International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 5, September-October 2014 **ISSN 2278-6856**

A Review on Hindi to English Transliteration System for Proper Nouns Using Hybrid Approach

Veerpal Kaur[1], Amandeep kaur Sarao[2] and Jagtar Singh[3]

[1]M.Tech Student (CE), YCOE, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab,India.

[2]Assistant Professor (CE), YCOE, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab,India.

[3]Associate Professor (ECE), YCOE, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab,India.

Abstract

This paper presents the review to transliterate proper nouns written in Hindi language into its equivalent English language. There are various approaches like direct mapping approach, rule based approach and statistical machine translation approach are available to transliterate proper nouns from Hindi language to English language. Transliteration is a process to generate the words from the source language to the target language. The reverse process is known as backward transliteration. Every approach for transliteration has the significant advantages and disadvantages in context of transliteration which will be discussed in this paper. Transliteration from Hindi language to English language plays a very important role as Hindi is official language of India and there is lot of data is present in Hindi which needs to convert into English for global usage.

Keywords: Transliteration, Statistical Machine Translation, Devanagri Script, Machine Translation, Mapping

1. INTRODUCTION

Machine transliteration System accepts characters of source language and map to the characters of the target language. The process is performed into two parts – Segmentation Phase , in which words of the source language are segmented into units and the – Assembly phase , in which segmented characters are mapped to the characters of target language with the help of rules. Transliteration and transcription are opposite to each other. Transcription is which maps the sounds of one language to script of another language. Transliteration maps the letters of source script to letters of pronounced similarly in target script. Transliteration is particularly used to translate proper names and technical terms from languages. Machine transliteration is classified into two categories: Forward transliteration and backward transliteration. For example transliterating the name "**अंबा**" to "amba" is known as forward transliteration while transliteration from "amba" to "**अंबा**" is known as backward transliteration.

2. RELATED WORK

It provides a description, summary and critical evaluation of each work. Following study has been carried out during this research work: Gurpreet Singh Josan et al. (2011) described a novel approach to improve Punjabi to Hindi transliteration system. The accuracy of the proposed technique described in this paper varies from 73% to 85% which can be improved further by using some modified technique. [1] Haque et al. (2009) developed English to Hindi transliteration system based on the phrase-based statistical method (PB-SMT). PB-SMT models have been used for transliteration by translating characters rather than words as in character-level translation systems. They have modelled translation in PB-SMT as a decision process, in which the translation a source sentence is chosen to maximize. They used source context modelling into the state-of-the-art log-linear PB-SMT for the English—Hindi transliteration task. To improve the system performance, they took source context into account substantially. An improvement of 43.44% and 26.42% has been reported respectively for standard and larger datasets. [2]Jia et al. (2009) developed Noisy Channel Model for Grapheme-based Machine Transliteration. They have experimented this model on English-Chinese. Both English-Chinese forward transliteration and back transliteration has been studied. The process has been divided into four steps: language model building, transliteration model training, weight tuning, and decoding. When building language model, data smoothing techniques Kneser-Ney and interpolate has been employed. In transliteration model training step, the alignment heuristic has been grown diag-final, while other parameters have default settings. In the Tunning method all the used parameters have default values. When decoding, the parameter distortion-limit has been set to 0, meaning that no reordering operation is needed. [3] Kamal Deep Singh et al. (2011) developed hybrid (statistical +rules) approach based transliteration system of person names; from a person name written in Punjabi (Gurumukhi Script), the system produces its English

(Roman Script) transliteration. Experiments have shown that the performance is sufficiently high. The overall accuracy of system comes out to be 95.23%. [4] Lehal et al. (2010) developed Shahmukhi to Gurmukhi transliteration system based on corpus approach. In this system, first of all script mappings has been done in which mapping of Simple Consonants, Aspirated Consonants (AC), Vowels, other Diacritical Marks or Symbols are done. This system has been virtually divided into two phases. The first phase performs pre-processing and rule-based transliteration tasks and the second phase performs the task of post-processing. Bi-gram language model has been used in which the bi-gram queue of Gurmukhi tokens has been maintained with their respective unigram weights of occurrence. The Output Text Generator packs these tokens well with other input text which may include punctuation marks and embedded Roman text. Finally, this generates a Unicode formatted Gurumukhi text. The overall accuracy of system has been reported to be 91.37%. [5] Malik et al. (2009) developed Punjabi Machine Transliteration (PMT) system which is rule-based. PMT has been used for the Shahmukhi to Gurmukhi transliteration system. PMT has preserved the phonetics of transliterated word and the meaning of transliterated word. Firstly, two scripts have been discussed and compared. Based on this comparison and analysis, character mappings between Shahmukhi and Gurmukhi scripts have been drawn and transliteration rules are formulated. The primary limitation of this system is that this system works only on input data which has been manually edited for missing vowels or diacritical marks (the basic ambiguity of written Arabic script) which practically has limited use. [6] Sumita Rani et al. (2013) presented various techniques for transliteration from Punjabi language to Hindi Language. Most of the characters in Punjabi language have their same matching part present in a Hindi language. There are some characters exist in Hindi which are double sounds but no such characters are available for Punjabi. The major inaccuracies in the transliteration are due to poor word selection. In this paper, transliteration system described is built on statistical techniques. This system can be developed with minimum efforts. [7] Verma et al. (2006) developed a Roman-Gurmukhi transliteration System and named it GTrans. He has surveyed existing Roman-Indic script transliteration techniques and finally a transliteration scheme based on ISO: 15919 transliteration and ALA-LC has been developed. Because according to linguistics, these systems are closer to the natural pronunciation of Punjabi words as compared with others. Most of the rules for transliteration in both schemes were same except for Bindi and tippi in case of vowels as compared with consonants. Some modifications have done like, bindi and tippi has been represented with the same symbol because both produce similar sounds and has been transliterated in the same way. He has also done reverse transliteration from Gurumukhi to Roman. The overall accuracy of system has been reported to be 98.43%. [8] Vijaya et al. (2009) developed English to Tamil transliteration system and named it WEKA.It is a rule based system and used the j48 decision tree classifier of WEKA for classification purposes. The feature patterns has been extracted from In this system, the valid target language n-gram (y_i) for a source language n-gram (xi) in the given source language input word is decided by considering the source language context features such as source language n-gram (xi) , two left context n-grams (xi-2, xi-1) and two right context n- grams (xi+1, xi+2). The transliteration process consisted of four phases: Pre-processing phase, feature extraction, training and transliteration phase .The accuracy of this system has been tested with 1000 English names that were out of corpus. The transliteration model produced an exact transliteration in Tamil from English words with an accuracy of 84.82%. [9] Vishal Goyal et al. (2009) presented the evaluation results of Hindi to Punjabi machine translation system. After evaluation, the accuracy of the system is found to be about 95%. The accuracy can be improved by improving and extending the bilingual dictionary. Even robust pre processing and post processing of the system can improve the system to greater extent. This system is comparable with other existing system and its accuracy is better than those. [10]

3. EXISTING APPROACHES

Hindi to English transliteration can achieve by using various techniques. In transliteration there are following techniques:

3.1 Direct mapping Approach:

When two languages are structurally similar and have similar vocabulary then direct approach is the best choice. The performance of a direct MT system depends on the quality and quantity of the source-target language dictionaries, text processing software, and word-by-word transliteration with minor grammatical adjustments on word order and morphology. Using direct approach system try to generate the result with the help of parallel corpus provided for training. It generates only those results which are in the parallel corpus. It is the base of the transliteration process. It is also known as character to character mapping.

Advantages

The main advantage of direct mapping approach is that is consumes minimum time to transliterate the proper noun of Hindi language into its equivalent English language as transliteration involves only in searching the source keyword.

Disadvantages

The major disadvantage of this approach is that it can transliterate only those proper nouns which are present in the database. It cannot transliterate those nouns which are not present in the database.

3.2 Rule-based Approach

A rule-based approach is the first strategy that was developed. In this approach various rules are created to perform the task of transliteration. Rules are created by considering the properties of the source and target language. This approach is not used so frequently because

Rule-based approaches take time, money and trained personnel to make and test the rules.
Some examples of rules of Rule Based System:-

1) Rule1: if length of proper noun is of 3 characters containing no vowel then 'A' is removed from the last position. For eg. नमक is transliterated to Namak.
2) Rule2: if length of proper noun is of 4 characters containing no vowel then 'A' is removed from the second and last position. For eg. अमजद is transliterated to Amjad.
3) Rule3: if proper noun ends with a consonant then 'A' should be removed from last position in English spelling. For eg. अमित is transliterated to Amit.
4) Rule4: if proper noun begins with 'आ' then it is replaced with 'A' during transliteration. For eg. आकृति is transliterated to Akriti.
5) Rule5: if ऊ ( ) occurs in middle of proper noun then it is replaced with 'OO'. For eg. दूज is transliterated to Dooj.
6) Rule6: if a name contains 'फ' then mapping of 'फ' becomes 'F' instead of 'PH'. For eg. फूफा is transliterated to Phupha.

Advantages

The main advantage of rule based approach is that if rules are properly created according to the features of both source and target language then system can transliterate those nouns also which are not present in the database.

Disadvantages

Rule based approach for transliteration is very difficult to implement as there are very large number of rules with various exceptions are there in this approach. These rules are created by the human beings are tends to produce errors if they are not properly developed. Another disadvantage of rule based approach is that is works only on the Indian origin names but not on the foreign names.

Table 1: Result Comparison of these two techniques

Approach For Transliteration	Avg. Result
Direct Mapping	47%
Rule Based Approach	67%

4. PROPOSED APPROACH

We use Statistical machine translation (SMT) Approach for transliteration purpose. SMT is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge. It is language independent. SMT has high accuracy of results as compared to rule based approach. It transliterates Indian origin naming entities as well as other names. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model. Our proposed system works in two phases. These two phases are: - System Training Phase and System Transliteration Phase.

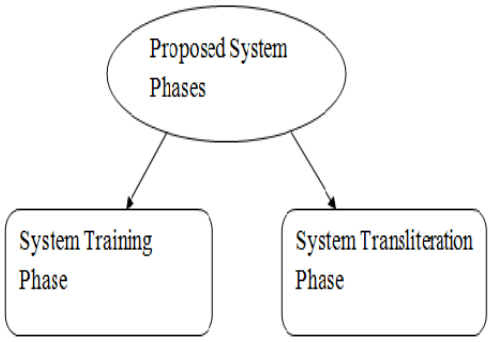

Figure 1 Proposed System

4.1 System Training Phase:

In System Training phase training is given to the system on the basis of names stored into the database and it generates the database tables. Database tables which are bi – gram table, tri – gram table, four – gram table, five – gram table and six – gram table will be filled with the data generated automatically in this phase. Tables are stored into the database.

4.2 System Transliteration Phase:

In System Transliteration Phase transliteration is actually takes places with the help of the data generated in the training phase. For this Purpose, we store more than 18,000 unique names on which the system is trained and in this phase system tries to find the word directly into the database and if word is found then system gives output otherwise with the help of generated tables, system can transliterate new word.

5. SOME EXPECTED OUTCOMES OF THE TRANSLITERATION SYSTEM FROM HINDI TO ENGLISH

Various examples of Hindi to English transliteration system shown in following table:-

Table 2: Examples of Hindi to English Transliteration System

Hindi Proper Noun	English Generated Name
अंबा	Amba
चितेश	Chitesh
अमिता	Amita
हसनदीप	Hasandeep
इंदरजीत	Inderjit
अंजली	Anjali
दर्शन	Darshan
दिनकर	Dinkar
सचिन	Sachin
जगजीत	Jagjeet
वीरपाल	Veerpal
अमनदीप	Amandeep
जगतार	Jagtar
मनप्रीत	Manpreet
शेरगिल	Shergill
कामेश	Kamesh
जयपाल	Jaipal
गोपन	Gopan
गुरमन	Gurman
उपजीत	Upjeet
अमनजोत	Amanjot
राज	Raj
गुरप्रीत	Gurpreet
किरणदीप	Kirandeep

6. CONCLUSION & FUTURE SCOPE

This paper presents the review on transliteration of proper nouns from Hindi language to English language. Various techniques are discussed along with advantages and disadvantages. It is shown that the result from these techniques is not satisfactory and hence there is a need to develop a new technique which can provide the expected results. To improve the results statistical Machine translation technique can be used for the transliteration system. In this type of technique transliteration takes place with the help of the stored examples in the database. Size of the database can be increase considerably to obtain the good results.

References

[1] Lehal and Singh, "A Punjabi to Hindi Machine Transliteration System", Computational Linguistics and Chinese Language Processing Vol. 15, No. 2, June 2010, pp. 77-102.

[2] Haque, Dandapat, Srivastava, Naskar and Way "English—Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009", Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 104–107,Suntec, Singapore, 7 August 2009. ACL and AFNLP.

[3] Jia, Zhu, and Yu, "Noisy Channel Model for Grapheme-based Machine Transliteration", Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 88–91,2009.

[4] Deep and Goyal, "Hybrid Approach for Punjabi to English Transliteration System", International Journal Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.

[5] Lehal and Singh "Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach", proceeding of Advanced Centre for Technical Development of Punjabi Language, Literature & Culture,Punjabi University, Patiala 147 002, Punjab, India, pp-151-162,2008.

[6] Malik, Besacier, Boitet, Bhattacharyya "A Hybrid Model for Urdu Hindi Transliteration", Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 177–185,Suntec, Singapore, 7 August 2009ACL and AFNLP.

[7] Rani and laxmi, "A Review on Machine Transliteration of related languages: Punjabi to Hindi", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 3, March 2013.

[8] Verma "A Roman-Gurmukhi Transliteration system", proceeding of the Department of Computer Science, Punjabi University, Patiala, 2006.

[9] Vijaya ,VP, Shivapratap and KP CEN "English to Tamil Transliteration using WEKA system", International Journal of Recent Trends in Engineering, May 2009, Vol. 1, No. 1, pp: 498-500,2009.

[10] Goyal and Lehal, "Evaluation of Hindi to Punjabi Machine Translation System", IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009 ISSN (Online): 1694-0784

[11] Josan & Kaur, "Punjabi to Hindi Statistical Machine Transliteration", International Journal of Information Technology and Knowledge Management July-December 2011, Volume 4, No. 2, 2011, pp. 459-463.

[12] Malik, "Punjabi Machine Transliteration System", In Proceedings of the 21st International Conference on

Computational Linguistics and 44th Annual Meeting of the ACL (2006), pp. 1137-1144.

[13] Manikrao L Dhore, "Hindi to english machine transliteration of named entities using conditional random fields", International Journal of Computer Applications (0975 – 8887), Volume 48– No.23, June 2012.

[14] Pankaj Kumar and Er.Vinod Kumar, "Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns" International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 8, August 2013, ISSN 2319-4847

[15] Sumita Rani, Vijay Laxmi, "A Review on Machine Transliteration of related language: Punjabi to Hindi", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 3, March 2013.

[16] Tejinder Singh Saini and Gurpreet Singh Lehal, "Word Disambiguation in Shahmukhi to Gurmukhi Transliteration", Proceedings of the 9th Workshop on Asian Language Resources, IJCNLP 2011, Chiang Mai, Thailand, pp. 79–87.

[17] UzZaman, Zaheenand , Khan, "A Comprehensive Roman (English)-To-Bangla Transliteration Scheme", International Conference on Computer Processing on Bangla (ICCPB-2006), 17 February, 2006, Dhaka, Bangladesh.

[18] Ali and Ijaz, "English to Urdu Transliteration System", Proceedings of the Conference on Language & Technology 2009, pp: 15-23.

[19] Antony, Ajith, Soman, "Kernel Method for English to Kannada Transliteration", International Conference on Recent Trends in Information, Telecommunication and Computing, 2010, pp: 336-338.

[20] Aswani, Robert, "English-Hindi Transliteration using Multiple Similarity Metrics", www.mt-archive.info/LREC-2010-Aswani.pdf, pp: 1786-1793

[21] Abbas Malik, Laurent Besacier, Christian Boitet, Pushpak Bhattacharyya, "A Hybrid Model for Urdu Hindi Transliteration", Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP, 2009, Singapore. pp. 177–185.

AUTHOR

Veerpal Kaur received her Bachelor of Engineering Degree in Information Technology from Guru Kashi University, Talwandi Sabo, Bathinda, Punjab, India in 2011 and pursuing Master of Technology Degree in computer science & Engineering from Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India. She is doing her thesis on Hindi to English Transliteration System for Proper Nouns Using Hybrid Approach. Her Research area is Natural Language Processing and Machine Translation

Amandeep Kaur Sarao received her Bachelor of Engineering Degree in Computer Engineering from Chandigarh University, Chandigarh, Punjab, India in 2006 and Master of Technology Degree in Computer Engineering from Punjabi University Patiala, Punjab, India in 2009. She is an assistant professor in the Computer Engineering Section, Yadavindra College of Engineering, Punjabi University, Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India. She has published research papers in various International Journals and Conferences. Her main research interests are in Network Security. She is a member of International Association of Engineers (IAENG).

Jagtar Singh Sivian received his Bachelor of Engineering Degree in Electrical and Electronics Communication from Punjab Technical University Jalandhar, Punjab, India in 1999 and Master of Technology Degree in Electronics Communication from Punjab Technical University Jalandhar, India in 2005. Currently He is pursuing Ph.D. degree from Sant Longowal Institute of Engineering and Technology from Longowal, Sangrur, India. He is an associate professor in the Electronics Communication Engineering Section, Yadavindra College of Engineering, Punjabi University, Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India. He has published more than 30 papers in various International Journals and Conferences. His main research interests are in Neural Networks, Genetic Algorithms, Antenna System Engineering, Electromagnetic Waves and Control Engineering. He is a member of Institution of Engineers (India) and International Association of Engineers (IAENG).