ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Implementation of Punjabi to English Machine Transliteration System

Richa, Dr.Vishal Goyal

Assistant Professor, Dept of Computer Science, DAV College, Bathinda, India

Assistant Professor, Dept of Computer Science, Punjabi University, Patiala, India

ABSTRACT: In Natural Language Processing one of the important and relatively less mature areas is transliteration. Basically we can say transliteration is also a translation of Nouns, Out of the Vocabulary (OOVs) Words or Technical Jargons but technically it is entirely different from translation. Detailed transliteration can be used in many well researched areas of language processing efficiently and effectively such as machine translation, information extraction, CLIR and MLIR. To develop a transliteration system, model should be designed to preserve the phonetic sounds of the text in target language. We have developed a Punjabi to English statistical approach based transliteration system for proper nouns. It has been performing adequately well. The overall accuracy of the system measured is 95.69%.

KEYWORDS: Corpus, Segmentation, Unigrams, Bigrams, Mapping, Fusing.

I. INTRODUCTION

Transliteration system converts the source text i.e. text written in source language script into target text which is target language script without losing the phonetic sound of the source text. We tend to preserve the meaning of the text in target language in case of translation where in transliteration the focus is to maintain inflection. Transliteration can be performed in two directions, forward and backward. Transliterating text from its original script to foreign script is called forward transliteration although in case of backward transliteration text form foreign script is brought back to its original script. In this paper we are going to discuss statistical approach based Punjabi to English transliteration system which is for forward transliteration. Transliteration becomes a challenging process when source script and target scripts are entirely different. In case of Punjabi which is written in Gurmukhi Script and English which is generated from Roman Script we had to encounter difficulties to bridge the gap. We have performed text segmentation and then mapping using training data.

The organization of the paper is as follows: In Section II, we briefly describe the Related work on this topic on various Indian and foreign languages. In Section III, we discuss the methodology of our system in detail. In Section IV, we discuss the performance, results and evaluation of the system. Section V concludes the paper and next Section contains the references.

II. RELATED WORK

Most of the machine transliteration work outside India is carried out for English to Korean, English to Japanese, English to Chinese, English to Russian, English to Pinyin, Pinyin to Chinese, Thai to Japanese, English to Chinese, English to Russian, English to Pinyin, Pinyin to Chinese, Thai to English, Chinese to English, English to Arabic, Arabic to English, English to Thai, Urdu to English, Persian to English, Spanish to Chinese, Japanese to English Swedish to Finnish, English to Hebrew, English to Spanish and Spanish to English language pairs. For Indian languages English to Hindi, English to Tamil, Shahmukhi to Gurmukhi, English to Telugu, Bengali to English, English to Kannada, English to Oriya, Hindi to English, English to Punjabi, Punjabi to Hindi language pairs are used.

Phoneme-based model considers transliteration as a phonetic process. One of the early works on transliteration is done by Arababi in 1994 by combining neural net and expert systems [3]. The grapheme-based and phoneme-based models are used for the machine transliteration. The grapheme based model treats transliteration as an orthographic process and tries to map the source language graphemes directly to the target language graphemes. Conceptually, it is a direct orthographical mapping from source graphemes to target graphemes [7]. Ali and Ijaz developed English to Urdu Transliteration System based on the mapping rules in 2009. The whole process has three steps. First step is the mapping rules. Pronunciation based transliteration was chosen. English text is converted to Urdu using both English pronunciation and mapping rules. In Second step, Urdu syllabification has been applied. To improve system's accuracy, they have applied the Urduization Rules in third step. Overall system's accuracy is 96% [1]. Antony et al in 2010 has developed English to Kannada Transliteration. The whole model has three important phases. Preprocessing, Segmentation and Alignment. All the English words has been converted into lowercase and then corresponding Kannada words. After this both English place names and Romanized Kannada place names has been segmented. Alignment has been based on the number of transliteration units in the segmented English and Romanized Kannada place names. The corresponding transliteration units in English and Romanized Kannada words have been aligned. Second training phase has generated the transliteration model for English to Kannada names by using the SVM. Third transliteration phase has generated Kannada transliterations for a given English Name. The overall accuracy of system has been reported to be 87.28% [2]. Das et al in 2010 developed English to Hindi Machine Transliteration System. They have applied three different methods: Joint Source-Channel model, trigram model and Modified Joint Source-Channel model for English to Hindi transliteration. Three transliteration models can generate the Hindi transliteration from an English named entity (NE). They have used one standard run & two nonstandard run for transliteration. The word accuracy has been obtained is 0.471(for standard run), 0.389(for first nonstandard run), 0.384(for second nonstandard run) [4]. Kamaldeep, Goyal V developed the system *"Punjabi to English Transliteration System"* using a rule based approach and achieved accuracy of 93.23% in 2011. This system addresses the problem of forward transliteration of person names from Punjabi to English by set of character mapping rules. This system is accurate for the Punjabi words but not for the words with the foreign origin. System evaluated for names from the different domains like Person names, City names, State names, River names, etc [5]. Manikrao, Dixit and Sonwalker in 2012 developed a system for Hindi to English machine transliteration of Indian named entities such as proper nouns, place names and organization names using conditional random fields (CRF). The system is given input in the form of syllabification as it applies the n-gram techniques on the input. As more than 50% named entities are formed as a combination of two and three syllabic units. Corpus is trained using the n-gram approach with unigrams, bigrams and trigrams of Hindi. Accuracy of 85.79% for the bi-grams of source language Hindi is calculated. Results are totally dependent on corpus size as approach is based on statistical probability [6]. Lehal and Singh developed Shahmukhi to Gurmukhi Transliteration System based on Corpus approach in 2008. This system has been virtually divided into two phases. The first phase performs pre-processing and Rule-based transliteration tasks and the second phase perform 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 bi-gram manager will search bi-gram probabilities from bigram table for all possible bi-grams and then will add the corresponding bi-gram weights. Then it identifies and marks the best possible bi-gram and pops up the best possible unigram as output. This Gurmukhi token is then returned to the Output Text Generator for final output. 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 Gurmukhi text. The overall accuracy of system has been reported to be 91.37% [8]. Jong-Hoon Oh and Key-Sun Choi developed English-to-Korean and English-to-Japanese system by using a model based on grapheme and phoneme altogether in 2005. Model used the correspondence between source grapheme and source phoneme to produce the target language grapheme. It gives the performance improvement of 15%~23% for English- to-Korean transliteration and 15%~43% for English- to-Japanese transliteration. [9]. Surana and Singh in 2008 developed Transliteration system for English to Hindi and English to Telugu by using the DATM. They had used a sophisticated technique and machine learning on the source language (English) side, while a simple and light technique on the target (IL) side. Based on the class of the word, the possible pronunciations (for foreign words) and the possible segmentations (for Indian words) have generated. The Mean Recoprocal rank for English to Hindi is 0.87 and English to Telugu is 0.82 [10]. Verma in 2006 has developed Gurmukhi to Roman Transliteration System and named it GTrans. It is a rule based system. Most of the rules for transliteration in both schemes were same except for bindi (◌ਂ) and tippi (◌ਂ). This system has used diacritic marks to represent some special speech which cannot be produced with normal Roman ASCII characters. He has also done reverse transliteration from Gurumukhi to Roman. The overall accuracy of system has been reported to be 98.43% [11]. Wei, Xu Bo in 2008 has developed a Chinese to English Transliteration system based on the Weighted Finite- state Transducers (WFST). WFST provides a unified framework for integrating the various components of a speech-to-speech translation system. Firstly they converted Chinese characters to its pronunciation Pinyin, which is almost deterministic. Then, they have built different models (grapheme-based model, a phoneme-based model, an extended phoneme- based model) to solve different problems for translating Chinese pinyin to English. Then, they have combined these models with unified framework of WFST, built a combining transliteration model for Chinese-English. For testing, they have used CMU pronunciation dictionary (CMU), & LDC Chinese- English lists. The overall accuracy has been reported to be 63.25%. [12]Yaser, Knight in 2002 developed Arabic to English Transliteration system based on the sound & spelling mapping using finite state machine. They have combined the phonetic based model & spelling based model into the single transliteration model. For testing they have used the development data set & blind data set. The overall accuracy with development data set has been reported to be 53.66% and with blind data set it showed 61% [13].

III. DESIGN AND IMPLEMENTATION

The system architecture consists of various stages through which source language text has to be passed to be converted into target language text. Punjabi text is passed through three phases to complete the transliteration i.e. Preprocessing, Training and Tuning and Post processing. It can be better described as three tier architecture as shown in Figure 1.

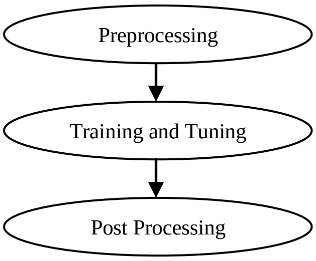

Figure 1. System architecture

a. PREPROCESSING

This phase includes the corpus collection and setting. We collected 30,860 names from various sources and after removing identical names we were left with around 26,500 names. Similarly bigrams were also collected or inserted manually in corpus, number of bigrams is 1214 till now but it can be further increased to improve the accuracy, moreover there are two possible matches added for bigrams which give us two unique outputs. Then unigrams 58 unigrams were added into corpus, it include all the 19 vowels, 38 consonants and sound of Punjabi language along with their corresponding match in English.

b. TRAINING AND TUNING

This phase acts as most important part of transliteration process. It takes the Punjabi text as input and then check for the match in corpus for full name. If match found then system gives the output and done but in case there is no match available then segmentation is performed on the text and bigrams i.e. two units of character set of language are generated. Then mapping takes place for bigrams. If match is found system checks for another possible match available, if any other possible match is available it works for two outputs otherwise it gives single output and if no match is available for bigram then bigram is further divided into unigrams i.e. single character of a character set and then mapping process repeats.

c. POST PROCESSING

Here some checks are performed on the source text i.e. Punjabi to check for the presence of addhak (ੱ). As we know the addhak (ੱ) has no sound equivalent to it in English so we have to apply rules for it instead of mapping. If addhak (ੱ) is detected following character is doubled. Then segments are fused together to produce the output.

IV. EVALUATION AND RESULTS

For evaluation we have taken two data set i.e. training data set and testing data set. Training data set was used to train the system, which include 26,500 full names, 1214 bigrams and 58 unigrams. Training data set helps the system to learn automatically. Whereas testing data set was used to test the system. Using testing data set the results of the system are compared to check the accuracy and evaluate the system. For evaluating the system we took names from the different domains like Person names, City names, State names and River names. To evaluate the performance of the system, transliteration accuracy rate is used. Accuracy Rate is the percentage of correct transliteration from the total generated transliterations by the system.

Accuracy Rate = (Correct Transliterations Generated / Total Transliterations Generated) * 100

The overall accuracy of our system is 95.69%. We have used two Test cases, one using both corpus names and segmentation process another using only segmentation process. Test cases are given below in Table 8 and its graphical representation is also described in Figure 4.

Table 1. Test cases

S.no.	Test Case	Accuracy Rate
1	Using Corpus Names and Segmentation Process	97.26%
2	Using Only Segmentation Process	94.12%

Our system provides two outputs to resolve the problem of multiple mapping. It improves the accuracy further more as user can choose among both options. Few such examples are given below.

Table 2. Examples

Punjabi Name	First Output	Second Output
ਵਸੀਮ	Waseem	Vasim
ਖੁਸ਼ਬੂ	Khushbu	Khushboo
ਇਜ਼ਰਾਇਲ	Israel	Izrael
ਫਿਲੀਪੀਨਸ	Philipins	Filipins
ਵੜਿੰਗ	Waring	Varing

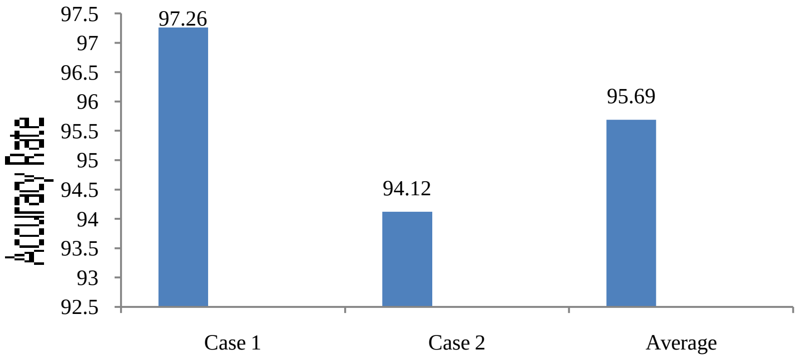

Figure 2. Graphical Representation of Accuracy Rate

V. CONCLUSION

In this paper, we presented machine transliteration of proper nouns for Punjabi-English language pair using statistical approach. Using segmentation and mapping to train the system, this approach is very good for the named entities of longer length. We have received very good accuracy 94.12% for the bi- grams and unigrams of source language Punjabi. As this approach is based on corpus, the results are always dependent on the training data size. This system can be used in various government offices where both the language Punjabi and English need to be used to maintain records. Our system can be further used by the researchers to improve it and produce multiple possible outputs for an input.

REFERENCES

1. Ali and Ijaz ,’English to Urdu Transliteration System’, Proceedings of the Conference on Language & Technology, pp. 15-23, 2009.
2. Antony, Ajith, Somam, ’Kernel Method for English to Kannada Transliteration’, International Conference on Recent Trends in Infirmation, telecommunication and computing, pp. 336-338, 2010.
3. Arbabi M et al,’ Algorithms for Arabic name transliteration’, IBM Journal of Research and Development, pp. 183-194, 1994
4. Das et al,’ English to Hindi Machine Transliteration System at NEWS 2009’, Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pp. 80–83, 2010.
5. Deep K, Goyal V,’ Hybrid Approach for Punjabi to English Transliteration System’, International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.
6. Dhore et al,’Hindi to English Machine Transliteration of Named Entities using Conditional Random Fields’, International Journal of Computer Applications (0975 – 8887) Volume 48– No.23, pp. 31-37, June 2012.
7. Karimi S, Scholer F and Turpin, ’Machine transliteration survey’, ACM Computing Surveys, Vol. 43, No.3, Article 17, pp.1-46, April 2011.
8. 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 147002, Punjab, India, pp. 151-162, 2008.
9. Oh Jong-Hoon and Choi Key-Sun, ‘An Ensemble of Grapheme and Phoneme for Machine Transliteration’, Springer-Verlag Berlin Heidelberg, IJCNLP 2005, LNAI 3651, pp. 450– 461, 2005.
10. Surana and Singh, ’A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages’, Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP-08), India. pp. 64-71, 2008.
11. Verma,’A Roman-Gurmukhi Transliteration system’, Proceeding of the Department of Computer Science, Punjabi University, Patiala, 2006.
12. Wei, Xu Bo, ’Chinese-English Transliteration Using Weighted Finite-State Transducer’, ICALIP, pp.1328-1333, 2008.
13. Yaser, Knight,’ Machine Transliteration of names in Arabic text’, Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, pp. 1-13, 2002.

BIOGRAPHY

Richa is currently working as an Assistant Professor in the department of Computer Science at D.A.V College, Bathinda. She holds degrees of B.Tech and M.Tech. Her research interests include natural language processing, artificial intelligence.

Dr. Vishal Goyal is currently working as Assistant Professor in the department of Computer Science at Punjabi University, Patiala in Punjab (INDIA). He received his MCA, M.tech(M.phil) and P.hd in computer sciences from Punjabi university, Rajasthan Vidyapeeth Deemed University, Udiapur and Punjabi university, Patiala respectively. He has more than 15 years of experience in teaching, research as well as in industry. He is associated with Punjabi university, Patiala since Feb, 2005 to present. He has published over 70 papers in referred journals and conferences (India and Abroad). He has published 4 books and 3 articles till now. He is member of various program committees for different International/National Conferences and is on the review board of various journals. He is editor-in-chief for IJES and editorial board member for JETWI, ‘Atti della Fondazione Giorgio Ronchi’ (Italy), JETCIS, IJCSI (France), ‘DAV Journal of Computing’. He has received ‘Young Scientist Award’ in 2005 and certificate of appreciation by JETWI. His biography has been published in Marquis Who’s Who in the World (USA) 2009 in Science and Engineering. His main areas of interests are: Natural language processing, database management systems, DNA computing, Theory of computation and psycho linguistic.