GRT: Gurmukhi to Roman Transliteration System using Character Mapping and Handcrafted Rules

Shailendra Kumar Singh, Manoj Kumar Sachan

Abstract*: *In the last two decades, the transliteration system has got significant research attention. It is observed that Punjabi to English transliteration for all type of part-of-speech words is comparably less studied. Currently, some research work in this area is carried out but only for proper nouns and some technical terms. So, there is need to focus on all type of words. The Gurmukhi to Roman transliteration (GRT) system is the first proposed system for transliteration of all kind of part-of-speech words. This system uses the handcrafted-rules and character mapping (CM) approach for transliteration between languages involved. The CM is done for Gurmukhi script with its equivalent to Roman script. It transliterates text written in Punjabi language into English language. It is tested on 65,130 Punjabi words and achieved accuracy of 99.27%, which is better than other state-of-art system results. The developed system can be used in social media text normalization, translation, sentiment analysis of multilingual text, text summarization of multilingual text, etc.

Index Terms*: *Transliteration, Gurmukhi script, Roman script, Character mapping, Handcrafted rule.

I. INTRODUCTION

The youth aged between 15-24 are using internet such as 94% in developed, 67% in developing and 30% in least developed countries. Out of 48% of the total world population using internet, 70% of the total youngster aged between 15-24 years age group are among internet users [1]. Nowadays, social media has become the part of people's daily life. The content generated on social media has exponentially increased [2] and contain multilingual text [3]. The content written on Facebook , more than 33% comments are written using phonetic text and more than 38% comments are written using code mixed phonetic text (bilingual) [4]. These text do not follow any standard spelling rules, but are based on the pronunciation of the words [5]. So, the development of phonetic dataset(s) is required for the text mining, opinion mining, information retrieval, feedback analysis, business intelligence, data analytics, etc. Transliteration is a process in which mapping is done for a word written in a language-script to another language-script. For example, Punjabi words ਘਰ and ਕਲਮ are equivalent to English words ghar and kalam, respectively [6]. Transliterating a word from its origin language to another (target) language is called as *forward transliteration*, but when transliterating a word written in another language back to its original language is known as *backward transliteration* [6]. The transliteration of a word from source language to target language with same sound of pronunciation as a source word sound is called as *phonetical translation (transliteration)*. The transliterated words in the target language are known as out-of-vocabulary or sometimes out-of-dictionary word. A general approach for transliteration is to create character (alphabets) sequence mapping rules or set of CM between both languages involved. The translation system translates source text to target text with same meaning but may have different pronunciation of words in both languages and this translation system was first time attempted during the World War II. The automatic translation system for almost all languages has been developed after two–three decade of that war [7]. The problem occurred in translation system, when word appeared as proper nouns or technical words, so to tackle this problem transliteration system was developed and considered to be part of translation system [8].

In recent years, script transliteration has been considered as significant research in area of machine translation [9]. There are three main approaches which are used for script transliteration such as grapheme, phoneme and hybrid-based [9], [10]. A grapheme is the basic unit of written language, which includes characters (alphabets), punctuation marks, numerals and all the individual symbols of that writing language system. For example, the word "she" contains three graphemes (i.e. 's', 'h' and 'e') [11]. In grapheme-based approaches, transliteration is seen as a process in which mapping of grapheme (character) sequence from a source language to target language without considering the phoneme-level processes. The grapheme-based models directly transliterate from source language graphemes to target language graphemes. The grapheme-based approaches are classified as rule-based, statistical machine transliteration, hidden markov model, finite state transducer based models, etc. [9]. A phoneme is the smallest unit of speech that distinguishes meanings. For example, if the sound [p] substituted with sound [b] in the word "pig", the word changes to "big". Hence here /p/ is a phoneme. In phoneme-based approaches, the mapping from source

Revised Manuscript Received on July 10, 2019.

Shailendra Kumar Singh*, Department of Computer science & Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Sangrur, Punjab, India.

Manoj Kumar Sachan, Department of Computer science & Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Sangrur, Punjab, India.

*Corresponding author
Email: sks.it2012@gmail.com

language graphemes to source language phoneme(s), source language phoneme(s) to target language phoneme(s), target language phoneme(s) to target language graphemes is done. In hybrid approaches, the both grapheme and phoneme models are combined [6]. Abbas Malik [12], the author has developed a "Punjabi Machine Transliteration" system using CM and dependency rules for transliteration of Shahmukhi words into Gurmukhi words. The author has used 45,420 words for testing of proposed system and claimed an overall accuracy of 98.95%. Komal Deep and Vishal Goyal [13], developed a transliteration system which translates proper names and technical terms from Punjabi to English language. This system is based on the grapheme method. They used direct mapping of vowels and consonant along with some developed different rules. The system is tested on two datasets; first on person names with accuracy of 95% and second dataset includes city names, state names, river names with accuracy of 91.40%. The overall accuracy of the system is 93.22%. Pankaj Kumar and Vinod Kumar [14], they have developed statistical machine transliteration system to transliterate proper nouns of Punjabi language into its equivalent English language. They have tested their system on 2000 names with accuracy of 97%. Manpreet Kaur [15], presented machine transliteration system for transliteration of Punjabi proper nouns into English language. The author has tested her proposed system on 5000 Punjabi names and claimed accuracy of 95.69%. There are two main technique used for language transliteration such as statistical-based and rule-base transliteration. As per research articles [16]–[21] and [6], statistical techniques work well for the resource rich languages while rule-base transliteration work better for the resource scarce language. So, the proposed system in this article based on handcrafted non-probabilistic character mapping rule-based transliteration from Gurmukhi script to Roman script. All characters of Gurmukhi script are mapped with its equivalent characters of Roman script. The proposed system is based on grapheme method.

Contribution & Paper Organization

1. The proposed system transliterates source text from Gurmukhi script to Roman (English/Phonetic) script, which helps non-Punjabi readers to understand the existing content written in Punjabi.
2. The Phonetic dataset(s) generated by this system, will be used by text mining, opinion mining, information retrieval and data analytics systems. Also, phonetic dataset(s) can be used to develop the machine learning system for transliteration.
3. The designed model using character mapping and rule-based approach can be used to develop dataset(s) for resource-scarce languages using transliteration.

This article focuses on the framework and approach of the proposed transliteration system, which transliterates Gurmukhi script text into Roman script text. There are some handcrafted rules are made and applied along with the CM of Gurmukhi alphabets with its equivalent to Roman script alphabets. This system achieved an accuracy of 99.27% on 65,130 Punjabi words and perform better than the Online Google Transliteration System. The organization of this article is as follows: section 2 describe the methodology of the proposed system; section 3 explains the framework and handcrafted rules; section 4 presents the experimental results and discussion; and section 5 includes conclusion and future scope.

II. METHODOLOGY

The proposed system used the handcrafted-rules and CM approach for transliteration between languages involved. The character mapping is done for Gurmukhi script characters with its equivalent to Roman script characters. This GRT system transliterate text written in Punjabi language into English language. For example- "ਆਮ" word of Punjabi language transliterated in English language as "aam".

A. *English Language (Roman Script)*

Roman script is used to write English language. It is a globally spoken and written language. In India, English is used as second language. There are 26 letters in English, out of which 5 are vowels and 21 are consonants as shown in table 1.

Table 1 Roman script letters

Vowels	A	E	I	O	U
Consonants	B	C	D	F	G
	H	J	K	L	M
	N	P	Q	R	S
	T	V	W	X	Y
	Z				

B. *Punjabi Language (Gurmukhi Script)*

Gurmukhi script is used to write the Punjabi language. There are 10 independent vowels (see table-2 (a)) and 9 dependent vowels (see table-2(b)). Those vowels are used alone are called independent vowels while dependent vowels depend on the consonants and used along with consonants.

Table 2 Gurmukhi script vowels (a) Independent (b) Dependent

(a) Independent Vowels				
ਅ	ਆ	ਇ	ਈ	ਉ
ਊ	ਏ	ਐ	ਓ	ਅੌ
(b) Dependent Vowels				
◌ਾ	◌ਿ	◌ੀ	◌ੁ	◌ੂ
◌ੇ	◌ੈ	◌ੋ	◌ੌ	

There are 32 distinct letters (consonants) as shown in table-3 and in addition to these , there are 6 consonants created by placing a dot (bindi) at the foot (pair) of the consonant as shown in table-4 [13].

Table 3 Gurmukhi script consonants

ਕ	ਖ	ਗ	ਘ	ਙ	ਚ	ਛ	ਜ
ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ	ਤ
ਥ	ਦ	ਧ	ਨ	ਪ	ਫ	ਬ	ਭ
ਮ	ਯ	ਰ	ਲ	ਵ	ੜ	ਸ	ਹ

Table 4 Gurmukhi script consonant placing dot at the foot

ਸ਼	ਡ਼	ਗ਼	ਜ਼	ਖ਼	ਫ਼	ਲ਼

C. *Character Mapping*

In character mapping process, one to one mapping

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

is used. The one to one mapping means, each character of Gurmukhi script is mapped equivalent to one unit of character(s) of English. Example: - "ੀ" equivalent to "ii"; "ਔ" equivalent to "au". But in article [13], the authors have used one to many mappings for some characters e.g. "ੀ" equivalent to "i" and "ee"; "ਵ" equivalent to "v" and "w". The mapping of one character with equivalent to many character(s) unit confuse the readers and need more rules to overcome this type of ambiguities. The CM of Gurmukhi script is shown in table-5 for vowels, table-6 for consonants and table -7 for special symbols.

Table 5 Gurmukhi script vowels character mapping of

Gurmukhi Character	ਅ	ਆ	ਇ	ਈ	ਉ	ਊ	ਏ	ਐ	ਓ	ਔ
		◌ਾ	◌ਿ	◌ੀ	◌ੁ	◌ੂ	◌ੇ	◌ੈ	◌ੋ	◌ੌ
Roman Character	a	aa	i	ii	u	uu	e	ai	o	au

Table 6 Gurmukhi script consonants character mapping

Gurmukhi Character	ਕ	ਖ	ਗ	ਘ	ਙ	ਚ	ਛ	ਜ	ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ	ਤ	ਥ	ਦ	ਧ	ਨ
Roman Character	k	kh	g	gh	ng	ch	chh	j	jh	yan	t	th	d	dh	n	t	th	d	dh	n
Gurmukhi Character	ਪ	ਫ	ਬ	ਭ	ਮ	ਯ	ਰ	ਲ	ਵ	ੜ	ਸ	ਹ	ਸ਼	ਫ਼	ਗ਼	ਜ਼	ਖ਼	ਫ਼	ਲ਼	
Roman Character	p	ph	b	bh	m	y	r	l	v	r	s	h	sh	dh	ghh	z	khh	f	ll	

Table 7 Mapping of special symbols

Gurmukhi Character	Mapping
◌ਂ	n
◌ੰ	n
◌ੱ	Doubles the following character

III. PROPOSED FRAMEWORK OF TRANSLITERATION SYSTEM

In this article, a hybrid approach is used based on both (i) rule-based and (ii) character mapping approach. The GRT system is designed, consisting of (a) segmentation (tokenization), (b) transliteration generation, and (c) target word generation as shown in fig. 1.

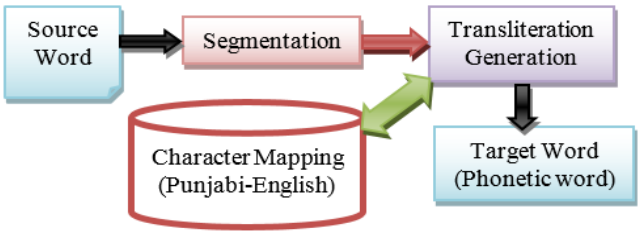

Figure 1 Framework of Gurmukhi to Roman Transliteration System

In segmentation, system tokenizes word into its constituent characters, then pass all the constituent characters to the next step. For example- "ਆਮ" word tokenize as "ਅ + ◌ਾ + ਮ". Now in transliteration generation step, system will generate equivalent character(s) unit in Roman script for Gurmukhis' character, which are constituents of the word. For example- "ਅ => a, ◌ਾ=> aa, ਮ=> ma". In final step, all transliterated character(s) unit are combined to generate target (phonetic) word based on some handcrafted rules as shown in fig. 1. Example- "a + aa + ma" apply some rule and generate as "aam" as target word.

A. *Dataset*

The Punjabi language word written in Gurmukhi script is considered as the input of the GRT system. The 65,130 most frequently used words of Punjabi language has been used for testing. This dataset is subset of 1,70,714 Punjabi words, which is discussed on website [1] and provided by Prof. G.S.Lehal[2].

B. *Handcrafted Rules*

The following handcrafted rules are made for the implementation of the proposed GRT system.

Rule 1 (R1): *Character (consonant) is last character of the word*

Each Gurmukhi consonant symbols is represented by consonant plus an inherent schwa vowel sound "ਅ" [13]. For example, ਕਮਲ=ਕ + ਅ+ ਮ+ ਅ + ਲ+ ਅ segmentation of word. If character is last character of the word, then map character without schwa as "kamal" in phonetic.

Rule 2 (R2): *Character (consonant) is followed by dependent vowel*

Each consonant written in Gurmukhi is followed by schwa vowel sound character i.e ਅ; but when consonant is followed by dependent vowel then no need to add schwa. Example: - ਕਿਤਾਬ = ਕ + ◌ਿ + ਤ + ◌ਾ + ਬ, and written in phonetic as "kitaab", here "ਕ" is followed by dependent vowel "◌ਿ" and "ਤ" is followed by dependent vowel "◌ਾ" so, there is no need to add schwa sound symbol with these consonants.

Rule 3 (R3): *Character is followed by consonant or independent vowel*

Each consonant written in Gurmukhi is followed by schwa vowel sound character i.e ਅ. Example: - In word "ਕਮਲ" charcater "ਕ" is followed by charcater "ਮ" then character "ਕ" is mapped with schwa sound symbol as "ka"; and the

[1] http://www.learnpunjabi.org/statistics.html
[2] Director, Research Centre for Punjabi Language Technology, Punjabi University, Patiala, Punjab, India

word is mapped as “kamal” in phonetic. In word “ਕਈ”, character “ਕ” is followed by vowel “ਈ”, so character “ਕ” is mapped with schwa as “ka” and “ਈ” as “ii”; and word “ਕਈ” mapped as “kaii” in phonetic.

Rule 4 (R4): - *“ਆ or ◌ਾ” is last character of word*

If vowel “ਆ or ◌ਾ” is last character of word, then map as “a” otherwise map as “aa”. For example, in word “ਪੈਦਾ”, “◌ਾ” dependent vowel is last character of word, so this word is mapped as “paida”; while in word “ਵਿਆਹ”, the vowel character “ਆ” is followed by “ਹ”, so this word is mapped as “viaah” in phonetic.

Rule 5 (R5): - *Character is “half character”*

In this case, character is basically followed by symbol “◌੍”, thus the character is treated as half character. The character is mapped without schwa as like Rule-2. E.g. word “ਪ੍ਰਾਪਤ” is segmented as “ਪ+◌੍+ਰ+◌ਾ+ਪ+ਤ” and mapped as “praapat” in phonetic.

IV. RESULT AND DISCUSSION

The proposed system directly transliterates the words written in Gurmukhi script into Roman script using CM along with the handcrafted rules. The transliteration of this system depends upon the constituents of words from different characters or symbols and implemented using Python 3. For example, the word ‘ਦਿੱਲੀ’ is transliterated as ‘dillii’, but it translated as ‘Delhi’. So, there is vast difference between transliteration and translation. This system is tested on 65,130 words including different part-of-speech words. The system correctly transliterated 64,657 words out of 65,130 Punjabi words.

To evaluate the performance of the GRT system, the handcrafted rules (as discussed in section 3) along with CM are categorized into six cases as given below:

CASE-1: CM + Rules (R2+R3+R4+R5)
CASE-2: CM + Rules (R1+R3+R4+R5)
CASE-3: CM + Rules (R1+R2+R4+R5)
CASE-4: CM + Rules (R1+R2+R3+R5)
CASE-5: CM + Rules (R1+R2+R3+R4)
CASE-6: CM + Rules (R1+R2+R3+R4+R5)

The performance of system is measured using accuracy parameter. The accuracy is defined as the ratio between the number of correctly transliterated words and total number of the words as shown in equation (1). The accuracy of the system varies depending upon the cases as shown in fig. 2. The accuracy of the system in case-6 is 99.27%, which is best among all the cases. The handcrafted rule-2 is excluded from case-2, then the accuracy of the system is decreased from 99.27% to 8.46%, so rule-2 is significant for the system. In case 2, 59621 words are incorrectly transliterated because these words consist dependent vowels. The rule-2 only can handle those words which consist dependent vowels. Thus, each rule has its own importance as the accuracy shows in fig. 2.

$$Accuracy = \frac{Correctly\ transliterated\ words}{Total\ words} * 100 \quad (1)$$

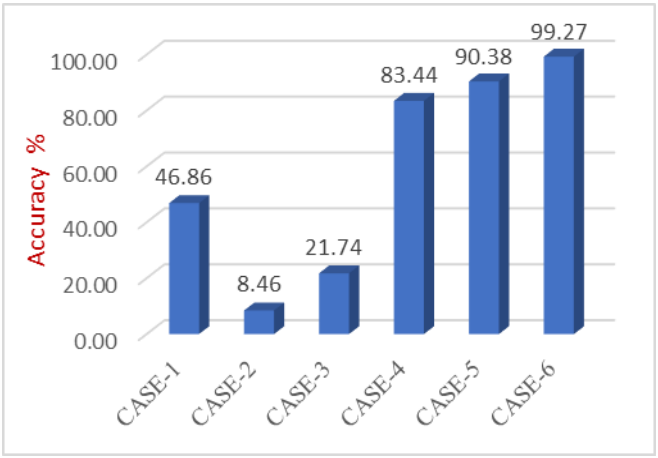

Figure 2 Accuracy of the system vary with different rules

A. *Comparison with Online Google Transliteration System*

This GRT system works better than the “OGTS[3]”. Table-8 shows transliterated words by both (i) OGTS and (ii) proposed GRT system.

- The special symbols (i.e. ‘◌ੰ’ and ‘◌ੱ’) of Gurmukhi script are recognized by the proposed system, but OGTS does not recognized. For example, the word ‘ਪਸੰਦ’ and ‘ਇੱਟਾ’ transliterated as ‘pasada’ and ‘itaa’ by OGTS, while transliterated as ‘pasand’ and ‘itta’ by proposed system as shown in table 8.
- The dependent vowel ‘◌ਾ’ transliterated as ‘aa’ on every position of the word by OGTS as in both word ‘neetaa’ and ‘saadi’, but the proposed system ‘◌ਾ’ transliterated as ‘aa’ at every position in the word except at last position as in word ‘saadi’ and ‘neta’ as shown in table-8.
- The OGTS use ‘sachwa’ even at last position in the word, as in word ‘preema’; but proposed system does not use ‘sachwa’ at end position in the word, as in word ‘prem’ as shown in table 8.

Table 8 Gurmukhi script words transliteration by OGTS and Proposed system

Gurmukhi script	Roman script	
	OGTS	Proposed GRT System
ਪਸੰਦ	pasada	pasand
ਇੱਟਾ	itaa	itta
ਨੇਤਾ	neetaa	neta
ਸਾਦਿ	saadi	saadi
ਪ੍ਰੇਮ	preema	prem

B. *Result comparison with state-of-art technique*

The transliteration system was developed for the transliteration of name entity and some technical terms [13]–[15]. But now, it is used (i) corpus / data acquisition for resource-scarce language,

[3] https://translate.google.com/?rlz=1C1NHXL_enIN727IN727&um=1&ie=UTF-8&hl=en&client=tw-ob#auto/hi/transliteration

(ii) to remove communication barrier for non-native language reader, (iii) for different applications such as information retrieval, text summarization, opinion mining, etc.

Table 1 State-of-art comparison with proposed system

Authors/References	Method/Approach	Language	Data size	Accuracy
Abbas Malik [12]	CM & Dependency Rules	Shahmukhi-Gurmukhi	45,420 words	98.95%
Komal Deep and Vishal Goyal [13]	Character Sequence Mapping Rules	Punjabi-English	2,046 names	93.22%
Pankaj Kumar and Vinod Kumar [14]	Statistical machine	Punjabi-English	2000 names	97%
Manpreet Kaur [15]	Parallel corpus	Punjabi-English	5,000 names	95.69%
Proposed system	*CM & Handcrafted Rules*	*Punjabi-English*	*65,130 words*	***99.27%***

As shown in table-9, Abbas Malik [12] used CM and dependency rules to transliterate Shahmukhi script words into Gurmukhi script words; Komal Deep and Vishal Goyal [13] used character sequence mapping rules, Pankaj Kumar and Vinod Kumar [14] used statistical machine, Manpreet Kaur [15] used parallel corpus for transliteration of Punjabi names into English names. The proposed system used CM and handcrafted rules for transliteration of Punjabi words into English language words and achieved best accuracy of 99.27% among all existing results as shown in table 9.

C. *Error Analysis*

1. **Difference in number of characters:** There is difference in number of character sets for both Gurmukhi and Roman. There are 5 vowels and 21 consonants in English while in Punjabi language 19 vowels as shown in table-2, 39 consonants as shown in table-6 and 3 special symbols as shown in table-7. So, the English language alphabets have multi mapping as shown in table-6 & 7. For example: the Punjabi characters 'ਣ', 'ਨ', 'ਂ' and 'ੰ' are transliterated as 'n'; 'ਟ', 'ਤ' are transliterated as 't'; 'ਠ', 'ਥ' are transliterated as 'th'; 'ਡ', 'ਦ' are transliterated as 'd'; 'ਢ', 'ਧ' are transliterated as 'dh'. This type of transliteration arises confuse for system at the time of back transliteration.
2. **Ambiguity in mapping of 'n':** Whenever 'n' character occurred as end character of the word then it is difficult to decide where it is half or full character. For example, the both word 'ਸਮੋਣ' and 'ਸਮੋਂ' has same transliterated word as 'samon'; so, there is confusion to decide the actual word for the back transliteration.
3. **Ambiguity due to occurrence of triple 'a':** Whenever a triple 'a' consecutively occurred in a word, then system or even human got confused to read that word. For example, transliteration of the both words, 'ਬਾਅਦ' and 'ਬਅਾਦ' results in 'baaad', so it is hard to decide that transliteration belongs to which word in actual.

V. CONCLUSION AND FUTURE SCOPE

Due to development of new web and mobile technologies, the content on web is generating in huge amount. Most of the content available on social media is in non-structured form (i.e. mix of image, text, audio and video). Also, internet users write text on social networking sites in their own native languages using Roman script. So, the development of phonetic (Roman script) dataset(s) using GRT system is required for natural language processing tasks and applications. Transliteration is a process in which mapping is done for a word written in one language-script to another language-script. The developed GRT system directly transliterate the text from Gurmukhi to Roman script based on grapheme approach. The transliteration of the system depends upon the constituents of words from different characters or symbols (discussed in section 4). The GRT system used CM along with handcrafted rules and tested over 65,130 Punjabi words. The performance of system is evaluated on different cases but gives best performance in case-6 with the accuracy of 99.27%. Also, system performs better than the OGTS (discussed in section 4). In future, the issues discussed in section 4 can be resolved for further advances in the GRT system. This article will provide the detailed concept of transliteration system and its methodology. So, the researcher who belongs from various countries and culture can easily understand the transliteration system. They can develop the new transliteration system for their native and other scripts. The transliterated text from other scripts into the native script of the reader will remove the language or script barriers throughout the world.

REFERENCES

1. International Telecommunications Union, "ICT facts and figures 2017," *Itu*, 2017. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf. [Accessed: 05-Aug-2018].
2. M. Kapko, "7 Staggering Social Media Use By-the-Minute Stats," *CIO From IDG*, 2015. [Online]. Available: http://www.cio.com/article/2915592/social-media/7-staggering-social-media-use-by-the-minute-stats.html. [Accessed: 01-Aug-2018].
3. J. Kaur and J. Singh, "Toward normalizing romanized gurumukhi text from social media," *Indian J. Sci. Technol.*, vol. 8, no. 27, pp. 1–6, 2015.
4. S. K. Singh and K. S. Manoj, "Importance and Challenges of Social Media Text," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 3, pp. 831–834, 2017.
5. D. K. Prabhakar and S. Pal, "Machine transliteration and transliterated text retrieval: a survey," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 43, no. 6, 2018.
6. M. K. Chinnakotla, O. P. Damani, and A. Satoskar, "Transliteration for Resource Scarce Languages," *ACMTransactions Asian Lang. Inf. Process.*, vol. 9, no. 4, pp. 1–30, 2010.
7. J. H. M. Daniel Jurafsky, *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed., vol. 1. USA: Person Prentice-Hall, Inc. Upper Saddle River, NJ, 2018.
8. S. Karimi, "Machine Transliteration of Proper Names between English and

Persian," RMIT University, Melbourne, Victoria, Australia, 2008.
9. K. Kaur and P. Singh, "Review of Machine Transliteration Techniques," *Int. J. Comput. Appl.*, vol. 107, no. 20, pp. 13–16, 2014.
10. K. Kaur, "Machine transliteration : A Review of Literature," *Int. J. Eng. Trends Technol.*, vol. 37, no. 6, pp. 327–336, 2016.
11. S. Karimi, F. Scholer, and A. Turpin, "Machine transliteration survey," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–46, 2011.
12. A. Malik, "Punjabi machine transliteration," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL.*, 2006, no. July, pp. 1137–1144.
13. K. Deep and V. Goyal, "Development of a Punjabi to English Transliteration System," *Ijcsc*, vol. 2, no. 2, pp. 521–526, 2011.
14. P. Kumar and V. Kumar, "Statistical machine translation based punjabi to english transliteration system for proper nouns," *Int. J. Appl. or Innov. Eng. Manag.*, vol. 2, no. 8, pp. 318–321, 2013.
15. M. Kaur, "A More Accurate Punjabi to English Machine Transliteration System for Proper Nouns," *Int. J. Adv. Res. Sci. Eng.*, vol. 6, no. 8, pp. 1455–1469, 2017.
16. S. S, "Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective," Aug. 2017.
17. S. Ganesh and S. Harsha, "Statistical Transliteration for Cross Language Information Retrieval using HMM alignment model and CRF," in *the Workshop on CLIA, Addressing the Needs of Multilingual Societies (IJCNLP'08)*, 2008, no. June, pp. 42–47.
18. A. Kumaran and T. Kellner, "A generic framework for machine transliteration," in *30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, 2007, pp. 721–722.
19. L. Haizhou, Z. Min, and S. Jian, "A joint source-channel model for machine transliteration," *Proc. 42nd Annu. Meet. Assoc. Comput. Linguist. - ACL '04*, p. 159–es, 2004.
20. A. Ekbal, S. K. Naskar, and S. Bandyopadhyay, "A modified joint source-channel model for transliteration," in *Proceedings of the COLING/ACL on Main conference poster sessions -*, 2006, no. July, pp. 191–198.
21. N. AbdulJaleel and L. S. Larkey, "Statistical transliteration for english-arabic cross language information retrieval," in *Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03*, 2003, pp. 139–146.

AUTHORS PROFILE

Mr S K Singh is currently Research Scholar in Department of Computer Science and Engineering at Sant Longowal Institute of Engineering and Technology (SLIET), Sangrur, Punjab, India. He obtained his B.Tech degree in Information Technology from UPTU (Lucknow) in 2012 and Master of Engineering in Software Engineering from Birla Institute of Technology (BIT), Mesra, Ranchi (India). He has qualified national level exams (GATE-13 & UGC-NET-2018). His research interests include handwriting recognition, sentiment analysis, natural language processing, human stress level prediction, personality detection and data mining.

Dr. M K Sachan is currently Associate Professor at Sant Longowal Institute of Engineering and Technology (SLIET), India. He is associated with the Department of Computer Science and Engineering. He did his B.Tech in Computer Science from Punjabi University, Patiala, India. He did M.E in Computer Science from Thapar Institute of Engineering & Technology, Patiala and Ph.D from Punjab Technical University, Jalandhar, India. His research interests include handwriting recognition, stress detection, opinion mining, medical image processing, natural language processing and data mining.