous consonant. Table IV presents examples to demonstrate the transliteration of letter 'R'.

Table IV. Examples Demonstrating Transliteration of Letter 'R'

S.No.	*English Word*	*Transliterated Word In Hindi*	*Remarks*
1.	Ram	राम	'R' is displayed as consonant '**र**'
2.	Kritika	कृतिका	'R' is displayed as a vowel '**ृ**'
3.	Triveni	त्रिवेनी	'T' followed by 'R' displayed as '**त्र**'
4.	Krushna	क्रुशना	'R' is placed as a vowel at the foot of the previous consonant '**क**'
5.	Priti	प्रीती	'R' followed by 'i' results in vowel '**ी**'
6.	Preeti	प्रीती	'R' preceded by 'p', but not followed by 'i', results as vowel at the foot of the previous consonant.

In addition, Hindi language supports one more vowel corresponding to alphabet 'R' which is displayed at the top of next consonant such as in "**कर्ण**" (Karan). The proposed system does not support this vowel as formulating rules for this vowel may result in contradiction at many places. In all such cases 'R' is displayed as consonant '**र**'.

Karan: **करन**

Aadarsh: **आदरश**

Arjun: **अरजुन**

Duryodhan: **दुरयोधन**

The 5-gram model, which takes into consideration five letters ($E_{i-2}E_{i-1}E_iE_{i+1}E_{i+2}$), is unable to correctly transliterate from English to Hindi in some cases. For example, the proposed system (5-gram model) correctly transliterates Hritik, Hritvik and Hriday in accordance with traditional pronunciation of Indian names.

Hritik: **रितीक** ('H' is ignored if followed by RI+consonant other than 'D')

Hritvik:**रितवीक** ('H' is ignored if followed by RI+consonant other than 'D')

Hriday: **हृदये** ('H' is pronounced if followed by "RID")

The following rules have been designed keeping in view the above cases:

- If 'H' is first character in named entity i.e. it is preceded by no character and 'H' is followed by "RID", 'H' is pronounced as '**ह**'.
- If 'H' is followed by "RI", which is followed by any consonant other than 'D' then 'H' is ignored.

However, the word "Hridhan (**हृधन**)" is an anomaly, which transliterates to "**हृधन**" {Hridhan} instead of "**रिधान**" {Ridhaan}.

Once the Hindi language consonants/vowels are determined for every phoneme in English language named entity, they are combined to form a word in Hindi language.

5. RESULTS AND DISCUSSION

Test run of the proposed transliteration system has been conducted on 2408 North-Indian names collected randomly from online sources. For comparative study, Google Input Tool [11] available on official website of Google as on Feb 28, 2017, has been used. The proposed transliteration software produces only one transliterated word per input. However, Google Input Tool provides multiple transliterated options for every input name of which only the first word is used for comparison with the result of the proposed software. Word accuracy of the proposed transliteration software has been found to be 70.22% as against 58.73% of Google Input tool as on Mar 04, 2017.

6. CONCLUSION

English and Hindi language follow completely different scripts and thus have a lot of variations in their sound systems. Furthermore, named entities do not follow a fixed pattern as far as their pronunciation is concerned. Thus only one set of rules cannot be used to transliterate these. The

performance of the system can be further improved by expanding the rule base.

REFERENCES

[1] S. Karimi, F. Scholer, and A. Turpin, "Machine Transliteration Survey," ACM Computing Survey, vol. 43(3), pp. 1-46, 2011.

[2] S. Singh, English – Hindi Translation Grammar, New Delhi, Prabhat Prakashan, 2010, pp. 69-81.

[3] A. Kumaran, M. M. Khapra and P. Bhattacharyya, "Compositional Machine Transliteration," ACM Journal on Transactions on Asian Language Information Processing (TALIP), vol. 9, no 4, pp. 1-29, 2010.

[4] G. Nicolai, B. Hauer, M. Salameh, A. S. Arnaud, Y. Xu, L. Yao and G. Kondrak, "Multiple System Combination for Transliteration," in Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP Beijing, China, July 26-31, pp. 72–77, 2015.

[5] S. Mathur and V.P. Saxena, "Hybrid Approach to English-Hindi Name Entity Transliteration," Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on March 1-2, 2014, pp.1-5, 2014.

[6] A. Das, A. Ekbal, T. Mandal, and S. Bandyopadhyay, "English to Hindi Machine Transliteration System at NEWS 2009," in Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, 2009, pp. 80–83.

[7] R. Haque, S. Dandapat, A. K. Srivastava,, S. K. Naskar, and A. Way, "English—Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009," in Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, 2009, pp. 104–107.

[8] D. Bhalla, N. Joshi and I. Mathur, "Rule Based Transliteration Sscheme for English to Punjabi," International Journal on Natural Language Computing (IJNLC), vol. 2, no.2, pp. 67-73, Apr 2013.

[9] B. J. Kang and K. S. Choi, "Automatic Transliteration and Back Transliteration by Decision Tree Learning," in Proceedings of Conference on Language Resources and Evaluation. Athens, Greece, pp. 1135–1411, 2000.

[10] The CMU Pronouncing Dictionary. [Online]. Available: https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict/. [Accessed: Jan 14, 2014].

[11] Google Input Tool. [Online]. Available: https://www.google.com/inputtools/windows/. [Accessed: Feb 28, 2107].