# Review of Machine Transliteration Systems

Kamaljeet Kaur
M.Tech. Student

Parminder Singh
Associate Professor

Department of Computer Science & Engineering
Guru Nanak Dev Engg. College, Ludhiana (Punjab) India

***Abstract* - Transliteration is the conversion of a text from one script to another, and thus representing words from one language using the approximate phonetic or spelling equivalents of another language. Machine Transliteration has come out to be an emerging and a very important research area in the field of machine translation. Transliteration systems are very beneficial for removing the language and scriptural barrier. It has gained prime importance as a supporting tool for machine translation and cross-language information retrieval, especially when proper names and technical terms are involved. This paper is intended to give a brief overview on the research work carried out on transliteration for Indian as well as for foreign languages.**

***Keywords: Transliteration, named entity, word accuracy rate, natural language processing.***

## I. INTRODUCTION

Machine transliteration has come out to be an emerging and a very important research area in the field of machine translation. Transliteration basically aims to preserve the phonological structure of words. Proper transliteration of name entities plays a very significant role in improving the quality of machine translation. The performance of machine translation and cross-language information retrieval depends extremely on accurate transliteration of named entities.

Transliteration is the conversion of a text from one script to another. It involves representing words from one language using the approximate phonetic or spelling equivalents of another language. Transliteration is totally different from translation, for instance English word "Silver" when translated into Punjabi language become 'ਚਾਂਦੀ'. The 'Silver' in English is transliterated to 'ਸਿਲਵਰ' into Punjabi instead of 'ਚਾਂਦੀ'. So we can say that translation tells the meaning of the word in another language and transliteration helps you to pronounce them.

From an information-theoretical point of view, systematic transliteration is a mapping from one system of writing into another, word by word, or ideally letter by letter. Transliterating a word from the language of its origin to a foreign language is called *Forward Transliteration*. For example English to Punjabi transliteration. On the other hand, transliterating a loan-word (a word borrowed from other language and incorporated) written in a foreign language back to the language of its origin is called *Backward Transliteration*.

Transliteration systems are very beneficial for removing the language and scriptural barrier. Machine transliteration has gained prime importance as a supporting tool for machine translation and cross language information retrieval especially when proper names and technical terms are involved. Machine 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.

## II. MACHINE TRANSLITERATION TECHNIQUES

Various techniques for transliteration are being used, each having its own advantages and disadvantages. These techniques are categorized as shown in Figure 1.

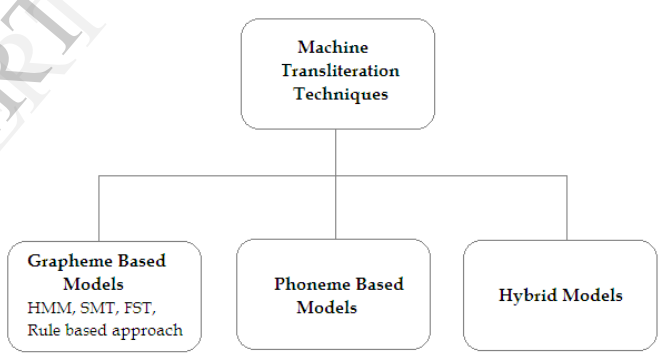


Figure 1. Machine Transliteration Techniques

Grapheme refers to the basic unit of a written language that has its own meaning or grammatical importance. Phonemes are the smallest significant unit of sound. In Grapheme based approaches, transliteration is viewed as a process of mapping a grapheme sequence from a source language to a target language ignoring the phoneme-level processes. Grapheme based models are classified into the Statistical Machine Transliteration based model, Decision Tree based model, Rule based models, Hidden Markov Model (HMM), Finite State Transducer (FST) based model etc.

Grapheme based models work by directly transforming source language graphemes into target language graphemes without explicitly utilizing phonology in the bilingual mapping. Phoneme based models, on the other hand, do not utilize orthographic information in the transliteration process. Phoneme based models are generally implemented in two steps- first obtaining the source language pronunciation and then converting that representation into the target language graphemes.

In phoneme based approaches, the transliteration key is pronunciation of the source phoneme rather than spelling or the source grapheme. This approach is basically source grapheme-to-source phoneme transformation and source phoneme-to-target grapheme transformation.

A rule based machine transliteration system consists of collection of rules called grammar rules, lexicon and software programs to process the rules. Rule based approach is the first strategy ever developed in the field of machine translation. RBMT (Rule Based Machine Transliteration) has much to do with morphological, syntactic and semantic information about the source and target language. Linguistic rules are built over this information. Rules play major role in various stages of translation: syntactic processing, semantic interpretation, and contextual processing of language. Rule based transliteration is based on linguistic information about source and target languages. The main approach of RBMT systems is based on linking the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning. There are three different types of rule-based machine translation systems: Direct Systems (Dictionary Based Machine Translation) map input to output with basic rules. Transfer RBMT Systems (Transfer Based Machine Translation) employ morphological and syntactical analysis. Interlingual RBMT Systems (Interlingua) use an abstract meaning. Transliteration in rule based system is done by pattern matching of the rules. The success lies in avoiding the pattern matching of unfruitful rules. General world knowledge is required for solving interpretation problems such as disambiguation.

Statistical based transliteration approaches tend to be computationally easier in language transliteration than trying to parse and evaluate grammatical rules. Statistical approaches employ various mathematical techniques. Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability. The best translation, of course, is the sentence that has the highest probability. The key problems in statistical MT are: estimating the probability of a translation, and efficiently finding the sentence with the highest probability. It works by finding most probable English sentence given a foreign language sentences, automatically align words and phrases within sentence pairs in a parallel corpus and then probabilities are determined automatically by training a statistical model using the parallel corpus. Based on the probabilities sentence get transliterated. Statistical approaches have a number of advantages over these non-statistical techniques. The primary advantage is that they have been shown to produce better transliteration. It has a way of dealing with lexical ambiguity.

A frequently used statistical model is the Hidden Markov Model (HMM). In Hidden Markov Models (HMM), Given some sentence $x=(x_1,x_2,..,x_n)$ in the language you want to transliterate from, you want to predict what the most likely sentence $y=(y_1,y_2,\ldots,y_m)$ is going to be in the language you want to transliterate to. The HMM works by looking at all possible combinations of a sequence of one language words in another language which computes probabilities of co-occurrence of words based on a given tagged corpus and then tags texts using these probabilities.

Finite State Transducers are models that are being used in different areas of pattern recognition and computational linguistics. In the area of machine transliteration the transducer based approaches that are based on building models automatically from training examples are becoming more and more attractive. A transducer has the intrinsic power of transducing or transliterating. Whenever the transducer shifts from one state to another, it will print the output word, if any. So, as a result, not only will it accept the sentence of one language, but it will print the transliteration in another language. Alternatively, a transducer can be seen as a bilingual generator.

In hybrid approaches, it simply combines the grapheme based transliteration probability and the phoneme based transliteration probability using linear interpolation. Hybrid machine transliteration approach strength the statistical and rule based transliteration methodologies. This reduces the rate of error in transliteration system to a great extent. It can be combination of grapheme based model and phoneme based model or can be combination of any grapheme based models. For example statistical machine transliteration with rule based approach.

## III. MACHINE TRANSLITERATION SYSTEMS FOR INDIAN LANGUAGES

This section gives a brief description of various approaches towards machine transliteration, used by various transliteration attempts for Indian languages.

### *(i) Punjabi to English transliteration*

Deep and Goyal (2011) have proposed a transliteration system that addresses the problem of forward transliteration of person names from Punjabi to English by set of character mapping rules. The proposed transliteration scheme uses grapheme based method to model the transliteration problem. System evaluated for names from the different domains like person names, city names, state names, river names, etc. The proposed technique has demonstrated transliteration from Punjabi to English for conman names of persons, cities, states, rivers etc. and achieved accuracy of 93.22%. The system is accurate for the Punjabi words but not for the foreign words.

Kumar and Kumar (2013) have presenting Statistical Machine Translation system to transliterate proper nouns written in Punjabi language into its equivalent English language. They have presented a statistical machine translation based approach to transliterate proper nouns of Gurumukhi script into its English equivalent names. The system is tested on various names belongs to various regions and overall accuracy of the system is very good. Accuracy of the system is depends on correctness of data stored into the database. The system can be furthered improved by adding unique names to the database. The proposed system is tested on more than 1000 names and system given as accuracy of 97%.

### *(ii) English to Hindi transliteration*

Rama and Gali (2009) have addressed the transliteration problem as a translation problem. They have used phrase based SMT technique for English-Hindi language pair. They

used SMT system, GIZA++, beam search based decoder for developing the transliteration model. They applied English-Hindi aligned word corpus to train and test the system. Results of proposed system show that these techniques can be successfully used for the task of machine transliteration. The achieved accuracy of the system on the test set is 46.3%.

Das et al. (2009) have proposed transliteration system that is based on news corpus. They have trained the transliteration system using the English-Hindi datasets obtained from the NEWS 2009 Machine Transliteration Shared Task. English named entities are divided into transliteration units. The proposed system also considers the English and Hindi contextual information in order to calculate the probability of transliteration from each English unit to various Hindi candidate transliteration units and chooses the one with maximum probability. They have also devised some post processing rules to remove the errors.

Sharma et al. (2012) have used phrase based statistical machine translation technique to transliterate English-Hindi language pair consisting of Indian names using different character encoding for target language (i.e. in UTF and wx-notation). They have applied two different statistical applications MOSES and Stanford Phrasal. They have performed experiments using English-Hindi parallel corpus consisting of Indian name entities. The overall performance accuracy of proposed system is quite good.

Joshi et al. (2013) have proposed system that can do transliteration from Roman script to Devanagari script. They have used the syllabification approach and considered the most probable term in the transliteration process. They have used backward transliteration that involved transliteration from Roman script to Devanagari script. Statistical machine learning approach was used for transliteration while TF-IDF model was used for Information retrieval. The syllable theory was used for transliteration. In the query labeling subtask, identification of English and Hindi words was performed using a hybrid approach that involved morphological analysis of English words and a corpus based approach to identify frequently occurring Hindi words.

*(iii) English to Punjabi transliteration*

Kaur and Josan (2011) have proposed a system that addresses the issue of statistical machine transliteration from English to Punjabi. Statistical Approach to transliteration is used for transliteration from English to Punjabi using MOSES that is a statistical machine transliteration tool. The system is improved by applying some transliteration rules at post processing stage. After applying transliteration rules average accuracy of this transliteration system comes out to be 63.31%.

Bhalla et al. (2013) have proposed rule based transliteration scheme for English to Punjabi. Some rules have constructed for syllabification. Syllabification is the process to extract or separate the syllable from the words. In this probabilities are calculated for name entities (proper names and location). For those words which do not come under the category of name entities, separate probabilities are being calculated by using relative frequency through a statistical machine translation toolkit known as MOSES. Using these probabilities the transliterating of input text from English to Punjabi is done. They have performed their experiment using Statistical Machine Translation tool. The average transliteration accuracy of 88.19% has been achieved.

*(iv) Shahmukhi to Gurmukhi transliteration*

Malik (2006) has developed Punjabi Machine Transliteration System (PMT) that is used to transliterate words from Shahmukhi script to Gurmukhi script. The Punjabi Machine Transliteration System uses transliteration rules (character mappings and dependency rules) for transliteration of Shahmukhi words into Gurmukhi. First basic character mapping is applied. Character mappings alone are not sufficient for PMT. There is need of certain dependency or contextual rules for producing correct transliteration. The basic idea behind these rules is the same as that of the character mappings. The PMT system can transliterate every word written in Shahmukhi. The PMT system gives more than 98% accuracy on classical literature and more than 99% accuracy on the modern literature.

Saini and Lehal (2008) have proposed a corpus based transliteration system for Punjabi language. The existence of two scripts for Punjabi language has created a script barrier between the Punjabi literature written in India and in Pakistan. The system has developed a new system for the first time of its kind for Shahmukhi script of Punjabi language. The proposed system for Shahmukhi to Gurmukhi transliteration has been implemented with various research techniques based on language corpus. The transliteration system was tested on a small set of poetry, article and story. The average transliteration accuracy of 91.37% has been obtained.

*(v) Marathi to English transliteration*

Dhore et al. (2012) have focused 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. Proposed phonetic based model transliterates Indian-origin named entities into English using full consonant approach and uses hybrid approach (that is rule based approach and metric based approach) stress analysis approach for schwa deletion. They have developed a rule-based phonetic model using Linguistic approach. Unicode encoded Hindi or Marathi named entity text input is given to the syllabification module which separates the input named entity into syllabic units. Transliteration module converts each syllabic unit in Devanagari into English by using phonetic map. Phonetic map is implemented by using the transliteration memory and mapping is done by writing the manual rules. The application can easily be ported on mobile devices because there is no need of bi-lingual and multilingual databases to be trained. Total 15,244 Named Entities are tested and the Top-1 accuracy of the proposed system is 74.14% and mismatch rate at Top-1 is 25.86%.

*(vi) Punjabi to Hindi transliteration*

Josan and Lehal (2010) have presented a novel approach to improve Punjabi to Hindi transliteration by combining a basic character to character mapping approach with rule based and Soundex based enhancements. On the basis of phonetic

sounds character mappings are determined. But character mapping alone is not sufficient for a Punjabi to Hindi machine transliteration system. Quite a reasonable improvement can be achieved by small amount of dependency or contextual rules. The rules are manually crafted. There are some characters in Hindi for which no rule applies. For generating these characters, the Soundex technique is employed. Soundex is a phonetic matching technique. Experimental results show that the approach followed effectively improves the word accuracy rate of various categories by a large margin.

Josan and Kaur (2011) have developed a statistical model that is used for transliterating the Punjabi text into Hindi text. They have described transliteration system build on statistical techniques. Their work presents empirical results for statistical Punjabi to Hindi transliteration system. Experimental results show that statistical approach effectively improves the transliteration accuracy rate and average levenshtein distance of the various categories by a large margin. The statistical method shows the improvements in performance by producing 87.72% accuracy rate.

Rani and Luxmi (2013) have proposed Direct Machine Translation System from Punjabi to Hindi for Newspapers headlines Domain. The similarity between Punjabi and Hindi languages is due to their parent language Sanskrit. Punjabi and Hindi are closely related languages with lots of similarities in syntax and vocabulary. Ambiguity is major problem in machine transliteration. Direct transliteration does not provide any solution for ambiguity. Another methodology or rules are developed for such types of problems. From the accuracy analysis total number of accurate sentence are calculated and it has given accuracy of 97%.

*(vii) English to Tamil transliteration*

Vijaya et al. (2009) have developed English to Tamil Transliteration system and named it WEKA. They have demonstrated a transliteration model based on multi class classification approach for English to Tamil transliteration. It is a Rule based system. During Transliteration phase, each word that is to be transliterated is first segmented into a sequence of English n-grams. The feature vector for each English n-gram is extracted. The trained model, corresponding to each n-gram in English word is used for prediction. The sequence of predicted class labels will form a transliterated word for the given English word. The accuracy of the model was tested with 1000 English names that were out of corpus. The accuracy can be increased by considering the next best output from the classifier. The transliteration model produced an exact transliteration in Tamil from English words with an accuracy of 84.82%.

*(viii) Hindi to Punjabi transliteration*

Goyal and Lehal (2009) have proposed system in which Hindi words are transliterated into Punjabi 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. The transliteration system is 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. In the post-processing phase bi-gram language model has been used. Developers claimed that the system is giving promising results and this can be further used by the researchers working on Hindi and Punjabi natural language processing tasks. The system will be openly available online for use.

*(ix) English to Kannada transliteration*

Antony et al. (2010) have developed English to Kannada transliteration system that is based on statistical method. Their work addresses the problem of transliterating English to Kannada language using a publically available transliteration tool that is known as Statistical Machine Translation (SMT). The purpose of statistical transliteration method is to find the transliteration of source language word into target language with a specific probability. The word in the target language with the highest probability is chosen that indicates the best transliteration. The tool named MOSES is used for English to Kannada transliteration. It is a complete statistical machine transliteration toolkit. The other open source tolls named SRILM and GIZA++ are used for creating language and transliteration model. The proposed system achieved exact Kannada transliteration for 89.27% of English names.

*(x) Bengali to English transliteration*

Mandal et al. (2007) have presented a cross-language retrieval system for the retrieval of English documents in response to queries in Bengali and Hindi. They have followed the dictionary based machine transliteration approach to result queries in English language. Transliteration is performed for good result. Phonetic transliteration is followed to generate equivalent English results. In order to achieve a successful retrieval a list-based named entity transliterations is adopted. The out-of-dictionary topic words were then transliterated into English using a phonetic transliteration system that works in the character level and converts every single Hindi or Bengali character in order to transliterate a word. The best values of Recall and MAP (Mean Average Precision) for the base run are 78.95% and 36.49%, respectively.

## IV. MACHINE TRANSLITERATION SYSTEMS FOR FOREIGN LANGUAGES

Significant work in the field of machine transliteration has been done for foreign languages. This section gives a brief description of various machine transliteration attempts for some foreign languages.

***(i)*** *English to Japanese transliteration*

Knight and Graehl (1998) have developed a phoneme based statistical model using finite state transducer that implements transformation rules to do back-transliteration. It is challenging to translate names and technical terms. These items are commonly transliterated that is replaced with approximate phonetic equivalents. For example, English word ‘computer’ is transliterated as ‘konpyuutaa’ in Japanese. Transliterating words from Japanese back to English is even more challenging. Knight and Graehl proposed method for automatic backward transliteration that can be used for transliteration of words from Japanese back to English. They

have mapped English sound sequences to Japanese sound sequences. A generative model of transliteration is built that involves sequence- An English phrase is written, Translator pronounces it in English, and the pronunciation is modified to fit the Japanese sound inventory. It port easily to new language pairs.

Yan et al. (2003) have proposed a method for automatically creating candidate Japanese transliterated versions of English words. They have used simple binary validation i.e. a large monolingual Japanese corpus. For given English word list, English word's phonetic representation is obtained. Then these English phonetic representations are mapped to Japanese phonetic sequences using a statistical based technique. Thus the Japanese phonetic sequences are mapped to katakana characters. These katakana sequences are all possible representations of the original English word.

Finch and Sumita (2009) have presented a phrased based transliteration technique for automatically transliterating between English and Japanese language pair. System is able to implicitly learn the correct character-sequence mappings through the process of character alignment. Proposed system is also able to reorder the translated character sequences in the output. The approach couches the problem of machine transliteration in terms of a character-level translation process. System generates phonetically correct transliterations around 80%.

***(ii)*** *Spanish to English transliteration*

Paul et al. (2009) have proposed Spanish-English statistical machine transliteration system. Source language input words that cannot be translated by the standard phrase based SMT models are either left un-translated or simply removed from the translation output. In their proposed work, they have applied a phrase based transliteration approach. The transliteration method is based directly on techniques developed for phrase based SMT and treats the task of transforming a character sequence from one language into another as a character-level translation process. The proposed transliteration model is applied as a post-process filter to the SMT decoding process, i.e. all source language words that could not be translated using the SMT engine are replaced with the corresponding transliterated word forms in order to obtain the final translation output. Experiments show that the incorporation of mixture models and phrase based transliteration techniques largely out-performed standard phrase based SMT engines gaining a total of 2.4% in BLEU and 2.1% in METEOR for the news domain.

***(iii)*** *English to Chinese transliteration*

Wan and Verspoor (1998) have described issues in the transliteration of proper names from English to Chinese. They have constructed a system for multilingual text generation supporting both languages. The proposed algorithm for mapping from English names to Chinese characters are based on heuristics about relationships between English spelling and pronunciation, and consistent relationships between English phonemes and Chinese characters. English to Chinese name transliteration occurs on the basis of pronunciation. That is, the written English word is mapped to the written Chinese character(s) via the spoken form associated with the word. The proposed algorithm is being implemented as a tool for the creation of Chinese lexical resources within a multilingual text generation project from an English language source database. Indeed, the result from the examples corresponds to a standard transliteration. Thus the algorithm produces results which are recognizable.

Lee and Chang (2003) proposed statistical machine transliteration based approach. For a given word in English the proposed method extracts the corresponding transliterated word from the aligned text in Chinese. First of all sentence alignment procedure is applied that align parallel texts at the sentence level. Then, a tagger is used to identify proper nouns in the source text. Some language dependent knowledge can be integrated to further improve the performance. They have presented unsupervised learning approach for machine transliteration. From the experimental results, it indicates that the proposed methods achieve excellent performance. Experimental results show that the average rates of word and character precision are 86.0% and 94.4%, respectively. The rates can be further improved with the addition of simple linguistic processing. The experiments can be extended to bi-directional transliteration and other different corpora.

***(iv)*** *Arabic to French transliteration*

Ben et al. (2011) have presented an approach for recognition and transliteration from Arabic into French of sports venues names. The proposed method integrates translation and transliteration together. The approach of recognition and transliteration is rule oriented. Based on rules, the implementation of the approach was performed using the NooJ platform. They focused on the transliteration of the proper names and the abbreviations and acronyms. Transliteration allows reducing the dictionary size of proper names. They have focused on sport domain. The corpus (i.e. learning corpus) made up of journalistic articles and lists of official sports venues names available on the Internet. This corpus is made of a hundred texts of which 200 are sports venues NE. The result of evaluation metric precision and recall is 69% and 67% respectively.

***(v)*** *Arabic to English transliteration*

Stalls and Knight (1998) have performed back-transliteration. They have built a model to transliterate names from Arabic into English. They applied probabilistic model that is based on the probabilities of particular word sequence. The proposed system's implementation is based on weighted finite state transducer. The resulting system contains all possible English transliterations, the best of which can be extracted by using graph based search algorithm.

Yaser and Knight (2002) have used a grapheme based approach that maps English letter sequences to Arabic letters. They have proposed a transliteration algorithm based on sound and spelling mappings using finite state machine. They have proposed a new spelling based model that is much more accurate than state-of-the-art phonetic based model. The proposed system is used to transliterate names from Arabic into English using probabilistic finite state machine that address transliteration of both Arab and foreign names into

English. They have evaluated transliteration algorithm using phonetic based and spelling based models. The proposed algorithm is most accurate for back transliterating English names.

Hermjakob et al. (2008) have presented a method to transliterate names in the framework of end-to-end statistical machine translation. The system is trained to learn when to transliterate for Arabic to English. They have divided the task of transliteration into two steps given an Arabic word or phrase to transliterate- identify a list of English transliteration candidates from indexed lists of English words and phrases with counts and then compute for each English name candidate the cost for the Arabic/English name pair. It shows that a state-of-the-art statistical machine translation system can benefit from a dedicated transliteration module to improve the translation of rare names.

***(vi)*** *English to Arabic transliteration*

AbdulJaleel and Larkey (2003) have developed a simple, statistical technique for building an English-Arabic transliteration model using Hidden Markov Model. The proposed n-gram transliteration model is a generative statistical model that produces a string of Arabic characters from a string of English characters. The model is trained from lists of proper name pairs in English and Arabic. They have demonstrated a simple technique for statistical transliteration that works well for cross-language IR, in terms of accuracy and retrieval effectiveness. Foreign words often occur in Arabic text as transliterations. They have proposed an approach to transliterate unknown words to deal with Out of vocabulary (OOV) words and called this a selected n-gram model. The proposed technique requires no heuristics or linguistic knowledge of either language. Good quality transliteration models can be generated automatically from reasonably small data sets.

***(vii)*** *Thai to English transliteration*

Khantonthong et al. (2000) have proposed a backward transliteration system from Thai into English. Thai text retrieval systems always involve documents that use loan words (borrowed from foreign language), especially in the area of science and engineering. Their work describes an algorithmic approach to backward machine transliteration aimed at improving the retrieval process. The approach consists of two main steps involving identifying loan words and back transliterating. Loan words, which are borrowed from foreign languages, are used in many languages such as Japanese, Chinese, Korean and Thai. These have effects on Thai Text Retrieval (TTR) system leading to inaccurate terms weight for indexing and text clustering. Therefore, there is a need to create automatic backward transliteration that can solve this problem. They have proposed hybrid model approach to an automatic backward transliteration system. The hybrid approach is the combination of a statistical model and a set of context sensitive rules. Proposed system has accuracy on document representation 70.41%.

***(viii)*** *Korean to English transliteration*

Jeong et al. (1999) have proposed a backward transliteration system that converts foreign words back to original words in English. Many foreign or English words appear in Korean text, especially in the area of science and engineering. Proposed system transliterates the Korean words to English words. They have proposed an algorithm that first identified the phrase containing foreign words and then extracts the foreign word part from the phrase. Statistical information is used for this purpose. Then they have presented a backward transliteration method that transliterates the foreign words to its English origin. After transliteration the generated English strings are probabilistic in nature. So in order to find correct English word among the candidates, the term matching technique i.e. HMM is used. The HMM based approach is implemented and the best word among the candidate words generated by HMM based algorithm is selected.

***(ix)*** *English to Korean transliteration*

Lee and Choi (1998) have presented a statistical transliteration model for the transliteration from English to Korean. They have compared two SMT based methods i.e. direct method and pivot method. After comparison they have proposed a hybrid method that is more effective for transliteration. In pivot method English words are converted into pronunciation symbols using SMT and then by using Korean standard conversion rules the pronunciation symbols are converted into Korean words. In case of direct method, English words are directly converted into Korean words. In this case intermediate step is not performed. They have used statistical transliteration model which can transliterate the words based on probability order. SMT is used as a base system for both direct and pivot method. SMT is a language-independent learning system that can learn rules automatically. This model is applied on English words and their pronunciation symbols pair and English word and Korean transliteration pair. At character accuracy level 86% of result is correct and at word accuracy level 53% of result is correct.

Kung and Choi (2000) have presented bi-directional and to some extent a language independent methodology for English-Korean transliteration and backward transliteration. The methodology is fully bi-directional, i.e. the same methodology is used for both transliteration and back transliteration. It consists of character alignment and decision tree learning. They have devised transliteration rules for each English alphabet and backward transliteration rules for each Korean alphabet. They have presented an automatic character alignment method for English word and Korean transliteration. The proposed alignment algorithm is capable of highly accurately aligning English word and Korean transliteration in a desired way. The method is highly accurate and more than 99% accuracy was obtained.

Oh and Choi (2002) have proposed English to Korean transliteration system using pronunciation and contextual rules. They have used phonetic information such as phoneme and its context as well as orthography. Phoneme-to-Korean (P-K) conversion method is based on English-to-Korean standard conversion rules. Previous works focused on an alphabet-to-alphabet mapping method. Because the transliteration is more

phonetic than orthographic, without phonetic information it may be difficult to acquire more relevant result. The proposed method shows significant performance increase by about 31% in word accuracy.

## V. CONCLUSION

A review of the different machine transliteration systems that have been developed for Indian languages as well as for foreign languages is presented in this paper. Various techniques for machine transliteration being used have also been described. Brief outline about the existing approaches those have been used to develop machine transliteration systems is summarized here. From this survey it is found that almost all existing Indian languages as well as foreign languages machine transliteration systems are based on statistical and hybrid approach.

## REFERENCES

[1] Deep, K. and Goyal, V. (2011), "Development of a Punjabi to English Transliteration System", International Journal of Computer Science and Communication, Vol. 2, No. 2, pp. 521-526.
[2] Kumar, P. and Kumar, V. (2013), "Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns", International Journal of Application or Innovation in Engineering & Management, Vol. 2, Issue 8, pp. 318-321.
[3] Rama T. and Gali K. (2009), "Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem", Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP, pp. 124-127.
[4] Das A., Ekbal A., Mandal T. and Bandyopadhyay S. (2009), "English to Hindi Machine Transliteration System at NEWS", Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. Association for Computational Linguistics, pp. 80-83.
[5] Sharma S., Bora N. and Halder M. (2012), "English-Hindi Transliteration using Statistical Machine Translation in different Notation", International Conference on Computing and Control Engineering.
[6] Joshi, H., Bhatt, A. and Patel. H. (2013), "Transliterated Search using Syllabification Approach", Forum for Information Retrieval Evaluation.
[7] Kaur, J. and Josan, G. (2011), "Statistical Approach to Transliteration from English to Punjabi", International Journal on Computer Science and Engineering, Vol. 3, No. 4, pp. 1518-1527.
[8] Bhalla, D. and Joshi, N. (2013), "Rule Based Transliteration Scheme For English To Punjabi", International Journal on Natural Language Computing, Vol. 2, No. 2, pp. 67-73.
[9] Malik, M. (2006), "Punjabi Machine Transliteration", Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 1137-1144.
[10] Saini, T. and Lehal, G. (2008), "Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach", Advances in Natural Language Processing and Applications Research in Computing Science, pp. 151-162.
[11] Dhore, M., Dixit, S. and Dhore, R. (2012), "Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis", Proceedings of COLING 2012: Demonstration Papers, pp. 111-118.
[12] Josan, G. and Lehal, G. (2010), "A Punjabi to Hindi Machine Transliteration System", Computational Linguistics and Chinese Language Processing, Vol. 15, No. 2, pp. 77-102.
[13] Josan, G. and Kaur, J. (2011), "Punjabi To Hindi Statistical Machine Transliteration", International Journal of Information Technology and Knowledge Management, Vol. 4, No. 2, pp. 459-463.
[14] Rani, S. and Luxmi, V. (2013), "Direct Machine Translation System from Punjabi to Hindi for Newspapers headlines Domain", International Journal Of Computers & Technology, Vol. 8, No. 3, pp. 908-912.
[15] Vijaya, M.S., Ajith, V.P., Shivapratap, G., and Soman, K.P. (2009), "English to Tamil Transliteration using WEKA", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 498-500.
[16] Goyal, V. and Lehal, G. (2009), "Evaluation of Hindi to Punjabi Machine Transliteration System", International Journal of Computer Science Issues, Vol. 4, No. 1, pp. 36-39.
[17] Antony, P., Ajith, V. and Soman, K. (2010), "Statistical Method for English to Kannada Transliteration", Communications in Computer and Information Science, Vol. 70, pp 356-362.
[18] Debasis Mandal, D., Dandapat, S., Gupta, M., Banerjee, P. and Sarkar, S. (2007),"Bengali and Hindi to English CLIR Evaluation", Cross-Language Evaluation Forum CLEF, pp. 95-102.
[19] Knight, K. and Graehl, J. (1998), "Machine transliteration", Proceedings of the 35th annual meetings of the Association for Computational Linguistics, pp. 128-135.
[20] Yan, Q., Grefenstette, G. and Evans, D. (2003), "Automatic transliteration for Japanese-to-English text retrieval", Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 353-360.
[21] Finch, A. and Sumita, E. (2009), "Phrase-based Machine Transliteration", Proceedings of the 2009 Named Entities Workshop, pp. 52-56.
[22] Paul, M., Finch, A. and Sumita, E. (2009), "Model Adaptation and Transliteration for Spanish-English SMT", Proceedings of the 4th EACL Workshop on Statistical Machine Translation, pp. 105-109.
[23] Wan, S. and Verspoor, C. (1998), "Automatic English-Chinese name transliteration for development of multilingual resources", Proceedings of the 17th international conference on Computational linguistics, Vol. 2, pp. 1352-1356.
[24] Lee, J. and Chang, S. (2003), "Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts using a Statistical Machine Transliteration Model", Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond, Vol. 3, pp. 96-103.
[25] Hamadou, A. and Piton, O. (2011), "Recognition and translation Arabic-French of Named Entities: case of the Sport places", International conference series Finite-State Methods and Natural Language Processing, pp. 134-142.
[26] Stalls, B. and Knight K. (1998), "Translating Names and Technical Terms in Arabic Text", COLING ACL Workshop on Computational Approaches to Semitic Languages, pp. 34-41.
[27] Yaser, O. and Knight, K. (2002), "Machine translation of names in Arabic text", Proceedings of the ACL conference workshop on computational approaches to Semitic languages.
[28] Hermjakob, U., Knight, K. and Daume, H. (2008), "Name Translation in Statistical Machine Translation Learning When to Transliterate", Proceedings of Association for Computational Linguistics, pp. 389–397.
[29] AbdulJaleel, N. and Larkey, L. (2003), "Statistical transliteration for English-Arabic cross language information retrieval", Proceedings of the 12th international conference on information and knowledge management, pp: 139-146.
[30] Khantonthon, N., Kawtraku, A. and Poovarawan, Y. (2000), "An Enhancement of Thai Text Retrieval Efficiency by Automatic Backward Transliteration", WAINS 7: E-Business for the new Millennium, Bangkok, Thailand.
[31] Jeong, K., Myaeng, S., Lee, J. and Choi, K. (1999), "Automatic identification and back transliteration of foreign words for information retrieval", In Proceedings of Information Processing and Management, Vol. 35, pp. 523-540.
[32] Lee, J. and Choi, K. (1998), "English to Korean Statistical transliteration for information retrieval", Computer Processing of Oriental Languages, pp. 17-37.
[33] Kang, B. and Choi, K. (2000), "Automatic Transliteration and Back-transliteration by Decision Tree Learning", In Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.
[34] Oh, J. and Choi, K. (2002), "An English-Korean Transliteration Model Using Pronunciation and Contextual Rules", Proceedings of the 19th international conference on Computational linguistics, Vol. 1, pp. 1-7.