ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

English to Sanskrit Transliteration: an effective approach to design Natural Language Translation Tool

Leena Jain
Department of Computer Application
IKG-Punjab Technical University
Kapurthala, India

Prateek Agrawal*
Department of Computer Science Engineering
IKG-Punjab Technical University
Kapurthala, India

***Abstract*:** This paper explains two approaches to develop transliteration tool which helps to convert text written in English language to equivalent Sanskrit language. First approach is through typing using computer keyboard while second approach explains typing through virtual keyboard method. Our objective is to implement a user friendly and robust tool that provides facility to users about to convert English text into Sanskrit language without any error. Major application of this tool is to train the students at school level and make them learnable. This tool promotes the use of native language among the users too. Another major application of this tool is to write govt notices, story writing, article writing etc**.**

Keywords: Sanskrit Transliteration, natural language processing, machine translation, Hindi transliteration, virtual keyboard

I. INTRODUCTION

Sanskrit is known as the oldest language spoken in India. It is also recognised as language of Lord Shiva in *Sanatana Dharma*. It follows *Devnagari* script. Similarly, Hindi - the major portion of language in India, is also having same script i.e *Devnagari* but still in its childhood stage concerning to natural language processing research and applications [1-3]. Nowadays, there is lots of work done for English Title Generation but not much work has been done for Hindi language because of less familiarity of Sanskrit or Hindi typing keyboards. English language has only two types of alphabets i.e. 21 consonants (e.g B,C,D) and 5 vowels (a,e,i,o,u), but in Hindi language there are 40 consonants, 10 vowels, which consists of various types of symbols (ten vowels has sign (matras), half letters & halant etc.). So the methods which are available for English language cannot be directly applied for Hindi language. Various novels, documents, newspaper, government notice, books, magazines, etc. are written in Hindi language so there is need for development of title generation tools for Hindi. Natural language processing (NLP) is a field of computer science and scientific study of language which deals with the interactions between computers and human languages [4].

There are many challenges for interaction between computers and humans. Computers understand only binary digits but humans can't deal with binary digits. So we require huge database stored in our system for processing of human understandable words by computers. Natural Language Processing (NLP) is one of the technique through which humans can interact with computers [6] Since, NLP is interconnected to the field of human–machine interaction there are many difficulties involved in making a computer understand a human language.

First and the foremost challenge is the platform or the framework through which humans can interact with computer. Natural language processing (NLP) technique can be used to generate the title of the given story in Hindi. One of the best suitable languages for working of Natural Language Processing is Java. The historical backdrop of NLP by and large begins in the 1950s. But despite of this fact the work of NLP can be found from very early time. In 1950, Alan Turing distributed an article titled Processing Machinery and Intelligence" which proposed what is currently called the Turing test as a rule of knowledge.

There are various tasks that can be performed using NLP. Some of these tasks are:-

Transliteration:

Transliteration is a process to convert the script written in one language into another language. The program transliterates this sentence into Sanskrit with the help of the transliterator that is encoded in the software application. The translator uses Unicode for the same. It takes Sanskrit / Hindi input (in English) and Transliterate it in Sanskrit later it translate it in Hindi. We took this step because writing in Hindi is a problem for every user. As Hindi keyboards are not usually available [7-9].

Automatic Summarization:

In this technique produce a meaningful outline of a summary of the content of a document [10]. For example, articles in the sports related segment of a daily newspaper. Summarization of content may be done with the help of paraphrasing of sentences [11-12].

Co-reference resolution

Co-reference resolution the assignment of all the aspects that refers to same element in a text or paragraph. It is one of the most important tasks for any NLP operation. It helps in information extraction from any paragraph. Here in the proposed system Stanford Core-NLP is used to resolve co references in any given text or paragraph.

For example: *John is a boy. He lives in Jalandhar. He is a student.*

In above example, there are 3 sentences and each sentence is having either noun pr pronoun as given in table 1.

Table 1: Co-reference resolution

Sentence Number	Word
1	John
2	He
3	He

Discourse Analysis

This discourse analysis incorporates various inter-related assignments. One undertaking is distinguishing the structure of associated content that is the way of the connections is made between the sentences. Other conceivable assignment is perceiving and ordering the discourse demonstrations in a summary of content that is yes-no queries, substance questions, explanation, declaration, etc.

Machine Translation

One of the most important tasks in NLP is automatic machine translations from one language to another language. It is one of the most difficult tasks and it requires different types of knowledge that a human possess such as grammar rules, semantic rules, and various other knowledge.

Morphological segmentation

Separating the words in different units and recognize the class of these units. The trouble of this assignment depends significantly on the intricacy of the word structure of the dialect being taken into consideration. English has genuinely straightforward word structure, particularly inflectional word structure, and hence it is frequently conceivable to overlook this undertaking totally and basically show every conceivable type of a word as partitioned words.

Name Entity Recognition (NER)

When a paragraph or text is given, Natural Language Processing technique is used to identify names entities in the paragraph. These name entities can be name of a place, individual, association, area etc. Name entities enable a machine to work like humans and process sentences in the same way as humans.

Natural language understanding

When a text has formal representations, in different way which a not defined in the databases, it is difficult for a computer to get the meaning of that paragraph. Here natural language processing plays an important role. NLP enables a computer get the paragraph meaning and apply different grammar rules accordingly [12-13].

Optical character recognition (OCR)

One of the most important tasks of Natural Language Processing is Optical Character Recognition. NLP techniques scan an image and identify the character in that image. This technique is very helpful to read texts from a banner and posters where direct text is not available and everything is in the form of an image.

Question answering

When question is presented to a machine it is very difficult for a machine to understand that question without any processing. Here comes the role of Natural Language Processing. NLP techniques process the questions and convert it in the form of an equation which is easily understandable of the machines.

Speech recognition

When a sound clip of a person is given to a computer, the computer finds the way in which that text is represented. It is very difficult for a machine to process speech without any NLP techniques. Many types of the languages are available that are spoken and represented in different ways. So the machine converts those analog signals into distinct characters and processes accordingly [14]. There is a certainly no pause between two words in speech to is it very difficult to segment a speech into subtask of speech. So natural languages processing plays an important role in speech processing.

Hindi language is much complex than English because Hindi language categories the alphabets into two parts:

Consonants: There are 36 consonants known as "व्यंजन" in Hindi language as shown in table 2.

Vowels: There are 12 vowels known as "स्वर" and 09 other secondary vowels in Hindi are shown in table 2 which include various types of symbols like sign / *maatraa*, half letters / *halants* etc.

So, due to these differences between both the languages the methods which are applicable for English language are not able to be directly used for the systems of Hindi language. Hindi, the major portion of language in India, is still in its childhood stage concerning to natural language processing research and applications. Various novels, documents, newspaper, government notice, books, magazines, etc. are written in Hindi language so there is need for development of title generation tools for Hindi.

NLP is the scientific study of natural languages and a field of computer science which makes computer interact-able with the human beings. NLP is used to train the computer for various natural languages. There are many challenges for interaction between computers and humans [5]. Computers understand only binary digits but humans can't deal with binary digits. So we require huge database stored in our system for processing of human understandable words by computers. NLP is one of the most promising technique through which humans can interact with computers. Since, NLP provides a platform through which human being can interact with computer in any natural language but making computer understandable about human language is a very difficult task.

User can input the text in Hindi by using Hindi keyboard provided by our interface. The user need to the click on the button of Hindi keyboard and type with the help of that keyboard [4]. This type of keyboard will be helpful to those who are not aware of using English keyword to type Hindi alphabets. This can also be use to teach students about various matras and word formation in an innovative and interesting way.

User can easily display the keyboard by clicking on the button of virtual keyboard and also easily exit from this keyboard by pressing the exit button on keyboard. There is also an option available for reset the whole text which will clear the existing text in the input box. An important feature of this keyboard is that the alphabets are arranged in this keyboard are in such an order in which teachers teach them. This order makes the keyboard user friendly because user can easily find the alphabets due to this sequence. Then, user can exit from this keyboard section by just clicking on the exit button present on the keyboard.

Table I. Table Type Styles

VOWELS									
a	अ	aa	आ	i	इ	ee	ई	u	उ
e	ए	ai	ऐ	o	ओ	au	औ	M	अं
AOM	ॐ	RiR	ऋ	RIR	ॠ	aA	अऽ	En	ऍ
AO	अँ	oo	ऊ	H	अः	Ao	ऑ	|	।
CONSONENTS									
k	क	kh	ख	g	ग	gh	घ	ND	ङ
ch	च	Ch	छ	j	ज	jh	झ	NY	ञ
T	ट	TH	ठ	D	ड	DH	ढ	N	ण
t	त	th	थ	d	द	dh	ध	n	न
p	प	ph	फ	b	ब	Bh	भ	m	म
y	य	r	र	l	ल	v	व		
sh	श	Sh	ष	s	स	h	ह		
kSh	क्ष	tr	त्र	jNY	ज्ञ				

METHODOLOGY

Here we have designed an algorithm which will automatically convert the text typed in English to Hindi language. This conversion is known as transliteration. This process will help the user to easily enter the input. For this algorithm we will make use of Hindi language Unicode's which will make the mapping between English and Hindi alphabets and will generate the desired output i.e. text in Hindi language.

Firstly we will discuss, what are Unicode?

Unicode are those numbers which are assigned to various characters of any language. Every language has different set of Unicode for each character of that language. The numbers which are assigned to any character is unique. These Unicode will help the computer system to understand various languages. Unicode do not depend on any platform or program. Unicode are helpful in typing any language in our computer system. With these we can do transliteration as well. All the languages like Hindi, English, Punjabi, Sanskrit all have their unique set of Unicode for each alphabet. Here we are discussing about the Unicode set of Hindi language which is very helpful in performing transliteration from English to Hindi i.e. using English keyboard we can type in Hindi language.

A. Transliteration Algorithm

***Method 1*:** Proposed algorithm to transliterate through tying keyboard is explained as follows:

```
(Input[ ]) : String
//To transliterate the English Input text to its Sanskrit equivalent
1. Initialise Hindi[ ] array
2. Initialise Unicode[ ] array
3. S[ ] ← NULL                          //String for the result
4. While(Input != NULL) do
        {
                I ← 0
                Key := Input[i]
                        If( Key == Hindi[i])
                        {
                                S[i] ← Unicode[i]
                        }
                Increment I;
        }
5. return s ;
```

Method 2:

The interface is also having virtual keyboard as shown below in Fig. 1:

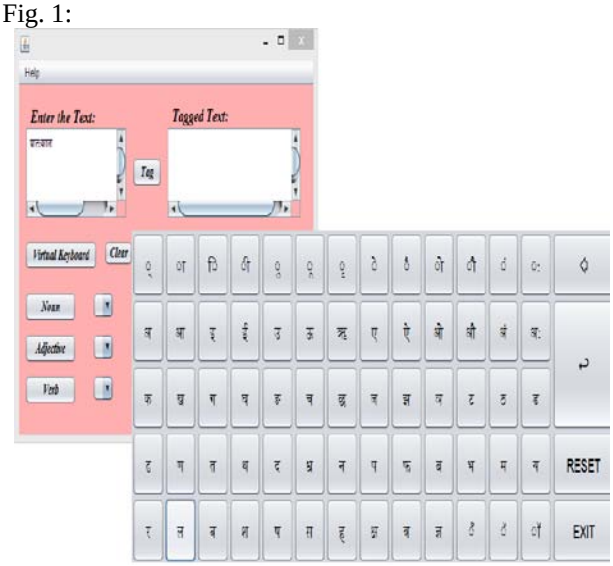

FIG. 1: TRANSLITERATION USING VIRTUAL KEYBOARDEASE OF USE

RESULTS & DISCUSSION

Proposed interface for transliterating the sentences are shown in following fig 2(a) and fig. 2(b). It is clearly shown that text typed in English are converted into equivalent Sanskrit Language. In fig 2 (a) & (b), user is able to type at left window panel of the tool. Testing on some complex and simple words were performed on the tool as shown in table 3. All the test cases are passed which indicate the strong robustness of the developed tool.

Fig. 2 (a): Interface to transliterate sentence

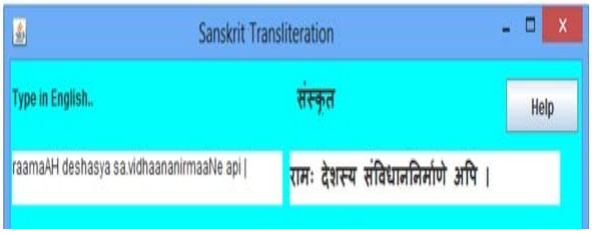

Fig. 2 (b): Interface to transliterate sentence

Table 3: English to Sanskrit Transliteration testing

Input	Output Obtained	Valid Output	True / False
raamaAH	रामः	रामः	T
aa`Dgladesheeyaa	आङ्गलदेशीया	आङ्गलदेशीया	T
prabhoota	प्रभूत	प्रभूत	T
digvijayaaya	दिग्विजयाय	दिग्विजयाय	T
karSha	कर्ष	कर्ष	T
neela paridhaana	नील परिधान	नील परिधान	T
kaalachakramiva	कालचक्रमिव	कालचक्रमिव	T
shuShkasaraAH	शुष्कसरः	शुष्कसरः	T
navavaaripoorNaaAH neelameghaaAH	नववारिपूर्णाःनीलमेघाः	नववारिपूर्णाःनीलमेघाः	T
viShakumbhapayomukham	विषकुम्भपयोमुखम्	विषकुम्भपयोमुखम्	T

CONCLUSION

Two methods for transliterating the text from English to Sanskrit language is proposed and implemented. 100% accuracy of the system is achieved. All the possible letters of Sanskrit and Hindi language are incorporated into it. The tool is user friendly and can be helpful in machine learning and natural language translation. Also it can be used as an effective tool to type documents in Hindi and Sanskrit language.

REFERENCES

[1] Dorr, B., Zajic, D., & Schwartz, R., Cross-language headline generation for Hindi. ACM Transactions on Asian Language Information Processing, 2(3), 2003, 270–289. doi.: 10.1145/979872.979878

[2] Knight, K., & Graehl, J., Machine Transliteration. Computational Linguistics, 24(4), 1998, 599–612. doi: 10.3115/976909.979634

[3] Kumaran A, Khapra, M. M. & Bhattacharyya P, Compositional Machine Transliteration, ACM Transactions on Asian Language Information Processing, 9(4), 2010, 1–29. doi: /10.1145/1838751.1838752

[4] Kumaran, A., & Kellner, T., A Generic Framework for Machine Transliteration. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 39, 2007, 721–722. doi: 10.1145/1277741.1277876

[5] Haque, R., Dandapat, S., Srivastava, A. K., Naskar, S. K., & Way, A. English-Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009. Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), 104–107, Aug 2009.

[6] Hersh, W. R., Campbell, E. H., Evans, D. A., & Brownlow, N. D., Empirical, Automated Vocabulary Discovery Using Large Text Corpora and Advanced Natural Language Processing Tools. Proceedings of the AMIA Annual Fall Symposium, 1996, 159–163.

[7] Chinnakotla, M. K., & Damani, O. P., Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009. Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), 2009, 44–47.

[8] Meurers, D., Natural Language Processing and. The Encyclopedia of Applied Linguistics, 28(1), 2013, 4193–4205. doi: 10.1002/9781405198431.wbeal0858

[9] Stokhof, M., Hand or hammer? On formal and natural languages in semantics. Journal of Indian Philosophy Springer Science, 35, 597–626. doi: 10.1007/s10781-007-9023-7

[10] Tiwary, U. S, A language independent approach to multilingual text summarization. Conference RIAO2007, 1(1), 2007, 123–132.

[11] Prateek Agrawal, Vishu Madaan, Nandini Sethi, Vikas Kumar, Sanjay Kumar Singh, "A Novel Approach to Paraphrase English Sentences using Natural Language Processing", International Journal of Control Theory and Applications, 9(11), pp 5119-5128, Aug 2016.

[12] Nandini Sethi, Prateek Agrawal, Vishu Madaan, Sanjay Kumar Singh, Anuj Kumar, "Automated Title

Generation in English Language using NLP", International Journal of Control Theory and Applications, 9(11), pp 5159-5168, Aug 2016.

[13] Nandini Sethi, Prateek Agrawal, Vishu Madaan, Sanjay Kumar Singh, "A novel Approach to Paraphrase Hindi sentences using Natural language Processing", Indian Journal of Science and Technology, 9(28), pp.1-6, July 2016.

[14] Leena Jain, Prateek Agrawal, "Text Independent Root word Identification in Hindi language using Natural Language Processing", International Journal of Advanced Intelligent Paradigm (IJAIP), vol 7, issue3-4, pp 240-249, Sept 2014.