IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Bodo To English Statistical Machine Translation System

Maheswar Daimary[1*], Shikhar Kumar Sarma[2], Mirzanur Rahman[3]

[1,2,3] Department of Information Technology, Gauhati University, Guwahati, India

**Corresponding Author: maheswardaimary45@gmail.com, Tel.: 919678585077*

DOI: https://doi.org/10.26438/ijcse/v7i5.17311736 | Available online at: www.ijcseonline.org

Accepted: 21/May/2019, Published: 31/May/2019

Abstract— Machine Translation (MT) is widely considered among the most difficult tasks and applications in the area of Natural Language Processing (NLP) and Computational Linguistics (CL). MT is the method of translating text from source language to target language using computer. The main objective of this proposed research work is to develop Bodo to English machine translation system for enhancing the translation result of Bodo to English Statistical Machine Translation (SMT) System by taking their respective parallel corpus. Here a statistical machine translation engine Moses is used to train statistical models of text translation from source language to a target language. We also used IRSTLM tool to develop the language model and GIZA++ tool to align the words respectively.

Keywords—Bodo Language, English Language, Machine translation, Moses, Corpus

I. INTRODUCTION

Machine Translation (MT) is fully automated translation that can translate one natural language (source language) into another (target language) without human intervention. Machine Translation is one of the most heavily research tasks in Natural Language Processing (NLP) and Computational Linguistics (CL). Nowadays, the MT is a very difficult research task in NLP and the demand of it is growing in the world, especially in India [1]. Lots of MT systems have been developed in India as well as all over the world using several pairs of major natural languages, such as English to (Urdu, Bengali, Chinese, French, Hindi, Japanese, Spanish, and Arabic). There are many approach to machine translation, one of the most significant and newest approach is the Statistical Machine Translation. Statistical Machine Translation is an approach to machine translation where translation from one language to another is generated on the basis of analyzing a huge amount of parallel corpus and corpora. The main aim of this proposed research work is to develop Bodo to English machine translation system for enhancing the translation result of Bodo to English Statistical Machine Translation (SMT) System [2]. Bodo is one of the major spoken languages in the North-East region of India. Though a considerable amount of work has already been done in different Indian languages in the field on NLP, still not much work has been done, especially on MT system for Bodo language due to the lack of a comprehensive set of parallel corpora. The SMT approach uses a huge amount of bilingual aligned parallel text corpora in both the source and target languages to attain high quality translation result [3]. The accuracy (adequacy and fluency) of the translation results in SMT system directly depend on the size and quality of a parallel corpus of a particular language pair. The SMT approach offers the best solution to ambiguity problem. The main benefits of SMT approach are: it is easy to build and maintain, less linguistic knowledge required, and reduces human efforts [4].

Bodo Language: Boro language is a Sino-Tibetan family's language and one of the major spoken languages of North-East of India, Nepal and Bengal. It is mainly spoken by the people of Assam in some districts of Kokrajhar, Chirang, Baksa and Udalguri. Bodo language is the official language of Bodoland Autonomous region (Assam) and is also one of the 22 recognized languages of India [5]. In 1963, the Bodo language was introduced in Assam as a medium of instruction in the primary school in some Bodo major dominated areas and it was the result of an intense political movement launched by the different Bodo organizations since 1913. It is written using Devanagari script and it has a total of 30 phonemes, 6 vowels, 16 consonants, and 8 dipthongs. The word order of this language is subject+verb or subject+verb+object. The Albhabets of Bodo Language is shown below-

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	ळ	ऱ	ऴ
	pure consonants								

English Language: English is the West Germanic language. It was the first spoken language in early medieval England. Now, it is an international language of the world and is the third most common native language in the world. The English language is mainly spoken by the people of Canada, Australia, United Kingdom, United States, Ireland and etc. In 1830, during the rule of the East India Company, English language was introduced in India. English is the associate official language of India as the declared of the constitution of India in 1951. It is written using Latin script and it contains 26 alphabets including 5 vowels and 21 consonants [6]. The Alphabets of English Language is shown below-

Table 2. *Alphabets of English Language.*

English Alphabets										
Letters	**Uppercase**					**Lowercase**				
Vowels	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				

II. Related Work

We have done a number of literature review throughout our work. A number of journal papers as shown below are found important for our work.

Assamese to English Statistical Machine Translation Integrated with a Transliteration Module: The Assamese to English MT system was developed by Pranjal Das and Kalyanee K. Baruah at the Department of Information Technology, Gauhati University, India in August 2014 [7]. In this paper, it is described how an Assamese sentence is translated to English using SMT. Here Moses is used as a platform for Statistical Machine Translation. GIZA++ is also used for word alignment and IRSTLM for language model training. They collected around 8000 Assamese/English parallel corpus from Tourism domain and obtained BLEU score is 11.32.

Assamese-English Bilingual Machine Translation: The Assamese-English Bilingual MT system was developed by Kalyanee K. Baruah, Pranjal Das, Abdul Hannan, Shikhar Kr. Sarma at the Department of Information Technology, Gauhati University, India in August 2014 [6]. In this paper, both Assamese to English and English to Assamese machine translation and transliteration is done with their respective parallel corpus and obtain BLEU score (Assamese/English) is 9.72 and (English/Assamese) is 5.02.

English to Hindi Machine Transliteration System: The English to Hindi Machine Transliteration (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 [8]. 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.

Punjabi to English Machine Transliteration system: The Punjabi to English Machine Transliteration (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% [9].

English to Manipuri Machine Transliteration system: The English to Manipuri Machine Transliteration (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 and the machine transliteration was performed from English to Manipuri Language [10].

English to Arabic Machine Transliteration system: The English to Arabic Machine Transliteration (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 [9].

Machine Translation System in Indian Perspectives: In this paper, Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve was described different types of machine translation projects as given below:

- Anusaaraka: Anusaaraka was started in 1995 at IIT Kanpur with the aim of translation from one Indian language to another. It gives translation from Telegu, Kannada, Bengali, Punjabi, and Marathi to Hindi [10].
- Anglabharati: Anglabharati is a pattern directed rule based system developed in IIT Kharagpur [12]. It produces translation between English to Hindi and was developed for a public Heath Campaigns. Anglabharati uses the rule based pseudo-interlingua approach for translation [10].
- Shiva and Shakti Machine Translation: Shiva and Shakti are two machine translation systems for English to Hindi. Shiva is an Example-based machine translation system while Shakti is a hybrid system composed of Rule-based and Corpus-based approaches [10].
- MaTra: MaTra is an indicative English to Hindi machine translation System which is fully automatic. The approach taken by MaTra is transfer-based [10].

III. TOOLS FOR IMPLEMENTING THE SYSTEM

MOSES: Moses is a free software Statistical Machine Translation System designed and developed by Philipp Koehn and Hieu Hoang at University of Edinburgh. It allows automatically train translation models of text translation from a source language to a target language. Moses requires a parallel corpus or translated texts of passages in the two languages (for e.g. Bodo and English) that are used in training the system [6]

GIZA++: GIZA++ is part of the Statistical Machine Translation word alignment toolkit used to develop the translation model of our system. It is an extension of the program GIZA, designed by Franz Josef Och. Giza++ is used to train different models like HMM (Hidden Markov Model) [11].

IRSTLM: IRSTLM is a language modeling tool that is used to develop the language model. The IRSTLM toolkit features algorithms and data structures suitable to estimates, represents and computes statistical languages models. It is used for improving the language model for the target language. It is open source [12].

BLEU: The BLEU (Bi-Lingual Evaluation Understudy) toolkit helps us to determine the quality of our translation system which has been machine-translated from one natural language to another. We can test how good our translation is by translating the text and then running the BLEU script on it [13].

IV. METHODOLOGY

Bodo to English Statistical Machine Translation system has been developed using Phrase-based SMT approach with the help of Bodo-English parallel text corpora. The Phrase-based-SMT approach is an exact and deeply used by many research communities all over the world. It can produce high-quality translation result using a huge amount of aligned parallel text corpus in both the source and target languages [14]. It has been noticed that the translation result can be enhanced by increasing the number of parallel sentences in each domain parallel corpus. Statistical Machine translation depends on huge amount of parallel corpus of source and target language pair of sentences. We have trained our system respectively with 8000 Bodo and English parallel sentences. When we increase the amount of corpus more amount of data leads to more accuracy in the translations. The main components of SMT approach include language model, translation model, and decoder. The SMT system has been developed individually for each domain parallel corpus. The following procedures have been performed to develop the Bodo-English SMT system using Phrase Based-SMT approach and Moses. The complete architecture of Bodo to English SMT system is shown in figure 1.

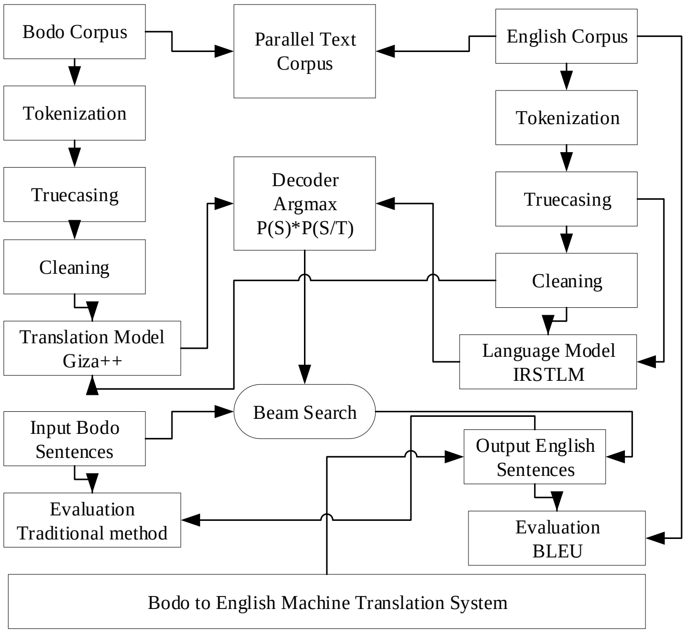


Figure 1.*Complete architecture of Bodo to English SMT System.*

Corpus pre-processing and preparation: A corpus is a collection of written text in linguistics, which is used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. The word ‘corpus’ comes from the Latin word ‘body’ and its plural form is corpora. The corpus preparation is a very essential task to train the SMT system. It can be considered as the primary resource for any linguistic analysis and NLP research and also can be classified into the following categories: Written corpus, Spoken corpus, General corpus, Monolingual corpus, Parallel corpus, Multilingual corpus, Learner corpus, Comparable corpus, Specialized corpus, Monitor corpus, and Annotated corpus. A parallel corpus consists of two or more monolingual corpus in one or more languages with their translation into another language that has been stored in the digital format. In the parallel text corpus, the texts of one corpus are the translation of another corpus. The order of the translation may be sentence by sentence, phrase by phrase, and word by word and the sentences, phrases, and words are needed to be aligned and matched, so that a user can find potential equivalents in each language and can investigate differences between the languages [16]. The process of parallel text corpus construction can be divided into three phases, namely translation, validation and sentence alignment. The Bodo and English sentences have been separated from the two domains parallel text corpora and created two separate files for Bodo and English languages. After that, the following steps have been performed on the Bodo and English files to build the language and translation models for every domain parallel text corpus: i) Tokenization (Performed to insert space between the words and punctuation), ii) True Casing (Performed to convert the first words of each sentence to their most probable case), and iii) Cleaning (Performed to remove the empty sentences and extra spaces). A sample of the corpus is shown in the following table.

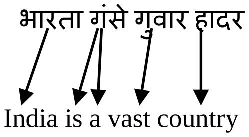
Table 3.*Bodo-English Parallel Corpus*

Bodo Sentences	**English Sentences**
भारत हादोरनि गलापि नोगोर मुडै मिथिजानाय मुंदांखा जयपुरा राजसथान रायजोनि राजथावनि	Jaipur, popularly known as the pink city, is the capital of rajasthan state, india
जयपुरा मारबलनि मुसुखा,गोथां दो-दोखोला, आरो राजसथाननि जथानि थाखायबो मुंदांखा	Jaipur is also famous for marble statues, blue pottery and the rajasthani shoes
जयपुर हाथाइनि नायगिदिंनायनि बयनिखृ◌ुयबो मोजांसिन समा जाबाय अक्टबरनिफृाय मारससिम	The best time to visit jaipur city is october to march
कनक बृन्दाबनआ जाबाय जयपुरनि गंसे मुंदांखा लावखार ओंखाम जानाय थावनि	kanak vrindavan is a popular picnic spot in jaipur
बुहुमनां दाब दाब जायगानि दावबायारिफोरनि नोजोर बोनो हानाय समायना खरं, महल आरो समायना बिलोमाफोरनि थाखाय बे नोगोरा मुंदांखा	The city is famous for its majestic forts, palaces and beautiful lakes which attract tourists from all over the world
राजस्थान राइजो हान्थामेला निगम (RSTC) आ साहा भारतनि गासैबो गाहाय गाहाय जायगासिम बास खारहोयो	The Rajasthan State Transport Corporation (RSTC) has bus services to all the major destinations of north India
बाहायजाथाव मोननैसो मुवाफोर बायनो हाथाव जयपुरनि गाहाय गाहाय हाथाइफोरा जाउह्रि बाजार, बापु बाजार, नेहरु बाजार, चाउरा रास्ता, त्रिपलिया बाजार आरो एम आइ रडाव दं	The main markets of Jaipur, from where you can buy some useful items are along Jauhri Bazar, Bapu Bazar, Nehru Bazar, Chaura Rasta, Tripolia Bazar and M.I. Raod

Language Model (LM): The purpose of the Language Model (LM) is to develop the most fluent output by calculating the probability of the target sentence. Here by using IRSTLM tool Language model is developed. The IRSTLM is suitable to estimates, represents and computes statistical languages models [9].

Translation Model (TM): The Translation Model (TM) computes the similarity between input and output. The Translation Model produces the probability of source sentence ‘S’, for a given target sentence ‘T’ i.e. P(S|T) by assigning probabilities to both source and target sentences. The Translation Model has been developed using Giza++ toolkit. Computation of Translation Model probabilities at sentence level is quite impossible, so the sentence is broken down into smaller units for e.g. words or phrases and their probabilities are learned [15]. For example, using the notation (T|S) to represent an input sentence S and its translation T, a sentence is translated as given below-

(भारता गंसे गुवार हादर | India is a vast country)



Decoder: Decoding is the process of finding a target translation sentence (Bodo/English) for a source sentence (English/Bodo) using translation model and language model. The output obtained from Language Model and translation Model is fed to the decoder and the decoder maximizes translation probability from the Bodo sentences into the corresponding translated English Sentences. The decoder finds the translation probability using the following Eq. (1):

$$P(S,T)=\text{argmax}P(T)*P(S|T) \quad \text{…………….(1)}$$

Where, P (T) and P(S|T) are the output results obtained from the both LM and TM respectively.

Preparing Data For Training: Training data has to be provided sentence aligned (one sentence per line), in two files, one for the Bodo sentences, one for the English sentences. Before training the data, we need to perform some tasks. We need to prepare the data for training the translation system. The following steps are performed.

- Tokenizing: We have to perform tokenize the data, so that spaces are inserted between words and punctuation [6].
- Truecasing: Truecasing does not apply for Bodo language because Bodo scripts do not have a distinction between uppercase and lowercase letters [6].
- Cleaning: Then we need to have clean the data i.e. removes empty lines, removes redundant space characters, extra spaces, drops lines (and their corresponding lines), that are empty, too short, too long [6].

Now, we can train our translation model. After training is done, a moses.ini file is generated which is used in running decoder. By tuning the data we can improve our translation system. Tuning the system again results in a moses.ini file. Then the file is ready to run the decoder [6].

V. RESULTS AND DISCUSSION

The Bodo to English SMT system has been examined several times with various numbers of Travel and Tourism domains Bodo-English 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. We have seen that whenever we increase the quantity of our corpus, the quality of the translation is also improved. The system is tested with a corpus of about 8000 sentences in bilingual text corpus. The results obtained from the final training with 8000 sentences are shown in the following table:

Table 4.*Translation result from Bodo sentence to English sentence*

Bodo Sentences	English Sentences
भारता गंसे गुवार हादर	India is a vast country
जयपुरा राजस्थाननि गंसे मुंदांखा नोगोर	Jaipur is one of the famous city of Rajasthan
रामा सासे मोजां मानसि	Ram is a good man
भारता जोंनि हादर	India is our country
जोंनि आयेन फरायसालि जोबोर समायना	Our law school is very beautiful
हाजोआ जोबोर समायना	Hills is very beautiful
दिल्लीआ भारतनि राजथावनि	Delhi is the capital of India
रामा सासे मोजां राजा	Ram is a good king
बियो ओंखाम जायो	He eats rice
मुंदांखा लावखार ओंखाम जानाय थावनि	Popular picnic spot

VI. CONCLUSION AND FUTURE SCOPE

This paper presented an approach of building a Statistical Machine Translation System that would give translated text from Bodo to English. Among the many approaches of Machine Translation, Statistical Machine Translation is the most widely used. Many of the researches on Machine Translation in India are Statistical-based. Statistical based systems require significant amount of corpus to achieve good translations. We have used a very small amount of data (about 8000 parallel sentences) to train our system. But this statistic is very small compared to better translation system, the size of parallel corpus should be large. There are not enough parallel corpora available between Bodo and English. As we are working on phrase-based machine translation, increasing the number of sentences will increase the quality of the translated text. We will work on this problem and try to increase the amount of corpus into our system, so that we can enhance our translation result. Also we will try to improve the transliteration module while dealing with (Out of Vocabulary) OOV words and also try to build both Bodo to English and English to Bodo translation using same parallel corpus.

ACKNOWLEDGEMENTS

The authors are thankful to the Department of Information Technology, Gauhati University for providing us the corpus, which helped us in building the SMT system and also the Department of Natural Language Processing, Gauhati University for their immense support.

REFERENCES

[1] Ananthi Sheshasaaye, Angela Deepa. V.R, *“The Role of Morphological Analyzer and Generator for Tamil language in Machine Translation Systems”*, International Journal of Computer Science Engineering Volume-**2**, Issue-**5**, May **2014**.

[2] Peter F. Brown et al., *“A Statistical Approach to Machine Translation”* Computational Linguistics Volume **16**, Number **2**, June **1990**.

[3] Gurpreet Singh Josan & Jagroop Kaur (2011)*’ Punjabi to Hindi Statistical Machine Translaiteration’*, International Journal of Information Technology and Knowledge Management, Volume **4**, No. **2**, pp. **459-463** July-December **2011**.

[4] Saiful Islam, Bipul Syam Purkayastha, *“English to Bodo Machine Transliteration System for Statistical Machine Translation”,* International Journal of Applied Engineering Research ISSN **0973-4562** Volume **13**, pp. **7989-7997** Number **10**, **2018**.

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

[6] Kalyanee K. Baruah, Pranjal Das, Abdul Hannan, Shikhar Kr. Sarma *“Assamese-English Bilingual Machine Translation”* International Journal on Natural Language Computing (IJNLC) Vol. **3**, No**. 3**, June **2014**.

[7] Pranjal Das and Kalyanee K. Baruah *“Assamese to English Statistical Machine Translation Integrated with a Transliteration Module”* International Journal of Computer Applications (0975 – 8887) Volume **100**– No.**5**, August **2014**.

[8] 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**.

[9] 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**.

[10] Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve, *“Machine Translation System in Indian Perspectives”*, Journal of Computer Science, Volume **6**, Issue **10**, pp. **1111-1116**.

[11] D. D. Rao, *“Machine Translation A Gentle Introduction”*, RESONANCE, July **1998**.

[12] Philipp Koehn et al, *“Moses: Open Source Toolkit for Statistical Machine Translation”*, In the Proceedings of the ACL, June **2007**, pp. **177-180**.

[13] Md. Zahurul Islam, “*English to Bangla Statistical Machine Translation”*, Master Thesis, Universitat des Saarlendes, August **2009**.

[14] N.Sharma, P.Bhatia, V.Singh, *“English to Hindi Statistical Machine Translation”*, June **2011**.

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

[16] Biswajit Brahma, Anup Kr Barman, Shikhar Kr. Sarma, Bhatima Boro*, “Corpus Building of Literary Lesser Rich Language-Bodo: Insights and Challenges”,* Conference: Proceedings of the 10th Workshop on Asian Language Resources, December **2012**.

Authors Profile

Maheswar Daimary

Maheswar Daimary is an M.Tech student in the Department of Information Technology, Gauhati University, Assam, India. He had done his B.Tech from Central Institute of Technology, Kokrajhar. His area of interest is Natural Language Processing.

Prof. Shikhar Kumar Sarma

Shikhar Kumar Sarma is a professor in the Department of Information Technology, Gauhati University, Assam, India. His area of research is Natural Language Processing, AI, and Language Technology in Assamese and Bodo languages.

Mirzanur Rahman

Mirzanur Rahman is currently working as assistant professor, Depament of Information Technology, Gauhati Univeristy, Guwhati, Assam,India-781014.