International Journal of Control and Automation
Vol. 12, No. 4, (2019), pp. 199-206

Hidden Markov Model based Punjabi to English Machine Transliteration System

Kamal Deep Garg[1], Dr. Umrinderpal Singh[2], Shivani Gupta[3]
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1]
GHG Khalsa College, GurusarSadhar, Ludhiana, Punjab, India[2]
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[3]

Abstract: Machine Translation is an ongoing research in area of natural language processing. To develop an accurate machine translation system there is a need of machine transliteration system also. This paper presents a hidden markov model based approach for Punjabi to English Machine Transliteration system. Our proposed system takes input written in Punjabi and transliterate into English (Roman script) preserving the phonetic structure of words. The overall accuracy of the proposed system is 90%.

Keywords*:Transliteration, Punjabi, English, HMM*

1. Introduction

Machine Translation is conversion of source language text into target language text with aid of computer program whereas Machine Transliteration is conversion of word of source language into target language without changing its phonetics[1][2]. E.g. a Punjabi ਸੁੰਦਰ is translated as “beautiful” in English, and transliterated as “Sunder” in English. Machine transliteration is a module of machine translation system which is used for handling named entities and out of vocabulary words (OOVW) of a target language. Forward transliteration transliterate a word from its own language to another language. Backward transliteration transliterate a loan word from another language to the language of the word. E.g. ਕਮਲ is Punjabi language word and ਜਿੰਨੀ is loan word from English. Our proposed system deals with both forward and backward transliteration.

There exists different techniques for machine transliteration but all are classified into three categories: grapheme based, phonetic based and hybrid based approach[3]. Grapheme based transliteration directly maps the graphemes of one language to another language[4]. Source channel model, Decision model and Maximum Entropy model are grapheme based models[5][6]. Phonetic based transliteration maps the phoneme of one language to another language. Weighted Finite State transducers and extended Markov window are phoneme based models. Hybrid based approach is combination of grapheme and phonetic based approach[7][8].

2. Gurmukhi and Roman Script

Each language is written by using script. Hindi and Sanskrit is written by using Devnagari Script whereas English is written by using Roman script. A language may have more than one script also. Punjabi language is written in Gurmukhi script in India and in Shahmukhi script in Pakistan[9].Shahmukhi is written from right to left, while Gurmukhi is written from left to right. Gurmukhi is standardized by second Guru Angad Dev[10]. Each script has its own set of vowels and consonants. We are discussing the consonants and vowels of Gurmukhi and roman script one by one. The Gurumukhi script contains thirty five distinct letters and two type of vowels: Basic Vowels (Independent Vowel) and Initial Vowels (Dependent Vowel).

Basic Vowels are:

ੳ(Ura)	ਅ (Aira)	ੲ(Iri)

Table 1: Basic Vowels

ISSN: 2005-4297 IJCA

There are 10 initial vowels in Punjabi. These are used with consonants in Punjabi. Following are the initial vowels of Punjabi language in Gurmukhi script

ਅਾ	ਆ	ਇ	ਈ	ਉ
ਊ	ਏ	ਐ	ਓ	ਔ

Table 2: Initial Vowels

Initial vowels when used with another constant make new letter. Table 3 list all new letters formed by combination of consonant ਕ with all initial vowels.

Initial Vowel	**With /k/**	**Name of letter**
ਅ	ਕ	Mukta
ਆ	ਕਾ	Kanna
ਇ	ਕਿ	Sihari
ਈ	ਕੀ	Bihari
ਉ	ਕੁ	Onkar
ਊ	ਕੂ	Dulankar
ਏ	ਕੇ	Lavan
ਐ	ਕੈ	Dulavan
ਓ	ਕੋ	Hora
ਔ	ਕੌ	Kanora

Table 3: Combination of consonants with initial vowels

Name		**Pron.**	**Name**		**Pron.**	**Name**		**Pron.**	**Name**		**Pron.**	**Name**		**Pron.**
									ਸ	Sussa	Sa	ਹ	Haha	Ha
ਕ	Kakka	Ka	ਖ	Khukha	Kha	ਗ	Gugga	Ga	ਘ	Ghugga	Gha	ਙ	Ungga	Nga
ਚ	Chuchaa	Ca	ਛ	Chhuchha	Cha	ਜ	Jujja	Ja	ਝ	Jhujja	Jha	ਞ	Neya	Nya
ਟ	Tainka	Tta	ਠ	Thutha	Ttha	ਡ	Dudda	Dda	ਢ	Dhudda	Ddha	ਣ	Nahnha	Nna
ਤ	Tutta	Ta	ਥ	Thutha	Tha	ਦ	Duda	Da	ਧ	Dhuda	Dha	ਨ	Nunna	Na
ਪ	Puppa	Pa	ਫ	Phupha	Pha	ਬ	Bubba	Ba	ਭ	Bhubba	Bha	ਮ	Mumma	Ma
ਯ	Yaiyya	Ya	ਰ	Rara	Ra	ਲ	Lulla	La	ਵ	Vava	Va	ੜ	Rharha	Rha

Consonants in Punjabi are also of two type one without dot at foot and one with dot at foot of consonant.

Table 4: Consonants of Punjabi

In addition to these, there are six consonants created by placing a dot (bindi) at the foot (pair) of the consonant (these are not present in Sri Guru Granth Sahib):

Name		Pron.
ਸ਼	Sussa pair bindi	Sha
ਖ਼	Khukha pair bindi	Khha
ਗ਼	Gugga pair bindi	Ghha
ਜ਼	Jujja pair bindi	Za
ਫ਼	Phupha pair bindi	Fa
ਲ਼	Lulla pair bindi	Lla

Table 5: Consonant with dot at foot

English language is written by using roman script. There are total 26 letters in English. Out of 26, 5 are vowels and 21 are consonants[11]. Vowels are:

A	E	I	O	U

Table 6: Vowels in English

Consonants are:

B	C	D	F	G	H	J
K	L	M	N	P	Q	R
S	T	V	W	X	Y	Z

Table 7: Consonants in English

3. Problems in Transliteration

1. There are various problems in Machine Transliteration[12]. Each language has its own set of vowels and consonants. So we cannot use direct character mapping for transliteration.

Language	Vowels	Consonants
Punjabi	3(independent vowel) 10(dependent vowel)	35
English	5	21

Table 8: Vowels and Consonants in Punjabi & English

2. Sound or phoneme is also different for same character in different languages. We may need to combine two or more character to create a sound in a language. For example

Punjabi Character whose sound is not present in English characters	Equivalent English Character
ਖ	Kh
ਛ	Chh
ਘ	Gh
ਞ	Yan
ਥ	Th

Table 9: Punjabi and Equivalnet English Character

3. For a single sound there can be more than one repersentation in different language. So we have to decide which have to choose when. In case of Punjabi and English it is also applicale as shown in below table 10.

Punjabi Character	Equivalent English Character
ਵ	v or w
ਫ	f or ph

Table 10: Punjabi and English Sound

4. **Hidden Markov Model**

Markov Model describe state sequence which are easily noticeable to observer[13]. In Hidden Markov Model identify the hidden states or unobserved events based on observed sequence. Input words are observed events or sequences where target transliteration is hidden or unobserved events which we need to generate based on source input text[14]. One can define HMM precisely as:

$\lambda = (P, Q, \pi)$ (1)

Where$P$a set of transition probabilities is, $Q$is set emission probabilities and $\pi$is set of initial probabilities.

First order Hidden Markov Model is the simplest one where possibility of event occur depends upon the previous event, one can described as:

Markov Inference: $P(t_j | t_1 \dots t_j - 1) = P(t_j | t_j - 1)$ (2)

Where$P(t_j | t_j - 1)$ can be computed as:

$$P(t_j | t_j - 1) = \frac{C(t_j - 1 t_j)}{C(t_j - 1)} \quad (3)$$

Emission probabilities event for observation obj depends only on the event which produced the observation t_j can be defined as:

Output Inference: $P(ob_j | t \ \ , \dots t_T, ob_1, \dots ob_j, \dots, ob_T) = P(ob_j | t_j)$ (4)

Where $P(ob_j | t_j)$can be computed as:

$$P(ob_j | t_j) = \frac{C(t_j \to ob_j)}{C(t_j)} \quad (5)$$

In HMM training we need to compute probabilities for emission, transition and initial states of model. To determine the likelihood of sequence of observation from training data we can use equation (2) and (3) for event sequence $T = t_0, t_1, t_{2\dots}, t_T$and observation sequence $OB = ob_1, ob_2, \dots ob_T$;to compute the joint probability density for first order hidden markov model is:

$$P(ob_1, \dots, ob_M, t_1, \dots, t_M) = P(t_1) \prod_{m=2}^{M} P(t_m | t_m - 1) * \prod_{m=1}^{M} P(ob_m | t_m) \quad (6)$$

Where$P(t_1) \prod_{m=2}^{M} P(t_m | t_m - 1)$transition probabilities and $\prod_{m=1}^{M} P(ob_m | t_m)$ emission probabilities.

Most of researcher shows that trigram HMM yields the better results and do not require much training time as compared to four gram or grater n-gram value. Trigram HMM can be defined as:

Transition probabilities: $P(t_j | t_1 \dots t_j - 1) = P(t_j | t_j - 1, t_j - 2)$ (7)

Where $P(t_j | t_j - 1, t_j - 2)$can be computed as:

$$P(t_j|t_j-1,t_j-2)=\frac{C(t_j-2t_j-1,t_j)}{C(t_j-2t_j-1)} \quad (8)$$

This markov assumption shows that current event's probability is based on previous two event already occurred as compared to bigram where current word is only depends upon previous one word.

Emission probability: $P(ob_j|t \ \ ,\dots t_T,ob_1,\dots ob_j,\dots,ob_T)=P(ob_j|t_j) \quad (9)$

Where output state in trigram model remain same as bigram model to yield the hidden state of current word or object.

Transition probabilities are bigram probabilities based on sequence and emission probabilities are likelihood of word and hidden states. HMM can be n-gram based model to learn probabilities like bigram, trigram etc. We also need to handle zero probabilities in emission and transitions probability calculation using various smoothing techniques. There are various available smoothing techniques to counter zero probabilities, such technique are add one smoothing, discounting method and interpolation.

For example, if we are computing the probability of trigram events $P(t_j|t_j-1,t_j-2)$ using equation (8) and got zero count for this trigram then we can approximate this count using bigram$P(t_j|t_j-1)$and then still no count present for bigram event then we go for unigram probability$P(t_j)$.

We can merge all these back off counting approach in interpolation which can be computed as:

$$P(t_j|t_j-1,t_j-2)=\begin{cases}P(t_j|t_j-1,t_j-2), if C(t_j-2,t_j-1,t_j)>0\\ \alpha_1 P(t_j|t_j-1), if C(t_j-2,t_j-1,t_j)=0 \wedge if C(t_j-1,t_j)>0\\ \alpha_2 P(t_j), otherwise\end{cases}$$

Back-off smoothing is non-liner method to estimate for n-gram is allowed to back-off through progressively shorter histories but problem with this approach is probability estimates may change when adding more data when back-off method select a different order of n-gram model on which to base the estimate. To overcome this we can use better smoothing method like liner interpolation defined as:

$$P(t_j|t_j-1t_j-2)=\lambda P(t_j|t_j-1t_j-2)+\lambda P(t_j|t_j-1)+\lambda P(t_j)$$

Such that summation of all $\lambda' sis1i.e\sum_j \quad \lambda_j=1$

Decoding Process: Decoding problem find the most likelihood state sequence from the given observation $O=(ob_1,\dots ob_j,\dots,ob_T)$to decoding the HMM model and find most probable state sequence with the maximum likelihood. To achieve this we have used Viterbi algorithm. Viterbi algorithm backtrack the whole sequence after decoding the hidden state sequence. Viterbi algorithm defined as:

$$input: charsequence x_1\dots x_n \text{and} Parameters P(t_j|t_{n-1}), P(ob_j|t_j)$$
$$Define K set of all tags. K_{-1}K_0=START$$
$$\pi(0,start,start)=1$$
$$fork=1\dots n$$
$$fora\epsilon K_{-1}, b\epsilon K_k$$
$$\pi(k,a,b)=argmax\left(\pi(k-1,ob_{j-1},ob_j)*P(t_j|t_{j-1})P(ob_j|t_j)\right)$$
$$Return argmax\left(\pi(n,ob_j)*P(stop|ob_j)\right)$$

5. Transliteration Methodology

Punjabi to English transliteration process divided into three different sub-modules pre-processing, segmentation, Model generation and model decoding. In pre-processing phase system identify word boundaries which usually separated by white-space. In pre-processing task algorithm also separate symbols from words like ਰਿਹਾ, → ਰਿਹਾ, by adding extra space in between. This phase also known as word tokenization. Next phase is to create segmentation of each word to be processed. We have used various rule to create character segmentation. These segments usually attach consonants with surrounding vowels like; word ਰੁਪਿੰਦਰwill divide into these segments ਰੁਪਿੰਦਰand its target transliteration will be segmented from RUPINDER to RU PIN D ER. We need to take care total number of segment into both words should be equal. There for it will be one to one mapping. This will be the training data for HMM model. For target word segments system will generate transition probabilities and from source and target segments it will generate emission probabilities. At last decoding process take Punjabi input word and based on training data it will generate possible target sequence in English characters.

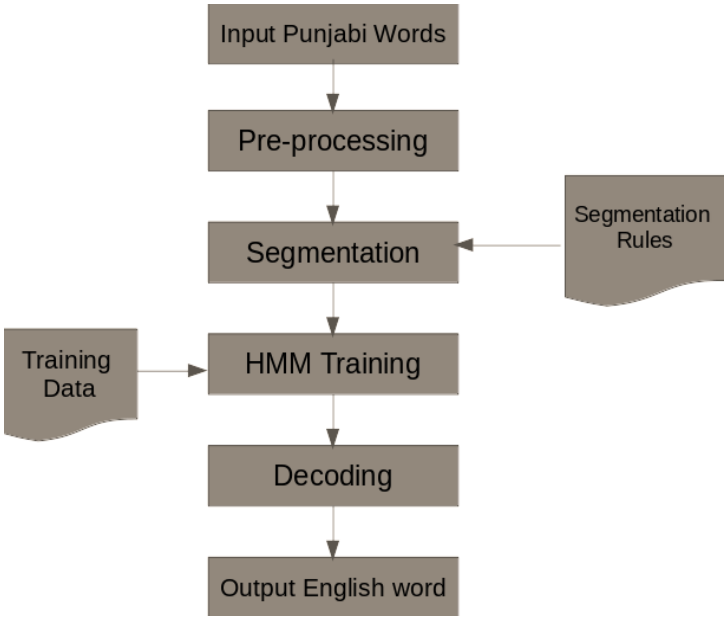

Figure 1: Proposed Methodology

6 Experiments and Results

We have created two data sets, one for training and another for testing. Training data set contains names in such a way that number of segment into both words should be equal. Our training data set contains 890 names and for test set contains 100 names. Both training and test set contains person names, city names, State names and River names etc.Some of Training data set is shown in below table 11.

Punjabi	English
ਨਿਤਿਨ	ni ti n
ਮਨਿੰਦਰ	ma nin de r
ਫੈਜ਼ਨ	fai za n
ਸਿਮੇਜੀਤ	si me ji t
ਸੁਦਰਸ਼ਨ	su da r sha n
ਨਾਮ	naa m
ਦੀਪਕ	dee pa k
ਖਾਨ	kha n

ਹਰਪ੍ਰੀਤ	ha r pree t
ਮੰਤਰ	man t r
ਪੰਕਜ	pan k j
ਗੋਇਲ	go ya l

Table 11: Training data set

To measure the accuracy of the transliteration, accuracy rate is used. Accuracy rate is the count of correct transliteration generated by proposed system divided by the count of total transliteration generated by proposed system. Accuracy of our proposed system is 90%. Some of names transliterated by our system.

Punjabi Word	Our Proposed System	Kamal.et.al(2011)[10]
ਅਮਨਜੋਤ	AMANJOT	AMANJOT
ਅੰਕਿਤਾ	ANKITA	ANKITA
ਰਾਜਬੀਰ	RAJBEER	RAJBEER
ਰਵੀਨਾ	RAVEENA	RAVEENA
ਅਭਿਸ਼ੇਕ	ABHISHEK	ABHISHEK
ਬਲਵਿੰਦਰ	BALWINDER	BALWINDER
ਲੀਜ਼ਾ	LIZA	LIZA
ਮਨੋਜ	MANOJ	MANOJ
ਲੱਕੀ	LUCKY	LAKKI
ਪ੍ਰਿੰਸ	PRINCE	PRINS

Table 12:Transliteration results and compare with Kamal.et.al [10] system

Conculsion

This paper address the transliteration problem using Hidden Markov Model. HMM is a statistical-based approach which is learned from annotated training data. Most of the transliteration system developed using handcraft rules and one should need to write language-specific rules are not transferable to other languages. Where the HMM system is a generic system can be trained for any language based on training data. We have compared our system with other available transliteration system and our system outperformed and yield higher accuracy. This transliteration system can be used in various application as a sub-system. We have used this system in the post-processing phase of Punjabi to English Machine Translation system to transliterate names and out of vocabulary(OOV) words to English.

References

[1] M. G. A. Malik, "Punjabi machine transliteration," *COLING/ACL 2006 - 21st Int. Conf. Comput. Linguist. 44th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, vol. 1, no. July, pp. 1137–1144, 2006.

[2] S. Mathur and V. P. Saxena, "Hybrid appraoch to english-hindi name entity transliteration," *2014 IEEE Students' Conf. Electr. Electron. Comput. Sci. SCEECS 2014*, 2014.

[3] P. Patel and P. Bhattacharyya, "Recent Work in Machine Transliteration for Indian Languages," pp. 1–13.

[4] A. Kumaran, M. M. Khapra, and P. Bhattacharyya, "Compositional machine transliteration," *ACM Trans. Asian Lang. Inf. Process.*, vol. 9, no. 4, pp. 1–28, 2010.

[5] A. Sharma and RattanDhavleesh, "Machine Transliteration for Indian Languages: a Review," *Int. J. Adv. Res. Comput. Sci.*, vol. 3, no. 7, pp. 274–279, 2017.

[6] D. F. Wong, Y. Lu, and L. S. Chao, "Bilingual recursive neural network based data selection for statistical machine translation," *Knowledge-Based Syst.*, vol. 108, pp. 15–24, 2016.

[7] P. H. Rathod, M. L. Dhore, and R. M. Dhore, "Hindi And Marathi to English Machine Transliteration

using SVM," *Int. J. Nat. Lang. Comput.*, vol. 2, no. 4, pp. 55–71, 2013.

[8] J. H. Oh, K. S. Choi, and H. Isahara, "A comparison of different machine transliteration models," *J. Artif. Intell. Res.*, vol. 27, pp. 119–151, 2006.

[9] G. S. Lehal, T. S. Saini, and S. K. Chowdhary, "An Omni-font Gurmukhi to Shahmukhi Transliteration System," *Coling*, vol. 3, no. December 2012, pp. 313–320, 2012.

[10] K. Deep and D. V. Goyal, "Hybrid Approach for Punjabi to English Transliteration System," *Int. J. Comput. Appl.*, vol. 28, no. 1, pp. 1–6, 2011.

[11] K. Deep and V. Goyal, "Development of a punjabi to english transliteration system," *Ijcsc*, vol. 2, no. 2, pp. 521–526, 2011.

[12] D. Chopra and S. Morwal, "H ANDLING U NKNOWN W ORDS I N N AMED E NTITY," vol. 2, no. 4, pp. 87–93, 2013.

[13] V. Sumalatha, "A Study on Hidden Markov Model (HMM)," pp. 465–469, 2014.

[14] A. Krogh, "Applications of Hidden Markov Model Approaches," vol. 235, no. August, pp. 1384–1391, 1994.