DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4620
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

MACHINE TRANSLITERATION FOR INDIAN LANGUAGES: A REVIEW

Anupama Sharma
Department of Computer Engineering
Punjabi University
Patiala, India

Dr. Dhavleesh Rattan
Department of Computer Engineering
Punjabi University
Patiala, India

Abstract: This paper addresses various developments in Indian language Machine Transliteration system, which is a very important task for many natural language processing (NLP) applications. There are a number of different transliteration systems available for different languages. Machine transliteration is receiving much research attention in recent years.

Keywords: Natural language processing; transliteration; phonemes; graphemes.

1. INTRODUCTION

Machine transliteration is the process of converting a word written in a source language into a word in a target language by preserving the word pronunciation. The script of the target word is different from that of the source language word. The important consideration for the transliteration is that the phonetic structure of word should be preserved as closely as possible [6]. Transliteration is used for handling named entities and out of vocabulary words in Machine Translation. The process of mapping of source language phonemes or graphemes into target language phonemes or graphemes is known as Forward Transliteration, the reverse process is called back transliteration [11]. Transliteration is used mainly for handling named entities and out of vocabulary words [11].

A. Machine Transliteration Approaches

Transliteration can be classified in to three categories Grapheme based, Phoneme based, hybrid based approaches. This approach treats transliteration as an orthographic process and the source graphemes are directly mapped to the target graphemes. In Grapheme based approach the following models are used (i) source channel model (iii) Maximum Entropy Model (iii) Conditional Random Field models and (iv) Decision Trees model. The grapheme based transliteration is also known as the direct method because it directly maps source language graphemes into target language graphemes without any phonetic knowledge.
Phoneme based approach treats transliteration as a phonetic process. In Phoneme based approach the following models are used (i) Weighted Finite State Transducers (WFST) and (ii) extended Markov window (EMW). The phoneme-based transliteration approach is also known as the pivot method because it uses source language phonemes as a pivot. In this approach transliteration is done in two steps: 1) source language graphemes are converted into source language phonemes and 2) source phonemes are converted into target language graphemes. A hybrid model uses combination of a grapheme based model and a phoneme based approach.

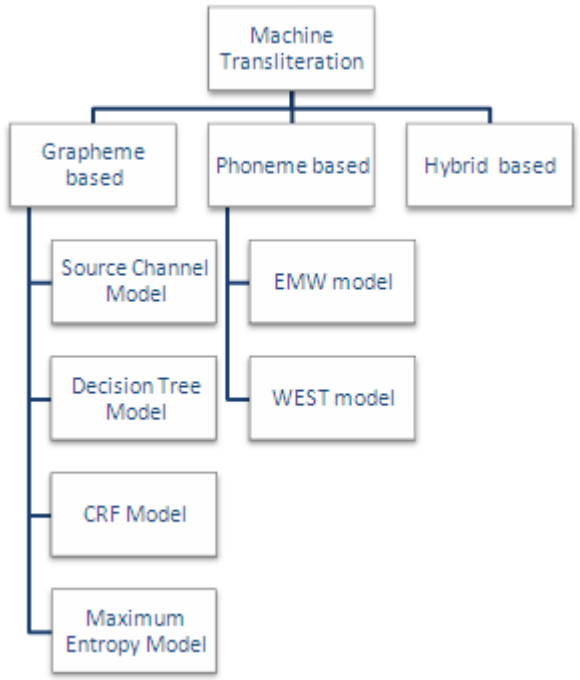

Figure 1. Machine transliteration approaches.

2. BACKGROUND

Rathod et al [1] discussed machine transliteration for Hindi to English and Marathi to English language pairs using Support Vector Machine (SVM). Their system is divided into three modules. 1. Preprocessing 2. Training of bilingual corpus 3. Testing of additional data. Preprocessing phase converts the input into the system's acceptable format. Syllabification is done in this phase. Classification is done in the training phase which is based on the n-gram.
Lehal [2] discussed Gurmukhi to Shahmukhi transliteration. In the pre-processing stage, the Gurmukhi word is prepared for transliteration according to Shahmukhi word. In processing stage transliteration is done using rule based approach. In post- processing stage, transliterated words are corrected by using Shahmukhi corpus. The transliteration accuracy was 98.6%.

Dhore et al [3] discussed Hindi to English machine transliteration of Indian named entities. They used conditional random fields (CRF). They also discussed the issues due to which the direct transliteration of Hindi to English is quite difficult. For syllabification word written in Devanagari script is divided into akshars. Their system accuracy is 85.79%.

Chinnakotla and Damini [4] discussed Hindi and Persian into English transliteration. They discussed Character Sequence Modeling (CSM) which is a probability distribution of characters in a word. They used character mapping table between the Hindi and English.

Shahnawaz [5] discussed conversion between Hindi and Urdu. She discussed the machine translation, approaches of Machine Translation and transliteration. She also discussed the impact of differences in spellings, pronunciation and writing style on conversion.

Antony and Soman [6] discussed various developments in Indian language machine transliteration system. They discussed various contributors of Machine Transliteration. They also discussed various Machine Transliteration approaches.

Ganesh et al [7] discussed Hindi to English transliteration. They used Hidden Markov Model (HMM) alignment and Conditional Random Fields (CRF). The task of HMM alignment is to maximize the probability of the source-target word and then the character level alignment. CRF is used for training and decoding which is conditioned on both the source and target language pairs.

Josan and Kaur [8] discussed Punjabi to Hindi transliteration. They discussed Gurumukhi and Devnagari scripts. They discussed baseline approach of Transliteration that is character to character mapping from source language alphabets to target language alphabets. They also discussed problems in this approach. Then they discussed Statistical approach of Transliteration.

Rani and Laxmi [9] discussed Punjabi to Hindi transliteration. They discussed direct character to character mapping approach of transliteration. They also discussed transliteration systems build by using statistical techniques.

Ramakrishnan et al [10] presented a browser plugin to Google Chrome that transliterates a website from any Indic script to Kannada. They used rule-based approach to transliterate to Kannada. The source languages supported by their systems are Tamil, Telugu, Malayalam, Bangla, Gujarati, Odiya, Punjabi, Sanskrit and Hindi. They also framed some rules to transliterate the above stated languages to Kannada language.

3. CONCLUSION

In this paper work, we have presented a review on work done in the field of machine transliteration systems for Indian languages. We also tried to give a brief idea on the approaches used to develop machine transliteration systems.

4. REFRENCES

1. P. H. Rathod, M L Dhore, R. M. Dhore, HINDI AND MARATHI TO ENGLISH MACHINE TRANSLITERATION USING SVM, International Journal on Natural Language Computing (IJNLC), 2(4), 2013, pp. 57-71.
2. G.S. Lehal, A GURMUKHI TO SHAHMUKHI TRANSLITERATION SYSTEM, in: proceedings of ICON-2009: 7th international conference on Natural Language Processing, 2009, pp.167-173.
3. M.L. Dhore, S.K. Dixit, T.D. Sonwalkar, Hindi to English Machine Transliteration of Named Entities using Conditional Random Fields, International Journal of Computer Applications (0975 – 8887), 48(23), 2012, pp. 31-37.
4. M.K. Chinnakotla, O.P. Damani, Character Sequence Modeling for Transliteration, in: proceedings of ICON-2009: 7th international conference on Natural Language Processing, 2009.
5. Shahnawaz, Conversion between Hindi and Urdu, in: Proceedings of International Conference on Computing, Communication and Automation (ICCCA2015), 2015, pp. 309-313.
6. Antony P. J, Soman K P, Machine Transliteration for Indian Languages: A Literature Survey, International Journal of Scientific & Engineering Research, 2(12), 2011.
7. S.Ganesh, S.Harsha, P.Pingali, V.Varma, Statistical Transliteration for Cross Langauge Information Retrieval using HMM alignment and CRF, in: proceedings of The 2nd International workshop on Cross lingual information access, 2008, pp. 42-47.
8. G.S. Josan, J. Kaur, Punjabi to Hindi Statistical Machine Transliteration, International Journal of Information Technology and Knowledge Management, 4(2), 2011, pp. 459-463.
9. S. Rani and V. Laxmi, A Review on Machine Transliteration of related languages: Punjabi to Hindi, International Journal of Science, Engineering and Technology Research (IJSETR), 2(3), 2013, pp. 733-736.
10. A.G. Ramakrishnan, S.S. Rao, R. D. Sequiera, S. Kumar H R, Transliteration of Indic Languages to Kannada with a User-Friendly Interface, in: proceedings of IEEE International Advance Computing Conference (IACC), 2015.
11. G.S. Josan and G.S Lehal, A Punjabi to Hindi Machine Transliteration System, Computational Linguistics and Chinese Language Processing, 15(2), 2010, pp. 77-102.