ISSN:0975-9646

Hybrid Approach for Hindi to English Transliteration System for Proper Nouns

Veerpal Kaur[#1], Amandeep kaur Sarao*[2], Jagtar Singh[#3]
[# 1]*M.Tech Student (CE), YCOE, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India.*
[*2]*Assistant Professor (CE), YCOE, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India.*
[#3]*Associate Professor (ECE), YCOE, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India.*

s

***Abstract*— In this paper hybrid approach is presented to transliterate proper nouns written in Hindi language into its equivalent English language. Hybrid approach means combination of direct mapping, rule based approach and statistical machine translation approach. Transliteration is a process to generate the words from the source language to the target language. The reverse process is known as backward transliteration. It shows that the performance is sufficiently high. This system can be used in various government organizations in India. Transliteration from Hindi language to English language plays a very important role as Hindi is official language of India and there is lot of data is present in Hindi which needs to convert into English for global usage.**

***Keywords*— Transliteration, Statistical Machine Translation, Devanagri Script, Machine Translation, Mapping**

1. INTRODUCTION

Machine transliteration System accepts characters of source language and map to the characters to the target language. The process is performed into two parts – Segmentation Phase , in which words of the source language are segmented into units and the – Assembly phase , in which segmented characters are mapped to the characters of target language with the help of rules. Transliteration and transcription are opposite to each other. Transcription is which maps the sounds of one language to script of another language. Transliteration is not translation. It‘s research area belongs to NLP (Natural Language Processing). Transliteration maps the letters of source script to letters of pronounced similarly in target script. Transliteration is particularly used to translate proper names and technical terms from languages. Machine transliteration is classified into two categories: Forward transliteration and backward transliteration. For example transliterating the name **"**गुरमन" to "Gurman" is known as forward transliteration while transliteration from "Gurman" to "गुरमन" is known as backward transliteration. But translation is the interpreting of the meaning of a text. For example इंसान is translation to Human.

II. RELATED WORK

Gurpreet Singh Josan et al. (2011) described a novel approach to improve Punjabi to Hindi transliteration System by using rule based approach. The accuracy of the proposed technique described in this paper varies from 73% to 85% which can be improved further by using some modified technique. [19] Haque et al. (2009) developed English to Hindi Transliteration system based on the phrase-based statistical method. PB-SMT models have been used for transliteration by translating characters rather than words as in character-level translation systems. They have modelled translation in PB-SMT as a decision process, in which the translation a source sentence is chosen to maximize. They used source context modelling into the state-of-the-art log-linear PB-SMT for the English—Hindi transliteration task. An improvement of 43.44% and 26.42% has been reported respectively for standard and larger datasets. [12] Jia et al. (2009) developed Noisy Channel Model for Grapheme-based Machine Transliteration. They have experimented this model on English-Chinese. Both English-Chinese forward transliteration and back transliteration has been studied. The process has been divided into four steps: language model building, transliteration model training, weight tuning, and decoding. In transliteration model training step, the alignment heuristic has been grown diag-final, while other parameters have default settings. When decoding, the parameter distortion-limit has been set to 0, meaning that no reordering operation is needed. [15] Kamal Deep Singh et al. (2011) developed hybrid approach based transliteration system of proper nouns written in Punjabi, the system produces its English transliteration. The performance of system is sufficiently high. The overall accuracy of system comes out to be 95.23%. [18] Lehal et al. (2010) developed Shahmukhi to Gurmukhi transliteration system based on corpus approach. In this system, first of all script mappings has been done in which mapping of Simple Consonants, Aspirated Consonants (AC), Vowels, other Diacritical Marks or Symbols are done. This system has been virtually divided into two phases. The first phase performs pre-processing and rule-based transliteration tasks and the second phase performs the task of post-processing. Bi-gram language model has been used in which the bi-gram queue of Gurmukhi tokens has been maintained with their respective unigram weights of occurrence. The Output Text Generator packs these tokens well with other input text which may include punctuation marks and embedded Roman text. The overall accuracy of system has been reported to be 91.37%. [11] Malik et al. (2009) developed Punjabi Machine Transliteration (PMT) system which is rule-based. PMT has been used for the Shahmukhi to Gurmukhi transliteration system. Firstly, two scripts have been discussed and compared. Based on this comparison and analysis, character mappings between Shahmukhi and Gurmukhi scripts have been drawn and

transliteration rules are formulated. The primary limitation of this system is that this system works only on input data which has been manually edited for missing vowels or diacritical marks which practically has limited use. [21] Sumita Rani et al. (2013**)** presented various techniques for transliteration from Punjabi language to Hindi Language. Most of the characters in Punjabi language have their same matching part present in a Hindi language. There are some characters exist in Hindi which is double sounds but no such characters are available for Punjabi. In this paper, transliteration system described is built on statistical techniques this system can be developed with minimum efforts. [25] Verma et al. (2006) developed a Roman-Gurmukhi transliteration System and named it GTrans. He has surveyed existing Roman-Indic script transliteration techniques and finally a transliteration scheme based on ISO: 15919 transliteration and ALA-LC has been developed. Because according to linguistics, these systems are closer to the natural pronunciation of Punjabi words as compared with others. Most of the rules for transliteration in both schemes were same except for Bindi and tippi in case of vowels as compared with consonants. He has also done reverse transliteration from Gurumukhi to Roman. The overall accuracy of system has been reported to be 98.43%. [31] Vijaya et al. (2009) developed English to Tamil transliteration system and named it WEKA. In this system, the valid target language n-gram (y_i) for a source language n-gram (xi) in the given source language input word is decided by considering the source language context features such as source language n-gram (xi) , two left context n-grams (xi-2, xi-1) and two right context n- grams (xi+1, xi+2). The transliteration process consisted of four phases: Pre- processing phase, feature extraction, training and transliteration phase .The accuracy of this system has been tested with 1000 English names. The transliteration model produced an exact transliteration in Tamil from English words with an accuracy of 84.82%. [29] Knight et al. (2005) presented English-Japanese Transliteration system. This system is a phoneme based as they converted English word to English sounds and then into Japanese sound. Japanese frequently imports vocabulary from other languages, primarily from English. It has a special phonetic alphabet called Katakana, which is used primarily to write down foreign names and loanwords. In process of transliteration, first step is to generate scored word sequences. The idea behind is that ice cream should score higher than ice crème, which should score higher than ace Kareem. In the second step, converts English word sequences into English sound sequences. [20].

III. OVERVIEW OF DEVANAGARI & ROMAN SCRIPT

In this section, we will study about Devanagri and Roman Script.

A. Devanagri Script

The script used for writing Hindi is called Devanagari. Devanagari evolved from the Brahmi script. The word Devanagari has been mystery to scholars, there is a hypothesis that it might be combination of two Sanskrit words 'Deva' (God, king) and 'Nagari' (city). Literally it combines to form 'City of Gods', 'Script of Gods'. Hindi uses only 34 consonantal syllables, 11 vowel letters, 9 vowel symbols, and 2 symbols for nasal sounds. The Devanagari script is an important and widely used script of India. It is mainly used to write Hindi, Marathi, Nepali and Sanskrit languages.

TABLE I
ALPHABET OF DEVANAGARI & ROMAN SCRIPTS
(CONSONANTS MAPPING)

क	k	ठ	Th	ब	b
ख	kh	ड	d	भ	bh
ग	g	ढ	dh	म	m
घ	gh	ण	n	य	y
ङ	n	त	t	र	r
च	ch	थ	th	ल	l
छ	chh	द	d	व	v,w
ज	j	ध	dh	श	sh
झ	jh	न	n	ष	sh
ञ	n	प	P	स	s
ट	T	फ	ph	ह	h

TABLE II
DEPENDENT VOWELS MAPPING

ा	aa	ृ	rri
ि	i	े	e
ी	ii	ै	ai
ु	u	ो	o
ू	oo	ौ	au

TABLE III
INDEPENDENT VOWELS MAPPING

अ	a	औ	au
आ	aa	ऋ	rri
इ	i	ॠ	
ई	ii	ऌ	l
उ	u	ॡ	l
ऊ	oo	अं	am
ए	e	अः	ah
ऐ	ai	अँ	n
ओ	o		

TABLE IV
SUPPLEMENTARY CONSONANTS MAPPING

क़	c,ch	ढ़	dh
ख़	kh	ऩ	n
ग़	g	फ़	Ph
ज़	z	य़	y
ड़	d	ऱ	rr

B. Roman Script

English Language is written in Roman script**.** In English Language 21 are consonants and 5 vowels.

Consonants are:-

B C D F G H J K L M N

P Q R S T V W X Y Z

Vowels are:-

A E I O U

IV. DESIGN AND IMPLEMENTATION

In the implementation, we use statistical machine translation approach. In this approach, systems try to transliterate new entries from the existing entries in the database. We use ASP.Net with C#.Net to implement the algorithm and MS-ACCESS to handle database contents. Hindi to English transliteration can achieve by using various techniques. In transliteration there are following techniques:

- Direct mapping
- Rule based approach
- Statistical machine translation (SMT) approach

A. Direct mapping Approach:-

When two languages are structurally similar and have similar vocabulary then direct approach is the best choice .Using direct approach system try to generate the result with the help of parallel corpus provided for training1) It generates only those results which are in the parallel corpus. It is the base of the transliteration process. It is also known as character to character mapping. The main advantage of direct mapping approach is that it consumes minimum time to transliterate the proper noun of Hindi language into its equivalent English language as transliteration involves only in searching the source keyword. The major disadvantage of this approach is that it can transliterate only those proper nouns which are present in the database. It cannot transliterate those nouns which are not present in the database.

B. Rule-based Approach:-

The rule-based approach is the first strategy that was developed. In this approach rules are created to perform the task of transliteration. Rules are created by considering the properties of the source and target language. Rules-based approaches take time, money and trained personnel to make and test the rules. The main advantage of rule based approach is that if rules are properly created according to the features of both source and target language then system can transliterate those nouns also which are not present in the database. The disadvantage of rule based approach for transliteration is very difficult to implement as there are very large number of rules with various exceptions are there in this approach. These rules are created by the human beings are tends to produce errors if they are not properly developed. Another disadvantage of rule based approach is that is works only on the Indian origin names but not on the foreign names.

C. Statistical machine translation approach:-

Statistical machine translation (SMT) is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge. It is language independent. SMT has high accuracy of results as compared to rule based approach. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model.

Our proposed system works in two phases. These two phases are: - System Training Phase and System Transliteration Phase.

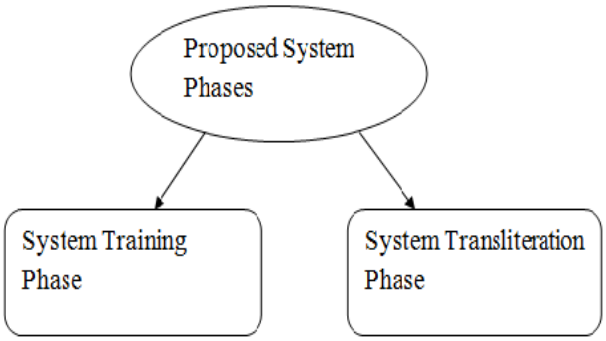

Fig. 1:- Figure of Proposed System

1) System Training phase: In System Training phase, training is given to the system on the basis of names stored into the database and it generates the database tables. Flowchart of System Training Phase is shown in following:-

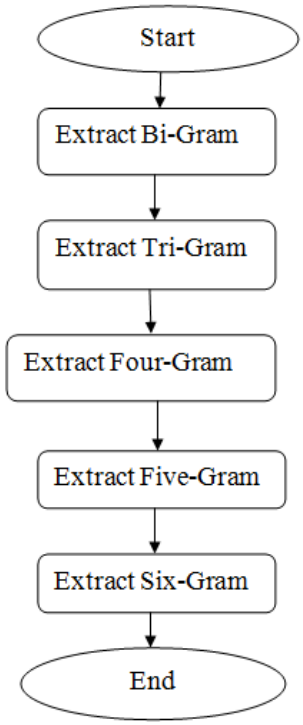

Fig. 2: - Flowchart of System Training Phase

In this Training Phase database tables are generated. Database tables which are bi – gram table, tri – gram table, four – gram table, five – gram table and six – gram table will be filled with the data generated automatically in this phase. Tables are stored into the database. Term Bi – Gram means combination of two characters of Hindi language word and their equivalent meaning in English Language. Bi-Gram table is shown as in the following:-

TABLE V
TABLE OF BI-GRAM

Hindi Gram	Eng Gram
शि	shi
मि	mi
िस	is
मि	mi
िन	in
दे	de
हव	hav
लि	Li

Term Tri – Gram means combination of three characters of Hindi language word and their equivalent meaning in English Language. Tri-Gram table is shown as in the following:-

TABLE VI
TABLE OF TRI-GRAM

Hindi Gram	Eng Gram
साल	Sal
लेद	Led
ेदा	Eda
गुण	Gun
बुर	Bur
अंध	And
धेर	Dher
कठि	Kathi

Term Four – Gram means combination of four characters of Hindi language word and their equivalent meaning in English Language. Four-Gram table is shown as in the following:-

TABLE VII
TABLE OF FOUR-GRAM

Hindi Gram	Eng Gram
रकाश	rakas
णधीर	nadhir
गराज	garaj
राजन	rajan
जेतु	jetu
खाकर	khakar
खदेव	khdev
वनिक	vanik

Term Five– Gram means combination of five characters of Hindi language word and their equivalent meaning in English Language. Five-Gram table is shown as in the following:-

TABLE VIII
TABLE OF FIVE-GRAM

Hindi Gram	Eng Gram
हारिश	haris
िवपुर	ivpur
पियूष	Piyus
पोइले	Poile
कटरमण	kataraman
चेरिल	cheril
रायणन	rayanan
पाटिल	Patil

Term Six – Gram means combination of six characters of Hindi language word and their equivalent meaning in English Language. Six-Gram table is shown as in the following:-

TABLE IX
TABLE OF SIX-GRAM

Hindi Gram	Eng Gram
पिनाकि	Pinakini
िनाकिन	inakin
ौदामिन	audamin
ुधामूर	udhamur
ुरपूजि	urapooji
ुवासिन	uvasin
्मलोचन	dmalochan
िशालाक	ishalak

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

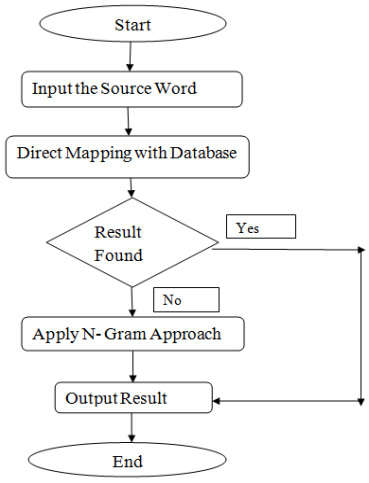

Fig. 3: - Flowchart of System Transliteration Phase

V. RESULTS

In this section, we will discuss about transliteration accuracy.

A. Statistics of work

We have more than 18000 entries in our database for Hindi to English proper nouns. And we have tested our software on various Hindi Proper nouns. Test cases developed and their results are shown in following sections. The system has been test thoroughly using test cases designed for number of various domains like proper names, City names, country names, river names, fruit names, color names, day names, literature, sports, and other subject's technical terms. Accuracy of system depends on data stored into the database.

System tests on more than 3000 names and system given as accuracy of 97%. System is also checked on those names which are not in the database of the system.

TABLE X
STATISTICS OF DATASET

Parameter	Frequency
Names Entity	18000+
N-Gram Extracted	90000+
Test	3000+
Result Accuracy	97%

B. Transliteration Accuracy

Accuracy Rate is the percentage of correct transliteration from the total generated transliterations by the system.

$$\text{Accuracy Rate} = \frac{\text{Number of Correct Transliteration}}{\text{Total no of Generated Transliteration}} * 100\%$$

TABLE XI
RESULTS OF OUR SYSTEM

Hindi Proper Noun	English Generated Name
दर्शन	Darshan
रीमा	Rima
वीरपाल	Veerpal
अमनदीप	Amandeep
जगतार	Jagtar
मनप्रीत	Manpreet
सरबजीत	Sarbjeet
अमनजोत	Amanjot
चितेश	Chitesh
Sरुब्बल	Rubbal
मनजीत	Manjeet
मनजोत	Manjot
मनमीत	Manmeet
खुशप्रीत	Khushpreet
आद्या	Aadya
आहना	Aahana
आलेयह	Aaleyah
अदरा	Adara
सोफिअ	Sophia
एम्मा	Emma
लिली	Lily
शेरगिल	Shergill
कामेश	Kamesh
जयपाल	Jaipal
गोपन	Gopan
गुरमन	Gurman
उपजीत	Upjeet
राज	Raj
गुरप्रीत	Gurpreet
नसीब	Nasib
सुखमन	Sukhman
आदेश	Aadesh
आदि	Aadi
अबनीत	Abneet
हसनदीप	Hasandeep
मंदीप	Mandeep
किरणदीप	Kirandeep

VI. CONCLUSION & FUTURE WORK

In this paper, a hybrid approach is developed to transliterate proper nouns of Hindi language into its English equivalent names. There is various machine transliteration models used for transliteration. After studying number of works done by various researches in the area, we have developed new algorithm based on statistical machine translation for transliteration from Hindi to English and the accuracy comes out to be approx. 97% and can achieved to 100% by improving database. Proper nouns from the State govt., Hindi Documents, Hindi Literature and other documents in Hindi can be transliterated into English for use on the click on a button. Size of the database can be increase considerably to obtain the good results. Now, as future work, database can be improved by including more names to improve the accuracy and increase the N-Gram Approach to Ten-Gram.

REFERENCES

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

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

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

[4] Abbas Malik, Laurent Besacier, Christian Boitet, Pushpak Bhattacharyya, "A Hybrid Model for Urdu Hindi Transliteration",

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP, 2009, Singapore. pp. 177–185.

[5] Das,Ekbal, Mandal and Bandyopadhyay, "English to Hindi Machine Transliteration System at NEWS 2009", *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009*, pages 80–83.

[6] Das, Saikh, Mondal, Ekbal, Bandyopa dhyay, "English to Indian Languages Machine Transliteration System at NEWS 2010", *Proceedings of the 2010 Named Entities Workshop*, pages 71–75,Uppsala, Sweden, 16 July 2010.

[7] Deselaers, Hasan, Bender and Ney "Deep Learning Approach to Machine Transliteration",*Proceedings of the Fourth Workshop on Statistical Machine Translation* , pages 233–241, Athens, Greece, 30 March – 31 March 2009.

[8] Ekbal Asif, Sudip Kumar Naskar and Sivaji Bandyopadhyay, "A Modified Joint Source-Channel Model for Transliteration", *Proceedings of ACL 2006*, pp 191-198, 2006.

[9] Josan, G.S, "Punjabi to Hindi Statistical Machine Transliteration", *International Journal of Information Technology and Knowledge Management*, vol 4, No. 2, pp. 459-463.

[10] Gurpreet Singh Lehal and Tejinder Singh Saini, "Conversion between Scripts of Punjabi:Beyond Simple Transliteration", *Proceedings of the coling 2012: Posters, Mumb*ai, pp.633-642.

[11] Gurpreet Singh Lehal and Tejinder Singh Saini, "Development of a Complete Urdu-Hindi Transliteration System", *Proceedings of the COLING 2012: Posters, Mumbai*, pp. 643-652.

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

[13] Hoonoh, Isahara & Sun, "A Comparison of Different Machine Transliteration Models", *proceeding of AI access foundation*, 2006, vol 27, pp: 119–151.

[14] 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, pp 1518-1527.

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

[16] Jasleen Kaur, "Machine Transliteration system in Indian perspectives", *International Journal of Science, Engineering and Technology Research (IJSETR*), Vol 2, Issue 5, May 2013

[17] Knight K. and J. Graehl, "Machine Transliteration", *Computational Linguistics*, 24(4): pp 599-612, 1998.

[18] Kamal Deep, Dr.Vishal Goyal, 2011, "Development of a Punjabi to English Transliteration System", *International Journal of Computer Science and Communication*, Vol. 2, No. 2, July-December 2011, pp. 521-526.

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

[20] Knight, Graehl, "English-Japanese Transliteration system", Computational Linguistics, Volume 24, Number 4, pp.599-612.

[21] Malik, "Punjabi Machine Transliteration System", In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (2006), pp. 1137-1144.

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

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

[24] Pankaj Kumar and Er.Vinod Kumar, "Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns" *International Journal of Application or Innovation in Engineering & Management*, Volume 2, Issue 8, August 2013, ISSN 2319-484

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

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

[27] "Transliteration Principles" Internet Source:-http://acharya. iitm.ac.in/ multi_sys /transli/translit.htm acessed on 2-12-2010

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

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

[30] V Goyal and G S Lehal, "A Machine Transliteration System for Machine Translation System: An Application on Hindi-Punjabi Language Pair", Atti Della Fondazione Giorgio Ronchi (Italy), Volume LXIV, No. 1, pp. 27-35.

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

[32] Wei, Xu Bo, "Chinese-English Transliteration Using Weighted Finite-state Transducers", 2008, pp- 1328 – 1333.

[33] Yaser ,Knight, "Machine Transliteration of names in Arabic text", Machine transliteration of names in Arabic text In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA.,2002, pp: 1-13