ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 5, May 2013

Machine Transliteration system in Indian perspectives

Jasleen Kaur

***Abstract*— This paper addresses the various progresses in Indian Machine Transliteration systems, which is considered as a very important part for many natural language processing applications. Transliteration is the general choice for handling OOV words. Accurate transliteration of named entities plays an important role in the performance of machine translation. Although a number of different transliteration mechanisms are present for world's top level languages like English, European languages and Asian languages like Japanese, Chinese, Korean and Arabic, still it is an initial stage for Indian languages. Literature shows that some attempts have done for few Indian languages like Hindi, Punjabi, Shahmukhi, Gurmukhi, Tamil, Kannada, Telugu and Bengali language.**

***Index Terms*— Named entities, transliteration, machine translation, Indo Aryan languages.**

I. Introduction

Machine processing of Natural (Human) Languages has a long tradition, benefiting from decades of manual and semi-automatic analysis by linguists, sociologists, psychologists and computer scientists among others. This cumulative effort has seen fruit in recent years in the form of publicly available online resources ranging from dictionaries to complete machine translation systems. In a large multilingual society like India, there is great demand for translation of documents from one language to another language. Though work in the area of machine translation has been going on for several decades. Efficient methods for machine transliteration, which is the one of the important part of machine translation, continue to be a challenging task. Machine transliteration is the practice of transcribing a character or word written in one alphabetical system into another alphabetical system. Machin**e** transliteration can play an important role in natural language application such as information retrieval and machine translation, especially for handling proper nouns and technical terms, cross-language applications, data mining and information retrieval system.

The topic of machine transliteration has been studied extensively for several different language pairs. Various methodologies have been developed for machine transliteration based on the nature of the languages considered. The development of algorithms specifically for machine transliteration began over a decade ago based on the phonetics of source and target languages, followed by approaches using statistical and language-specific methods. Most of the current transliteration systems use a generative model based on alignment for transliteration and consider the task of generating an appropriate transliteration for a given word. Such model requires considerable knowledge of the languages.

The transliteration is straightforward if all the phoneme representations are present in both languages e.g. transliteration of name "ਕਬੀਰ" [kabir], in source language Punjabi, is "कबीर", in target language Hindi, which is essentially pronounced in the same way. But in real world, this barely happens. Generally the two scripts vary and some of the sounds are missing or are extra in the target language. Some type of mapping is done for the missing or extra phonemes to the most phonetically similar letter, *e.g.*, in Hindi for alphabet "थ", no such corresponding letter is present in English. So generally a similar sounding letter or combination of letters is used to denote such sounds for example "th" for representing above alphabet. For example, "ph" and "f" both map to the same sound of (f).The transliteration model should be design while considering all these complexities

II. CONTRIBUTORS TO MACHINE TRANSLITERATION

One of the works on Transliteration is done by Arababi et al [1]. Arababi etal [1] model forward transliteration through a combination of neural net and expert systems for transliteration from Arabic-English in 1994. In 1997, Knight and Graehl proposed generative model for back transliteration from English to Japanese katakana [2].This statistical based approach was used by stall and knight[3].In year 2000,Jung et al[4] proposed a statistical English Korean transliteration model that exploits various information sources. This model is a generalized mod-el from a conventional statistical tagging model by extending Markov window with some mathematical approximation techniques. An alignment and syllabification method is developed for accurate and fast operation. In year 2000, Kang [5] presented an automatic character alignment method between English word and Korean transliteration. Aligned data is trained using supervised learning decision tree method to automatically induce transliteration and back-transliteration rules. This methodology is fully bi-directional, i.e. the same methodology is used for both transliteration and back transliteration. In year 2002, Y. Al-Onaizan and K. Knight [6] developed a hybrid model based on phonetic and spelling mappings using Finite state machines. The model was designed for transliterating Arabic names into English.In 2003 Nasreen and Larkey[7] had presented a method for automatically learning a transliteration model from a sample of name pairs in English and Arabic languages. In their paper, simple statistical technique for English to Arabic transliteration was evaluated. The technique learns translation probabilities between English and Arabic characters from a training sample of pairs of transliterated words from the two languages. The accuracy of this system increases with the size of the training set in both aligned conditions. Aligned training is more effective than unaligned training, and bigrams are more effective than monograms. This system was evaluated with respect to how well it can

Manuscript received April 2013

Jasleen Kaur: *Computer science departmenty/Shroff S.R.Rotary Institute of chemical Technology., Ankleshwar, India,Country /+91-8128672321*

generate correct Arabic transliterations from the Arabic Proper Names dictionary for a test set of English words, after training on a non-overlapping set of word-pairs from the same source. In 2004, Li Haizhou, Zhang Min, Su Jian[8] presents a new framework that allows direct orthographical mapping (DOM) between two different languages, through a joint source-channel model, also called n-gram transliteration model (TM). It generates probabilistic orthographic transformation rules using a data driven approach. By skipping the intermediate phonemic interpretation, the transliteration error rate is reduced significantly. The bilingual aligning process is integrated into the decoding process in n-gram TM, which allows us to achieve a joint optimization of alignment and transliteration automatically. The new framework greatly reduces the development efforts of machine transliteration systems. Although the framework is implemented on an English-Chinese personal name data set, without loss of generality, it well applies to transliteration of other language pairs such as English/Korean and English/Japanese. In 2005 Jong-Hoon et al. [9] presents Hybrid transliteration model which is based on both grapheme and phoneme information. Through this combination they achieved performance improvements. In this paper, they showed both grapheme as well as phoneme information is useful for machine transliteration. They showed Machine based Learning is the best machine learning method and correct pronunciation is very helpful to generate a correct Korean transliteration. Their method uses both grapheme and phoneme information in English-to-korean transliteration which achieves 13%-78% performance improvements.

III. CLASSIFICATION OF TRANSLITERATION APPROACHES

Transliteration is classified into three types namely, Grapheme based, Phoneme based, Hybrid models and Correspondence based transliteration models. These models are classified in terms of the units to be transliterated. Grapheme-based transliteration model perform direct orthographical mapping from source graphemes to target Graphemes. This is also referred to as the direct method because it directly transforms source language graphemes into target language graphemes without any phonetic knowledge of the source language words. Several transliteration methods based on this model have been proposed, such as those based on a source-channel model, a decision tree, a transliteration network, and a joint source-channel model.

In Phoneme based transliteration model the transliteration key is pronunciation or the source phoneme rather than spelling or the source grapheme. This model is basically source grapheme-to-source phoneme transformation and source phoneme-to-target grapheme transformation. WFST (weighted Finite State transducers), and extended Markov window are the approaches belong to the phoneme based models. Phoneme based models treat transliteration is treated as phonetic process rather than an orthographic process. It needs two steps: 1) produce source language phonemes from source language graphemes; 2) produce target language graphemes from source phonemes.

Hybrid transliteration model and correspondence-based transliteration model use both source graphemes and source phonemes in machine transliteration. The correspondence based transliteration model makes use of the correspondence between a source grapheme and a source phoneme when it produces target language graphemes; the hybrid based models simply combines grapheme information and phoneme information through linear interpolation. Note that the hybrid model combines the grapheme-based transliteration probability and the phoneme-based transliteration probability using linear interpolation

IV. MACHINE TRANSLITERATION FOR INDIAN LANGUAGES

India is home to several hundred languages. Most Indians speak a language belonging either to the Indo-European (ca. 74%), the Dravidian (ca. 24%), the Austro-Asiatic (Munda) (ca. 1.2%), or the Tibeto-Burman (ca. 0.6%) families, with some languages of the Himalayas still unclassified. The SIL Ethnologue lists 415 living languages for India [9].As per The Times of India, the number of internet users worldwide is expected to touch 2.2 billion by 2013.For providing the internet access to rural Indians ,the government of India had taken number of projects. This signifies the importance for providing the information in regional languages so that it can be easily understandable to its regional users. Many technical terms and proper names, such as personal name, location name, organization name, are translated from one language into another language with approximate phonetic equivalents. The following subsections describe various machine transliteration developments in Indian Languages.

A. English to Indian Language Machine Transliteration

Various machine transliteration systems are developed from English to various Indian languages. The following are the noticeable developments in English to Hindi or other Indian language or vice-versa machine transliteration.

--English to Hindi Transliteration system was developed by Amitava Das, Asif Ekbal, Tapabrata Mandal and Sivaji Bandyo-padhyay based on NEWS 2009 Machine Transliteration Shared Task training datasets [10]. The proposed transliteration system uses the modified joint source channel model along with two other alternatives to translate English to Hindi transliteration. The system also uses some post processing rules for the purpose of removing the errors in the system to improve the accuracy. They performed one standard run and two nonstandard runs in the developed English to Hindi transliteration system. The results showed that the performance of the standard run was better than the non standard one. Hindi to English.

--The paper [11] addresses the issue related to statistical machine transliteration from English to Punjabi. Statistical Approach to transliteration is used for transliteration from English to Punjabi using MOSES, a statistical machine translation tool. The Efficiency of this transliteration system is evaluated manually as well as using BLEU metrics. The system is improved by applying some transliteration rules at post processing stage Average %age accuracy and BLEU score of this transliteration system without applying transliteration rules is 50.22% and 0.4123 respectively. After applying transliteration rules, average %age accuracy and BLEU score comes out to be 63.31% and 0.4502 respectively. Further improvements can be done in this transliteration system from English to Punjabi. One of major weakness of transliteration from English to Punjabi is dealing with multiple mapped characters.

--Kamaldeep ,Vishal Goyal in [12] have addressed the problem of transliterating Punjabi to English language using statistical rule based approach. Punjabi to English transliteration system is very beneficial for removing the language and scriptural barrier. They 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%.

--Vijaya, VP, Shivapratap and KP CEN [13] has developed English to Tamil Transliteration system and named it WEKA. It is a Rule based system and is used the 48 decision tree classifier of WEKA for classification purposes. The transliteration process consisted of four phases: Preprocessing 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%.

UzZaman, Zaheenand, Khan [14] has developed Roman (English) to Bangla transliteration system. Two mappings, one is direct phonetic mapping and second location enabled phonetic mapping, has been used. The user is provided with multiple options of lettergroups in the source script (which, in this case, is Roman) to represent one letter in the goal script (Bangla), (many-to-one mapping scheme). This scheme can be used in applications such as cross language information query and retrieval.

--English to Kannada transliteration system was developed using a publically available translation tool called Statistical Machine Translation (SMT) [15].The model is trained on 40,000 words containing Indian place names. During the training phase the model is trained for every class in order to distinguish between examples of this class and all the rest. The SVM binary classifier predicts all possible class labels for a given sequence of source language alphabets and selects only the most probable class labels. Also SVM generate a dictionary which consists of all possible class labels for each alphabet in the source language name. This dictionary avoids the excessive negative examples while training the model and training become faster. This transliteration technique was demonstrated for English to Kannada Transliteration and achieved exact Kannada transliterations for 89.27% of English names.

B. Indian to another Indian Language Machine Transliteration

Various machine transliteration systems are developed from Indian to other Indian languages.Brief summary of that is given below:

--A Punjabi to Hindi transliteration system was developed by Gurpreet Singh Josan and Jagroop Kaur based on statistical approach in 2011 [16]. The system used letter to letter mapping as baseline and try to find out the improvements by statistical methods. They used a Punjabi – Hindi parallel corpus for training and publically available SMT tools for building the system.

--Vishal Goyal Gurpreet Singh Lehal[17] has taken Hindi as source language and Punjabi as target language to generate transliteration for out of vocabulary words.Hindi and Punjabi are closely related languages and hence it is comparatively easy to develop than the system between very different language pairs like Hindi and English. They have implemented approximately fifty complex rules for making the transliteration between Hindi-Punjabi language pair accurate after studying both the languages in details. The system found to give accuracy of about 98%.

-- M. G. Abbas Malik [18] developed Punjabi Machine Transliteration System that is used to transliterate words from Shahmukhi script to Gurmukhi script. Punjabi Machine Transliteration (PMT) is a special case of machine transliteration and is a process of converting a word from Shahmukhi (based on Arabic script) to Gurmukhi (derivation of Landa, Shardha and Takri, old scripts of Indian subcontinent), two scripts of Punjabi, irrespective of the type of word. The Punjabi Machine Transliteration System uses transliteration rules (character mappings and dependency rules) for transliteration of Shahmukhi words into Gurmukhi. The PMT system can transliterate every word written in Shahmukhi.

--Finite-state Transducers (FST), very efficient to implement inter-dialectal transliteration, are used to perform transliteration from Hindi to Urdu language. Malik, M G Abbas; Boitet, Christian; Bhattacharyya, Pushpak. in[19] introduce UIT(universal intermediate transcription) for the same pair on the basis of their common phonetic repository in such a way that it can be extended to other languages like Arabic, Chinese, English, French, etc.They describe a transliteration model based on FST and UIT, and evaluate it on Hindi and Urdu corpora. The Hindi –Urdu Machine Transliteration system gives 97.50% accuracy when it is applied on the Hindi-Urdu corpora containing 412,249 words in total.

V. CONCLUSION

In this paper, Survey on developments of different machine transliteration systems for Indian languages are presented. Various different existing approaches that have been used to develop machine transliteration tools are also explained. From the survey I found out that almost all existing Indian language machine transliteration systems are based on statistical and hybrid approach. The main effort and challenge behind each and every development is to design the system by considering the agglutinative and morphological rich features of language

REFERENCES

[1] Arbabi, M., Fischthal, S. M., Cheng, V. C., And Bart, E. Algorithms for Arabic Name Transliteration.IBM Journal of Research and Development 38, 2, 183, 1994.

[2] Knight, Kevin and Graehl, Jonathan.. Machine Transliteration. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. 1997, pp. 128-135.

[3] Stalls, B. G., & Knight, K.. Translating names and technical terms in arabic text.In Proceedings of COLING/ACL Workshop on Computational Approaches to Semitic Languages, 1998, pp. 34–41.

[4] Jung, S. Y., Hong, S., & Paek, E.. English to Korean transliteration model of extended markov window. In Proceedings of the 18th conference on Computational linguistics, 2000, pp. 383 – 389.

[5] Kang, B. J., & Choi, K. S. Automatic transliteration and back-transliteration by decision tree learning. In Proceedings of the 2nd International Conference on Language Resources and Evaluation, 2000, pp. 1135–1411.

[6] Y. Al-Onaizan and K. Knight ,"Machine Transliteration of Names in Arabic Text", Proc. of ACL Workshop on Computational Approaches to Semitic Languages, 2002.

[7] Nasreen AbdulJaleel Leah S.Larkey "English to Arabic transliteration for Cross Language Information Retrieval: A Statistical Approach"in Proceedings of the 12th international conference on information and knowledge management, 2003, pp-139-146.

[8] Jong-Hoon Oh Key-Sun Choi"Machine Learning Based nglish-to-Korean Transliteration using Grapheme and Phoneme information" IEICE TRANS.INF.& SYST., VOL.E88-D, NO.7,july2005,pp 1737-1748.

[9] H.Li, M.Zhang, and J.shu "A Joint Source-channel model for Machine Transliteration", Proc.ACL2004. pp 160-167.

[10] Article "List of languages by number of native speakers in India" accessed from http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India on July 2012.

[11] Amitava Das, Asif Ekbal, Tapabrata Mandal and Sivaji Bandyopadhyay "English to Hindi Machine Transliteration System at NEWS 2009" Proceedings of the 2009 Named Entities workshop, ACL-IJCNLP 2009, pages 80–83, Suntec, Singapore.

[12] Jasleen Kaur,Gurpreet Singh josan "Statistical Approach to Transliteration from English to Punjabi" International Journal on Computer Science and Engineering, Vol. 3 No. 4 Apr 2011,pp1518-1527.

[13] Kamal Deep, Vishal GoyalHybrid Approach for Punjabi to English Transliteration System, International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.

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

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

[16] Antony P J, Ajith V P and Soman K P, 'Statistical Method for Eng-lish to Kannada Transliteration', Lecturer Notes in Computer Science-Communications in Computer and Information Science (LNCS-CCIS), Vo-lume 70, 2010,356-362, DOI: 10.1007/978-3-642-12214-9_57.

[17] Gurpreet Singh Josan & Jagroop 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.

[18] Goyal V., Lehal G. S., 'Hindi-Punjabi Machine Transliteration System (For Machine Translation System)", George Ronchi Foundation Journal, Italy, **64**, n.1, 2009.

[19] M. G. Abbas Malik "Punjabi Machine Transliteration" Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1137–1144, Sydney, July 2006.

[20] Malik, M G Abbas; Boitet, Christian; Bhattacharyya, Pushpak. 2008. *Hindi Urdu Machine Transliteration using Finite-state Transducers*. In proceedings of the 22nd International Conference on Computational Linguistics, August 18 - 22, 2008, Manchester, UK