International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7989-7997

English to Bodo Machine Transliteration System for Statistical Machine Translation

Saiful Islam
Research Scholar, Department of Computer Science,
Assam University, Silchar, Assam, India

Bipul Syam Purkayastha
Professor, Department of Computer Science,
Assam University, Silchar, Assam, India

Abstract
Machine translation and transliteration are both the challenging research tasks and applications in the field of Computational Linguistics (CL) and Natural Language Processing (NLP) nowadays. Machine Transliteration (MTn) is a technique to convert script of a text from a source natural language to a target natural language using computer without changing the pronunciation of the source text. On the other hand, Machine Translation (MT) is a technique to translate text from a source natural language to a target natural language using computer. MTn is a highly helpful technique for handling OOV (Out-Of-Vocabulary) words in MT system, especially in statistical and neural machine translation systems. Bodo is a recognized language of India and one of the major spoken languages of North-East India. Still, not much work has been done on MTn system as well as MT system for Bodo language. Therefore, it has been decided to develop English to Bodo machine transliteration and translation systems. The primary objective of this proposed research work is to develop English to Bodo machine transliteration system for enhancing the translation result of English to Bodo Statistical Machine Translation (SMT) system. The MTn system has been developed using hybrid approach with the help of multi-domain English-Bodo parallel words or terms. The accuracy of the translation result of the SMT system has been evaluated using automatic evaluation technique BLEU.

Keywords: Bodo language, English language, Machine translation, Machine transliteration

INTRODUCTION

Machine translation and machine transliteration are both the greatest significant applications of NLP and CL. At present, the MT and MTn systems are very important for a multilingual country like India. A large amount of research work has been done on both the systems for some popular natural language pairs, such as Arabic-English, Chinese-English, French-English, Hindi-Punjabi, Japanese-Spanish, and Urdu-Malayalam. We have been developed English to Bodo SMT system using phrase-based SMT approach with the help of Agriculture, Health, and Tourism domains English-Bodo parallel Text Corpus (E-BPTC). The different domains E-BPTC have been collected from TDIL (Technology Development for Indian Languages) as in MS Excel format [http://tdil-dc.in/index.php]. The SMT system has been developed individually for each domain E-BPTC. In the translation results, it has been noticed that some unknown or OOV words like proper nouns, abbreviations, and technical terms are available in every domain translated Bodo corpus in English scripts which may be reduced the translation accuracy in the MT system. For the lack of availability of English to Bodo MTn system and to enhance the translation accuracy in the SMT system, it has been decided to develop English to Bodo machine transliteration system with the help of Agriculture, Health, and Tourism domains English words or terms and their corresponding transliterated Bodo words or terms. The MTn system has been implemented using hybrid approach which is the combination of Direct Transliteration Technique and Sequential Search Technique.

Machine Transliteration

Machine transliteration is the process of automatically converting the script of a word or term from a source natural language to a target natural language without changing the pronunciation of the source word or term. The main goal of transliteration is to preserve the phonological structure of words or phrases between two particular natural languages. It can perform a very good significant role in the various applications of NLP, such as Machine Translation, Cross-Language Information Retrieval (CLIR), Text to Speech Conversion, Named Entity Recognition, and Part of Speech Tagging (POST) [18]. MTn is completely different from MT; still, it is a very helpful technique for MT system, especially for SMT and Neural Machine Translation (NMT) systems. Both the SMT and NMT approaches use a huge amount of aligned parallel text corpora to achieve high-quality translation results. Every domain parallel text corpus may have some proper nouns, abbreviations, and technical terms which are written in same language script (in our case, English script) in both the source and target corpora. This type of parallel corpora can reduce the quality and accuracy of the translation result. The MTn is one kind of system which can handle the OOV words like proper nouns (people name, place name and so on), abbreviations, and technical terms in MT system.

Machine transliteration can be classified basically into two types namely Forward Transliteration (FT) and Backward

Transliteration (BT). Forward transliteration is simply known as transliteration and backward translation is known as back-transliteration. For example, if English to Bodo transliteration is FB then Bodo to English transliteration will be BT. More precisely; suppose, S is a text of source language and T is a text of target language, then FT means transliteration from S to T and BT means transliteration from T to S [14].

Machine transliteration can be developed using various techniques or models. The different techniques or models of MTn are shown in figure 1 [2].

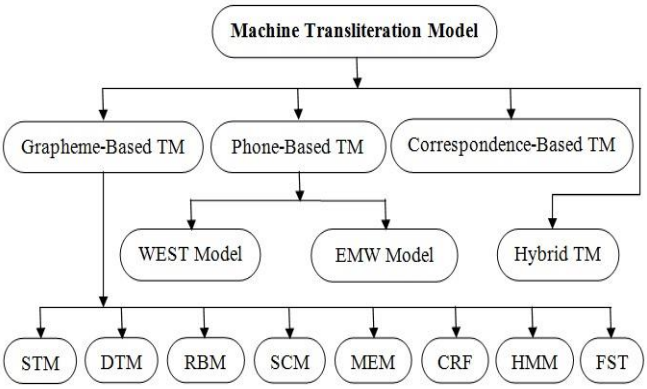


Figure 1: Different techniques of MTn system

The most commonly used techniques or models of MTn include Grapheme-Based Transliteration Model (GBTM), Phoneme-Based Transliteration Model (PBTM), Hybrid Transliteration Model (HTM), and Correspondence-Based Transliteration Model (CBTM) [8, 22]**.** In GBTM, grapheme is the smallest unit of a writing system of a particular language that has its own meaning or grammatical importance. The GBTM is the simplest method for direct orthographic mapping from source language graphemes to target language graphemes. For that, it is also known as Direct Transliteration Technique (DTT). The GBTM can be classified into several categories, such as Statistical Transliteration Model (STM), Decision Tree Model (DTM), Rule-Based Model (RBM), Source Channel model, Maximum Entropy Model, Conditional Random Fields (CRF) model, Hidden Markov Model (HMM), and Finite State Transducer (FST) model [15]. In PBMT, phoneme is a significant unit of sound that distinguishes one word from another in a specific natural language. It is based on pronunciation or the source phoneme rather than spelling or source grapheme. This model is basically used in source grapheme to source phoneme transformation and source phoneme to target grapheme transformation. The PBTM can be classified into two categories, such as Weighted Finite State Transducers (WFST) and Extended Markov Window (EMW) [15]. The HTM and CBTM use both the source grapheme and source phoneme in transliteration [29].

Machine Translation

Machine translation is the process of automatic translation of text from one natural language to another natural language. It is one of the most important research tasks of NLP and CL. MT is done either unidirectional or bidirectional mode between two particular natural languages. It is very necessary for a multilingual country like India because it can translate huge amount text from a source language to a target language within a short period of time which is not possible by human translators. Research work on MT is not a new one. A large number of research works have been done on MT using different techniques in different countries and at different times. There are various approaches of MT, but nowadays the most commonly used approaches include SMT, Rule-based MT, Example-based MT, and Hybrid MT. The Rule-based MT is classified into three categories, such as Direct MT, Transfer MT, and Interlingua MT. The Statistical MT is also classified into three categories namely Word-based SMT, Phrase-based SMT (PB-SMT), and Hierarchical Phrase-based MT [11]. At present time, many researchers have been worked and working on MT system using NMT approach for several language pairs. The NMT approach is a very new and highly active approach in MT.

Comparative Study of Bodo and English Languages

Bodo is a Sino-Tibetan family's language and one of the major spoken languages of North-East India [10]. It is mainly spoken by the people of Kokrajhar, Chirang, Baksa, and Udalguri districts of Assam. Bodo language is also known as Mech and is the fundamental language of Bodo people. The Bodo language was introduced in Assam as a medium of instruction in the primary school in Bodo dominated areas in 1963 and it was the result of an intense political movement carried out by different Bodo organizations since 1913 [20]. Bodo is the official language of Bodoland Territorial Council (Assam) and one of the 22 recognized languages of India. It is a low resource language and is written using Devanagari script. Though Bodo language is written using Devanagari alphabets or script, but based on phonemes it has pure 6 vowels and 16 consonants which are shown in Table 1. The word order of this language is subject-object-verb.

Table 1: Alphabets of Bodo language

BODO ALPHABETS									
Vowels	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ
	ओ	औ	अं	अः	Pure Vowels				
Consonants	क	ख	ग	घ	ङ	च	छ	ज	झ
	ञ	ट	ठ	ड	ढ	ण	त	थ	द
	ध	न	प	फ	ब	भ	म	य	र
	ल	व	श	ष	स	ह	क्ष	त्र	ज्ञ
	श्र	Pure Consonants							

English is the West Germanic language and was the first spoken language in early medieval England. Now, it is an international language and used in all over the world. The English language is spoken mainly by the people of Canada, Australia, United Kingdom, United States, Ireland, and New Zealand. It is an official language of almost sixty sovereign

states. It is the third most common native language in the World. English language was introduced in India (1830) during the rule of the East India Company. In 1951, the Constitution of India declared English as the associate official language of India. Now, it is the third most spoken language in India [9]. It is a high resource language and written using Latin script. It contains 26 alphabets including 5 vowels and 21 consonants which are shown in Table 2. The word order of this language is subject-verb-object.

Table 2: Alphabets of English language

ENGLISH ALPHABETS										
Letters	Uppercase					Lowercase				
Vowel	A	E	I	O	U	a	e	i	o	u
Consonant	B	C	D	F	G	b	c	d	f	g
	H	J	K	L	M	h	j	k	l	m
	N	P	Q	R	S	n	p	q	r	s
	T	V	W	X	Y	t	v	w	x	y
	Z					z				

RELATED WORK

Lots of machine transliteration research works have been done by the researchers of the educational institutions and organizations in most of the countries of the world using various techniques for human languages. Even though a large number of MTn systems have been developed for European and Asian natural languages, such as English to (Arabic, Chinese, Hindi, Korean, and Spanish) MTn systems; still it is an infancy state for Indian natural languages. There are 22 recognized languages and almost 720 dialects in India, but a small number of MTn systems have been developed for English and Indian languages like English to Hindi, English to Malayalam, English to Telugu, Hindi to Urdu, and Panjabi to Hindi MTn systems.

In this section, we briefly discuss the existing MTn systems which have been developed in all over the world and for the Indian natural languages.

MTn System Developed in the World

A wide range of machine transliteration systems has been developed in all over the world by the research communities in the different institutions/organizations using different techniques or models for natural languages. The first transliteration work was done by Arababi through a combination of neural network and expert systems for transliterating from Arabic to English language in 1994 [2]. The next development in transliteration was based on a statistical based approach proposed by Knight and Graehl for back transliteration from English to Japanese Katakana in 1998. This approach was adopted by Stalls and Knight for back transliteration from Arabic to English. Some of the existing MTn systems developed in the world are briefly discussed below:

- ***Arabic to English MTn system:*** The Arabic to English MTn system was developed by Yaser All-Onaizan and Kevin Knight at the Department of Information Sciences Institute, University of Southern California, US in 2002. They have been presented a transliteration algorithm based on sound and spelling mapping using finite state machine approach. The PBTM approach has been also used to develop the system [31].

- ***English to Arabic MTn system:*** The English to Arabic MTn system was developed by N. A. Jaleel and L. S. Larkey at the Department of Computer Science, University of Massachusetts, US in 2003. The MTn system has been developed using STM approach and n-gram model for handling named entities and technical terms or OOV words in the English-Arabic CLIR system [12].

- ***English-Chinese MTn system:*** The English-Chinese MTn system was developed by Ying Qin and Guohua Chen at the Department of Computer Science and National Research Centre for Foreign Language, Beijing Foreign Studies University, China in 2011. They have been developed forward and backward MTn systems using CRF approach between English and Chinese languages for the shared task of NEWS (Named Entities Workshop) 2011. In the forward transliteration (English to Chinese), they have been achieved 0.312 transliteration accuracy and in the backward transliteration (Chinese to English), they have been achieved 0.167 transliteration accuracy [25].

- ***English to Farsi (Persian) MTn system:*** The English to Farsi MTn system was developed by Najmeh Mousavi Nejad, Shahram Khadivi, and Kaveh Taghipour at the Department of Engineering, Islamic Azad University and the Department of Computer Engineering, Amirkabir University of Technology, Iran in 2011. The MTn system has been developed for the shared task NEWS 2011using MEM approach, Sequitur g2p tool, Moses, and PB-SMT approach [21].

- ***English to Korean MTn system:*** The English to Korean MTn system was developed by Jae Sung Lee and Key Sun Choi at the Department of Computer Science, Korea Advanced Institute of Science and Technology, Korea in 1998. The MTn system has been developed using STM approach for Information Retrieval [19].

- ***English to Sinhala MTn system:*** The English to Sinhala MTn system was developed by Budditha Hettige and Asoka Karunananda at the Department of Statistics and Computer Science, Faculty of Applied Sciences, University of Sri Jayawardenepura, Sri Lanka in 2007. The MTn system has been developed using FST approach to improve the translation result of English to Sinhala MT system [8].

- ***Japanese to English MTn system:*** The Japanese to English MTn system was developed by Isao Goto, Naoto Kato, Terumasa Ehara, and Hideki Tanaka at NHK Science and Technology Research Laboratories, Tokyo,

Japan in 2004. The MTn system has been developed using GBMT approach for transliterating the proper nouns from Japanese to English language [7].

- ***Spanish-English MTn system:*** The Spanish-English MTn system was developed by Michael Paul, Andrew Finch, and Eiichiro Sumita at National Institute of Information and Communications Technology, Japan in 2009. The system has been developed using STM approach for handling OOV words and to improve the accuracy of translation result in Spanish-English PB-SMT system. The Spanish-English SMT system has been developed for WMT (Workshop on Statistical Machine Translation) shared task [24].

MTn System Developed for Indian Languages

Although machine transliteration is an important application of NLP, but a small number of MTn system have been developed for Indian languages as compare to MT system developed for Indian languages. The MTn research works have been developed by the researchers of the different institutions/organizations using various approaches or models for English and Indian natural languages. Some of the MTn systems which have been developed for English and Indian languages using various techniques are mentioned below:

- ***English to Bengali (Bangla) MTn system:*** The English to Bangla MTn system was developed by Khan Md. Anwarus Salam, Setsuo Yamada, and Tetsuro Nishino at Graduate School of Informatics and Engineering, University of Electro-Communications, Tokyo, Japan in 2013. The MTn system has been developed using IPA based transliteration technique for transliterating unknown words from English to Bangla language to improve the translation result in the English to Bangla MT system [26].

- ***English to Hindi MTn system:*** The English to Hindi MTn system was developed by Amitava Das, Asif Ekbal, Tapabrata Mandal, and Sivaji Bandyopadhyay at the Department of Computer Science and Engineering, Jadavpur University, Kolkata, India in 2009 [4]. The MTn system has been developed for NEWS 2009 MTn shared task datasets using Modified Joint Source Channel model. The transliteration accuracy of the MTn system was 0.471.

- ***English to Kannada MTn system:*** The English to Kannada MTn system was developed by P. J. Antony, V. P. Ajith, and K. P. Soman at CEN, Amrita University, Coimbatore, India in 2010. The MTn system has been developed using STM approach and Moses for transliterating English names into Kannada language. The transliteration accuracy of the system was 89.27% [3].

- ***English to Kashmiri MTn system:*** The English to Kashmiri MTn system was developed by Mir Aadil and M. Asger at the Department of Computer Science, Baba Ghulam Shah Badshah University, Jammu & Kashmir, India in 2017. The MTn system has been developed using hybrid approach which consists of DTT and PBTM and tested with 15000 words of medical sciences. The transliteration accuracy of the system was 86% [1].

- ***English to Malayalam MTn system:*** The English to Malayalam MTn system was developed by Sumaja Sasidharan, Loganathan, R., and Soman, K. P at CEN, Amrita Vishwa Vidyapeetham, Coimbatore, India in 2009. They have been used Sequence Labeling approach and English-Malayalam parallel corpus with 20000 person names for developing the system. They have been tested the system using 1000 parallel names and achieved 90% transliteration accuracy [27].

- ***English to Manipuri MTn system:*** The English to Manipuri MTn system was developed by Mayanglambam Premi Devi, Irengbam Tilokchan Singh, and Haobam Mamata Devi at the Department of Computer Science, Manipur University, Manipur, India in 2017. The system has been developed based on syllabification process to transliterate from English to Manipuri language [6].

- ***English to Tamil MTn System:*** The first English to Tamil MTn system was developed by Kumaran A. and Tobias Kellner in the year 2007 [2]. Another English to Tamil MTn system was developed by Vijaya M. S., Loganathan R., Shivapratap G., Ajith V. P., and Soman K. P. at CEN, Amrita Vishwa Vidyapeetham Coimbatore, India in 2008. The system has been developed using Sequence Labeling approach and Support Vector Machines (SVM) approach for transliterating English to Tamil language. The transliteration accuracy of the system was 88% [30].

- ***Hindi to English MTn system****:* The Hindi to English MTn system was developed by Veerpal Kaur, Amandeep Kaur Sarao, and Jagtar Singh at Punjabi University Guru Kashi Campus, Punjab, India in 2014. The MTn system has been developed to transliterate proper nouns of Hindi language into equivalent English language using hybrid approach which consists of DTT, RBM, and SMT approaches. The transliteration accuracy of the system was 97% [16].

- ***Manipuri-Bengali MTn system:*** The Manipuri-Bengali MTn system was developed by Thoudam Doren Singh at CDAC Mumbai, India in 2012. The system has been developed using Rule-based Technique for transliterating Meetei Mayek script to Bengali script and Bengali script to Meetei Mayek script of web-based Manipuri news text [28].

- ***Punjabi to English MTn system:*** The Punjabi to English MTn system was developed by Kamal Deep and Vishal Goyal at the Department of Computer Science, Punjabi University, Patiala, India in 2011. The MTn system has been developed using GBTM (RBM) approach for transliterating the common names from Punjabi to English language. The transliteration accuracy of the system was 93.22% [5].

- ***Punjabi to Hindi MTn system:*** The Punjabi to Hindi MTn system was developed by Gurpreet Singh Josan and Jagroop Kaur at the Department of Computer Science, Punjabi University, Patiala, India in 2011. The MTn system has been developed using DTT (Letter to letter mapping approach) and STM for transliterating OOV words from Punjabi to Hindi language [13].

IMPLEMENTATION OF ENGLISH TO BODO MACHINE TRANSLITERATION SYSTEM

The English to Bodo MTn system has been designed using HTML (Hyper Text Markup Language), JavaScript, CSS (Cascading Style Sheets), and PHP (Hypertext Preprocessor) as Front-end; MySQL and SQL as Back-end; Notepad++ as text editor, and WampServer. The MTn system has been developed using hybrid approach which is the combination of Direct Transliteration Technique and Sequential Search Technique.

Direct Transliteration Technique

The Direct Transliteration Technique (DTT) is a popular technique of GBTM. The GBTM is a process of mapping a grapheme sequence from a Source Language (SL) to a Target Language (TL) ignoring the phoneme level processes. In this technique, characters of SL are directly orthographic mapped to the characters of TL. The DTT is based on the direct mapping between the letters of SL and TL words. The DTT has been used to enter words or terms of English language and their corresponding transliterated words or terms of Bodo language into the English-Bodo transliteration database. It is a simplest and more accurate transliteration technique [1].

Architecture and Algorithm of DTT

The architecture and Algorithm of DTT in English to Bodo MTn system are briefly discussed as below:

Architecture: The architecture of DTT is shown in figure 2.

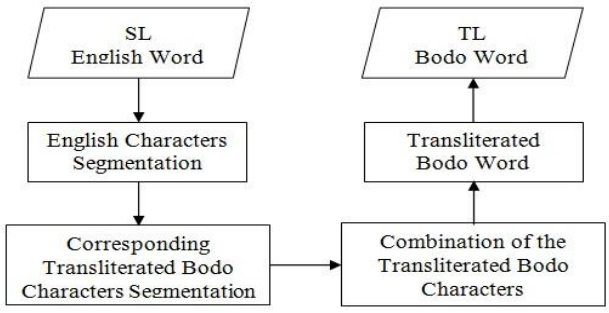


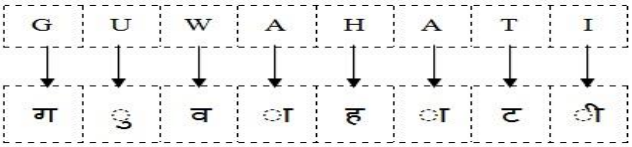
Figure 2: Architecture of DTT

Algorithm: Suppose the word "GUWAHATI" of English Language (SL) is to be transliterated into Bodo Language (TL), then the DTT will work as the following steps.

Step1: Segmentation of the SL word as below:

G	U	W	A	H	A	T	I

Step 2: Mapping the letters of the SL word with its corresponding transliterated letters of TL word as below:



Step3: Combination of the transliterated TL letters to form a word.

Step 4**:** Finally, "**गुवाहाटी**" is formed as a word of TL.

Letter Mapping

The letters or characters mapping between the English and Bodo languages in English to Bodo MTn system is shown in Table 3. In some cases, the characters mapping depends on the word of SL and the characters mapping may be changed. For example, characters mapping between English and Bodo languages is "Laptop ⟶ लैपटॉप".

Table 3: Characters mapping between the English and Bodo languages

English (En) to Bodo (Bd) Character Mapping											
En	Bd	En	Bd	En	Bd	En	Bd	En	Bd	En	Bd
a	अ	o	ओ	C	छ	Q	ौ	>	ं	1	१
b	ब	p	प	D	ध	R	ष	~	ँ	2	२
c	च	q	औ	E	े	S	श	!	ं	3	३
d	द	r	र	F	ै	T	थ	@	ः	4	४
e	ए	s	स	G	घ	U	उ	#	्	5	५
f	ऐ	t	त	H	ट	V	ढ	$	़	6	६
g	ग	u	ु	I	ि	W	ऊ	%	ृ	7	७
h	ह	v	ड	J	झ	X	ी	^	ॉ	8	८
i	इ	w	व	K	ख	Y	आ	,	,	9	९
j	ज	x	ई	L	ठ	Z	ऋ	`	`	0	०
k	क	y	य	M	ङ	&	इ	|	|	,	,
l	ल	z	ञ	N	ण	*	ढ़	?	?	.	.
m	म	A	ा	O	ो	+	ृ	=	ॅ	(	(
n	न	B	भ	P	फ	<	ॅ	_	ॐ	)	)
k+shift⁺+R(क्ष)		t+shift⁺+r (त्र)				j+shift⁺+z(ज्ञ)			S+shift⁺+r(श्र)		

Sequential Search Technique

The Sequential Search Technique (SST) has been used in the English to Bodo MTn system to look up or search the transliterated Bodo words or terms of the corresponding given English words or terms. The SST is the simplest and popular word search technique. It is a very useful and efficient search technique to look up the words from a database easily and quickly. If a user wants to search a particular word in a system, then the SST checks each word one by one in sequential order from the beginning of the database table until

the desired word is found in the system. The average number of comparisons in SST is (N+1)/2, where N is the size of the row in the table. If the searching word is in the 1st position, the number of comparisons will be 1 and if the word is in the last position, the number of comparisons will be N. Its worst-case cost is proportional to the number of elements in the list. The searching time for SST is O(n) [9].

Architecture of SST

The architecture of SST in the English to Bodo MTn system is shown in figure 3.

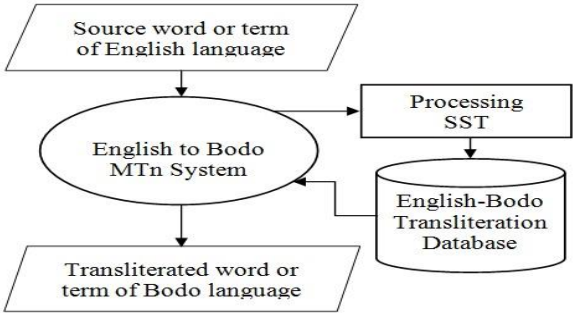


Figure 3: Architecture of SST

Suppose, a user wants to search an English word and it corresponding transliterated Bodo word in the MTn system, then the given word will compare with each existing word in the field (or column) of English language one by one in sequential order until the desired word is found in the table. If the given word is found, then the given word and its corresponding transliterated Bodo word will be displayed. Otherwise, the given word is not available in the system.

Algorithm of SST

The algorithm of SST in the English to Bodo MTn system is shown as below:

Step 1: Initialize sarray, sword, leng;
Step 2: Initialize pos=0;
Step 3: Repeat step 4 until pos<=leng
Step 4: if (sarray[pos])==sword)
return pos (Print sword is found);
else
pos=pos+1;
Step 5: if (pos>leng)
Print sword is not found;
Step 6: Stop

Where,

sarray= Search array (Column or field of the English language)

sword=Search word (Keyword, or searching word)

leng=Length (Number of words exist in the field of the English language in the database table)

pos=Position (Position of the word in the field of the English language in the database table)

Advantages of SST

The main advantages of SST are as follows:

i. The main advantage of SST is its simplicity.
ii. It is very simple to implement, easy to understand, and is straightforward.
iii. SST provides good performance in both the sorted and unsorted database tables.

Result and Discussion

The English to Bodo machine transliteration system has been developed for storing OOV words like proper nouns, abbreviations, and technical terms of Agriculture, Health, and Tourism domains and for improving the translation accuracy of the English to Bodo SMT system. A Bodo hard keyboard has been developed based on Unicode format for entering Bodo alphabets into the system with the help of English hard keyboard. The typing instructions and mapping of the Bodo and English characters are available in the MTn website. We have been also designed a soft (or virtual) keyboard for typing Bodo alphabets in the MTn system. The snapshot of the user interface of English to Bodo MTn system is shown in figure 4. Here, a user can search English words and their corresponding transliterated Bodo words easily.

Figure 4: Snapshot of the user interface of English to Bodo MTn system

The English to Bodo MTn system contains total 10450 English words or terms and their corresponding transliterated Bodo words or terms under Agriculture, Health, and Tourism domains. Some of the existing different domains English words or terms and their corresponding transliterated Bodo words or terms in English to Bodo MTn system are shown in Table 4.

Table 4: Existing words or terms in the English to Bodo MTn system

Name of the different domains English-Bodo parallel corpus	Input words or terms [English Script]	Output words or terms [Bodo Script]
Agriculture domain	Urea	इउरिया
	FAO	एफ.ए.अ.
Health domain	Vitamin A	भिटामिन ए
	ORS	अ.आर.एस.
Tourism domain	ASTC	ए.एस.टि.सि.
	7Km	७कि:मि:

ENGLISH TO BODO STATISTICAL MACHINE TRANSLATION SYSTEM

English to Bodo SMT system has been developed using Phrase-based SMT approach with the help of Agriculture, Health, and Tourism domains English-Bodo parallel text corpora. The PB-SMT approach is an accurate and deeply used by many research communities all over the world. It can produce high-quality translation result using an enormous amount of aligned parallel text corpus in both the source and target languages [11]. The primary components of SMT approach include language model, translation model, and decoder. The SMT system has been developed individually for each domain parallel corpus. The Agriculture, Health, and Tourism domains English-Bodo parallel text corpora contain 3500 (Three thousand five hundred), 12000 (Twelve thousand), and 9000 (Nine thousand) parallel sentences of each English and Bodo language respectively.

The complete architecture of English to Bodo SMT system is shown in figure 5.

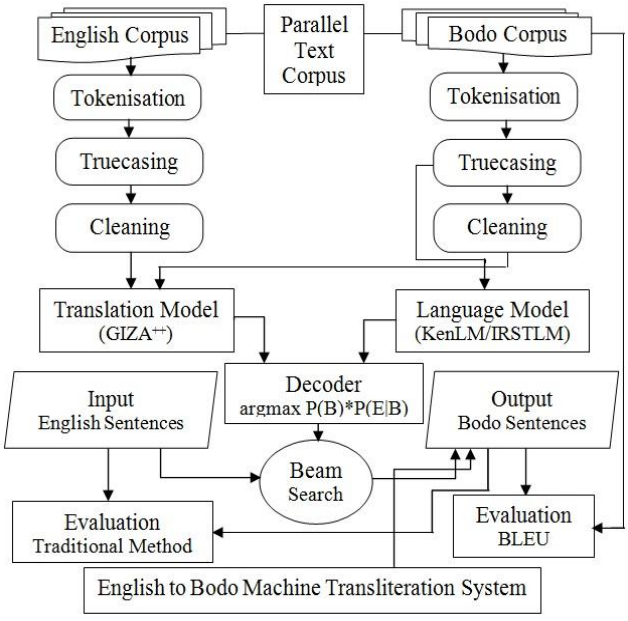


Figure 5: Complete architecture of English to Bodo SMT system

The following operations have been performed to develop the English-Bodo SMT system using PB-SMT approach and Moses.

Corpus Pre-processing and Preparation: Corpus pre-processing and preparation are very much essential task to train the SMT system using Moses. The English to Bodo MT system has been developed individually for each domain English-Bodo parallel text corpus namely Agriculture, Health, and Tourism. First, the SMT system has been developed using Tourism domain E-BPTC. Since the Tourism domain has been collected from TDIL as MS Excel format. Therefore, to train the English to Bodo SMT system using Moses, we have prepared 9000 parallel sentences of each English and Bodo language form the Tourism domain E-BPTC and two separate text files have been created in UTF-8 format for English and Bodo corpus on Linux operating system. We have also checked and maintained the alignment of the parallel sentences in the English and Bodo corpora. In the similar way, we have been prepared two separate text files for English and Bodo corpus from each domain (Agriculture and Health) E-BPTC to train the SMT system. After that, Tokenization, True Casing, and Cleaning have been performed to build the language model and translation model for each domain parallel text corpus separately [17].

Language Model: The Language Model (LM) has been built using KenLM and IRSTLM toolkits to compute the probability of the Bodo sentences. The LM is used to ensure the fluency of the translated Bodo sentences in the system.

Translation Model: The Translation Model (TM) has been developed using Giza++ toolkit to compute the probability of the English and Bodo sentences. Giza++ has been used for word or phrase alignment. The TM is used to ensure the adequacy of the translation result in the system.

Decoder: The decoder can find the maximum translation probability from the English sentences into the corresponding translated Bodo sentences with the help of beam search technique. The decoder finds the translation probability using the following Eq. (1):

$$P(E, B) = \text{argmax } P(B) * P(E|B) \quad (1)$$

Where, P (B) and P (E|B) are the output results obtained from the LM and TM respectively.

Result and Evaluation

The English to Bodo SMT system has been examined several times with various numbers of Agriculture, Health and Tourism domains English-Bodo parallel corpora individually. It has been noticed that the translation result can be enhanced by increasing the number of parallel sentences in each domain parallel corpus. Finally, the number of parallel sentences of each domain E-BPTC has been used for training, tuning, and testing the SMT system is shown in Table 5. It has been also

noticed that a few numbers of unknown words or OOV words like abbreviations, proper nouns, and technical terms are available in English scripts in every domain translated Bodo corpus which may be reduced the translation result. Therefore, we have been converted the unknown words which exist in every domain translated Bodo corpus in English script into phonetically equivalent Bodo script manually with the help of English to Bodo MTn system for better translation result. Though the manual transliteration is a time-consuming task, still we rely on manual transliteration because manual transliteration gives us accurate transliteration result.

The accuracy of the translation results of the English to Bodo SMT system has been evaluated individually for each domain E-BPTC using BLEU (Bilingual Evaluation Understudy) method. It is the best method to evaluate the accuracy of the translation result in any MT system [23]. The BLEU score has been determined from the reference Bodo corpus (human translated corpus) and the candidate Bodo corpus (machine translated corpus). The BLEU scores which have been achieved for each domain E-BPTC in English to Bodo SMT system before using and after using the English to Bodo MTn system are shown in Table 5.

Table 5: BLEU scores and corpus statistics of the SMT system

Multi-domain English-Bodo Parallel Text Corpora	Corpus Statistics (Sentences)			BLEU scores	
	Training	Tuning	Testing	Before using the MTn system	After using the MTn system
Agriculture	3500	500	3500	30.18	31.92
Health	12000	1000	12000	38.87	40.08
Tourism	9000	1000	9000	37.50	38.35

From the above BLEU scores, it can be claimed that transliteration can improve the translation result in any SMT system, because a higher BLEU score represents better translation result.

CONCLUSION AND FURTHER WORK

Machine translation and machine transliteration are two totally distinct and the most important applications of CL and NLP universally nowadays. Machine transliteration is a very helpful technique in SMT system for handling unknown words or OOV words like abbreviations, proper nouns, and technical terms. There are several techniques available for machine transliteration, but the English to Bodo MTn system has been developed using hybrid approach. The English to Bodo SMT system has been developed using PB-SMT approach and English to Bodo MTn system with the help of Agriculture, Health, and Tourism domains English-Bodo parallel text corpora. Since Bodo is a low resource language and not much work has been done on machine transliteration as well as machine translation. Therefore, it can be expected that the English to Bodo MTn system and the SMT system would be beneficial for the people of India and abroad.

The proposed research work can be extended by adding more number of English-Bodo parallel abbreviations, proper nouns, and technical terms under Agriculture, Health, and Tourism domains in the English to Bodo MTn system for better translation result in the English to Bodo SMT system.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ismail Hussain, Assistant Professor, Department of Bodo, Gauhati University, Guwahati, India and Mr. Dwipen Baro, Assistant Professor, Department of Bodo, Dhamdhama Anchalik College, Nalbari, India for their great support, suggestion, and help to develop both the English to Bodo MTn and SMT systems.

REFERENCES

[1] Aadil, M. and Asger, M., "English to Kashmiri Transliteration System: A Hybrid Approach", International Journal of Computer Applications, 162(12), 2017.

[2] Antony, P. J. and Soman, K. P., "Machine Transliteration for Indian Languages: A Literature Survey", International Journal of Scientific and Engineering Research, 2(12), 2011.

[3] Antony, P. J., Ajith, V. P., and Soman, K. P., " Statistical Method for English to Kannada Transliteration", In: Das V. V. et al. (eds) Information Processing and Management. Communications in Computer and Information Science, Springer, Berlin, Heidelberg, vol. 70, 2010.

[4] Das, A., Ekbal, A., Mandal, T., and Bandyopadhyay, S., "English to Hindi Machine Transliteration System at NEWS 2009", Proceedings of the Named Entities Workshop-2009, ACL-IJCNLP 2009, Suntec, Singapore, pp. 80–83, 2009.

[5] Deep, K. and Goyal, V., "Development of a Punjabi to English Transliteration System", International Journal of Computer Science and Communication, 2(2), pp. 521-526, 2011.

[6] Devi, M. P., Singh, I. T., and Devi, H. M., "English to Manipuri machine transliteration system based on syllabification", Journal of global research in computer science, 8(8), 2017.

[7] Goto, I., Kato, N., Ehara, T., and Tanaka, H. "Back transliteration from Japanese to English using target English context", In the proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, pp. 827–833, 2004.

[8] Hettige, B. and Karunananda, A., "Transliteration System for English to Sinhala Machine Translation", In the proceedings of Second International Conference on Industrial and Information Systems (ICIIS2007), IEEE, Sri Lanka, 2007.

[9] Islam, S., "An English to Assamese, Bengali and Hindi multilingual E-dictionary", International Journal of Current Engineering and Scientific Research, 3(9), pp.74-80, 2016.

[10] Islam, S., Devi, M. I., and Purkayastha, B. S., "A Study on Various Applications of NLP Developed for North-East Languages", International Journal on Computer Science and Engineering, 9(6), pp. 368-378, 2017.

[11] Islam, S. and Purkayastha, B. S., "English to Bodo Phrase-Based Statistical Machine Translation", Advanced Computing and Communication Technologies. Advances in Intelligent Systems and Computing, Springer, Singapore, vol. 562, pp. 207-217, 2018.

[12] Jaleel, N. A. and Larkey, L. S., "Statistical Transliteration for English-Arabic Cross-Language Information Retrieval", In the proceedings of the 12th International Conference on Information and Knowledge Management, pp. 139-146, 2003.

[13] Josan, G. S. and Kaur, J., "Punjabi to Hindi Statistical Machine Transliteration", International Journal of Information Technology and Knowledge Management, 4(2), pp. 459-463, 2011.

[14] Karimi, S., Scholer, F., and Turpin, A., "Machine Transliteration Survey", ACM Computing Surveys, 43(3), pp. 1-46, 2011.

[15] Kaur, K. and Singh, P., "Review of Machine Transliteration Systems", International Journal of Engineering Research and Technology, 3(5), 2014.

[16] Kaur, V., Sarao, A. K., and Singh, J., "Hybrid Approach for Hindi to English Transliteration System for Proper Nouns", International Journal of Computer Science and Information Technologies, 5(5), pp.6361-6366, 2014.

[17] Koehn, P., "MOSES (User Manual and Code Guide)", Statistical Machine Translation System, University of Edinburgh, UK, 2016.

[18] Kumaran, A., Mitesh, M. K., and Bhattacharyya, P., "Compositional Machine Transliteration", ACM Transactions on Asian Language Information, 2010.

[19] Lee, J. S. and Choi, K. S., "English to Korean Statistical Transliteration for Information Retrieval", Computer Processing of Oriental Languages, pp.17-37, 1998.

[20] Narzary, M., "English to Bodo Student's Dictionary (Printed)", Published by Nilima Prakashani, Baksa, BTAD (Assam), India, 2015.

[21] Nejad, N. M., Khadivi, S., and Taghipour, K., "The Amirkabir Machine Transliteration System for NEWS 2011: Farsi-to-English Task", In the proceedings of the Named Entities Workshop-2011, IJCNLP 2011, pp. 91–95, Chiang Mai, Thailand, 2011.

[22] Oh, J. H., Choi, K. S., and Isahara, H., "A Comparison of Different Machine Transliteration Models", Journal of Artificial Intelligence Research, pp.119–151, 2006.

[23] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., "BLEU: A Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318, 2002.

[24] Paul, M., Finch, A., and Sumita, E., "NICT@WMT09: Model Adaptation and Transliteration for Spanish-English SMT", Proceedings of the 4th EACL Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece, pp.105–109, 2009.

[25] Qin, Y. and Chen, G., "Forward-Backward Machine Transliteration between English and Chinese Based on Combined CRFs", In the Proceedings of the 2011 Named Entities Workshop, IJCNLP, pp. 82–85, Chiang Mai, Thailand, 2011.

[26] Salam, K. M. A., Yamada, S., and Nishino, T., "How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration", Journal of Computers, 8(5), pp.1167-1174, 2013.

[27] Sasidharan, S., Loganathan, R., and Soman, K. P., "English to Malayalam Transliteration using Sequence Labeling Approach", International Journal of Recent Trends in Engineering, 1(2), 2009.

[28] Singh, T. D., "Bidirectional Bengali Script and Meetei Mayek Transliteration of Web-Based Manipuri News Corpus", In the proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING 2012, Mumbai, pp. 181–190, 2012.

[29] Sunitha, C. and Jaya, A., "A Phoneme-based Model for English to Malayalam Transliteration", International Conference on Innovation Information in Computing Technologies (ICIICT), IEEE, Chennai, India, 2015.

[30] Vijaya, M. S., Loganathan, R., Shivapratap, G., Ajith, V. P., and Soman, K. P., "English to Tamil Transliteration using Sequence Labeling Approach", International Conference on Asian Language Processing, Thailand, 2008.

[31] Yaser, O. and Knight, K., "Machine Transliteration of Names in Arabic Text", In the proceedings of the Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages, 2002.