Construction of English-Bodo Parallel Text Corpus for Statistical Machine Translation

Article *in* International Journal on Natural Language Computing · October 2018

DOI: 10.5121/ijnlc.2018.7509

4 authors, including:

Saiful Islam
Sonargaon university
18 PUBLICATIONS **98** CITATIONS
SEE PROFILE

Bipul Syam Purkayastha
Assam University
47 PUBLICATIONS **273** CITATIONS
SEE PROFILE

CONSTRUCTION OF ENGLISH-BODO PARALLEL TEXT CORPUS FOR STATISTICAL MACHINE TRANSLATION

Saiful Islam[1], Abhijit Paul[2], Bipul Syam Purkayastha[3] and Ismail Hussain[4]

[1,2,3]Department of Computer Science, Assam University, Silchar, India
[4]Department of Bodo, Gauhati University, Guwahati, India

ABSTRACT

Corpus is a large collection of homogeneous and authentic written texts (or speech) of a particular natural language which exists in machine readable form. The scope of the corpus is endless in Computational Linguistics and Natural Language Processing (NLP). Parallel corpus is a very useful resource for most of the applications of NLP, especially for Statistical Machine Translation (SMT). The SMT is the most popular approach of Machine Translation (MT) nowadays and it can produce high quality translation result based on huge amount of aligned parallel text corpora in both the source and target languages. Although Bodo is a recognized natural language of India and co-official languages of Assam, still the machine readable information of Bodo language is very low. Therefore, to expand the computerized information of the language, English to Bodo SMT system has been developed. But this paper mainly focuses on building English-Bodo parallel text corpora to implement the English to Bodo SMT system using Phrase-Based SMT approach. We have designed an E-BPTC (English-Bodo Parallel Text Corpus) creator tool and have been constructed General and Newspaper domains English-Bodo parallel text corpora. Finally, the quality of the constructed parallel text corpora has been tested using two evaluation techniques in the SMT system.

KEYWORDS

Bodo Language, Corpus, English Language, Statistical Machine Translation

1. INTRODUCTION

Machine translation is an important application in the field of Computational Linguistics and NLP whose aim is to translate texts from one natural language to another natural language in an automatic fashion. Nowadays, the MT is a very challenging research task in NLP and the demand of it is growing in the world, especially in India. 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 (Arabic, Bengali, Chinese, French, Hindi, Japanese, Spanish, and Urdu). 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. Therefore, it has been decided to construct General and Newspaper domains English-Bodo Parallel Text Corpora (E-BPTC) to develop the English to Bodo SMT system using Phrase-Based SMT approach.

DOI: 10.5121/ijnlc.2018.7509

In this section, Bodo language, English languages, corpus, and SMT approach are also briefly discussed.

1.1 Bodo Language

The Bodo language is one of the recognized languages of India and the co-official language of Assam (Bodoland Territorial Council). The word 'Bodo' is also pronounced as Baro and denotes both the Bodo language and Bodo community. The Bodo language is one of the major spoken languages of North-East India and belongs to Sino-Tibetan language family [13]. It is mainly spoken by the maximum population of Kokrajhar, Chirang, Baksa, and Udalguri districts of Assam, India [14]. It is also spoken by some people of West Bengal and Nepal. Bodo is a tonal language and has two kinds of tone, namely Low and High [3]. The language is written using Devanagari script. However, earlier it was written using Assamese script. There are many ambiguous words in Bodo language and the word order of the language is SOV (Subject+Object+Verb). Bodo is rich, ancient and divergent natural language. It has not sufficient written literature, web information and corpus.

1.2 English Language

The English language is an Indo-European language, widely used by native speakers as well as non-native speakers all around the world, thus it is also known as international natural language as well as lingua franca. English was the first spoken language in England and is now third most widely used language all around the world [12, 14]. It is an official language of sixty seven countries and is the most commonly spoken language in Australia, Canada, Ireland, New Zealand, United Kingdom and the United States. The English language was introduced in India in 1830 during the rule of the East India Company. In 1951, the Constitution of India declared English as the associate official language of India. At present, English is the third most spoken language in India. The language is written using Latin script. There are many ambiguous words in the language and the word order of the language is SVO (Subject + Verb +Object). Unlike Bodo, the English language has sufficient written literature, web information and corpus.

1.3 Corpus

Corpus is an immense systematic collection of homogeneous and authentic written texts (or speech) of a particular natural language which exists in digital form. The word 'corpus' comes from the Latin word 'body' and its plural form is corpora [6]. The scope of a corpus is very vast and can be considered as the primary resource for any linguistic analysis and NLP research. The quality and structure of a corpus can directly influence the performance of various NLP tasks like Machine Translation, Spelling Checker, Grammar Checker, Speech Recognition, Text-to-Speech Synthesis, Part of Speech Tagging, Electronic Dictionaries, Text Summarization, Word Sense Disambiguation, WordNet, Text Annotation, and Information Retrieval [12]. A text corpus can provide better descriptions about a natural language to authors, grammarians, lexicographers, and other interested people. The corpus 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, Historical corpus, Reference corpora, Multimedia corpus, Annotated corpus, and Un-annotated corpus [7].

A parallel corpus consists of two or more monolingual corpus in one or more language(s) 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. A parallel corpus is very much useful for language learning process, Cross-language information retrieval, Electronic dictionary, and Machine translation systems; especially for SMT system [5].

1.4 Statistical Machine Translation

Statistical machine translation is a popular and one of the widely used approaches of MT. It can be used for translating immense texts from one natural language to another language. It comes under Empirical (or Corpus-Based) MT system. In 1949, Warren Weaver introduced the first idea about the SMT [16]. Today, it has gained tremendous potential in the research community as well as in the commercial sector. The SMT approach uses an enormous amount of bilingual aligned parallel text corpora in both the source and target languages to achieve high quality translation result [15]. 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 [1]. The SMT approach offers the best solution to ambiguity problem. The main advantages of SMT approach are: it is easy to build and maintain, less linguistic knowledge required, and reduces human efforts. The SMT approach can be classified into the following three categories: Word-Based SMT, Phrase-Based SMT, and Hierarchal Phrase-Based SMT [14]. The SMT approach contains the following three important modules: Language model, Translation model, and Decoder.

2. REVIEW OF RELATED CORPUS CONSTRUCTION

A large amount of monolingual corpus for the English language has been built by many developers all over the world. The Brown corpus was the first machine readable general corpus of the English language which was developed by W. N. Francis and H. Kucera at Brown University [16]. Some well known English corpora available all over the world, such as BNC (British National Corpus), COCA (Corpus of Contemporary American English), ANC (American National Corpus), COHA (Corpus of Historical American English), ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts), ICLE (International Corpus of Learner English), LLC (London-Lund Corpus), LOB (Lancaster-Oslo/Bergen Corpus), WWC (Wellington Corpus of Written New Zealand English), ACE (Australian Corpus of English), ICE (International Corpus of English)-New Zealand, ICE-Singapore, ICE-Philippines, ICE-Canada, ICE-Ireland, ICE-Hong Kong, and ICE-India [6]. Apart from the monolingual corpus, lots of parallel corpora have been developed for popular natural languages all over the world, such as Arabic⇔English, English-Bulgarian, English⇔Chinese, English-German, English-Italian, English-Russian, English-Swedish, English-Turkish, French⇔English, Greek⇔English, and Spanish⇔English [10]. Some of the parallel text corpora which are available on the web and can be downloaded freely, such as Bulgarian-English, French-English, German-English, Spanish-English, and Hong Kong (English-Chinese) parallel corpus [5].

A large number of monolingual corpus and parallel corpora have been also constructed in India for English and Indian natural languages. The Kolhapur corpus is the first Indian English corpus which was developed at Shivaji University, Kolhapur, India in 1988 by Prof. S.V. Shastri and his colleagues [8]. The EMILLE (Enabling Minority Language Engineering)/CIIL (Central Institute of Indian Languages) corpus was constructed at Mysore, India which consists of Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu, and Urdu monolingual corpus [9]. The EMILLE/CIIL corpus also consists of some parallel corpus like English-Hindi, English-Bengali, and English-Urdu. The TDIL (Technology Development for Indian Languages), CDAC (Centre for Development of Advanced Computing), MCIT (Ministry of Communications and Information Technology), and CIIL are

playing a major role in developing the corpora for Indian languages. The TDIL has been constructed some multi-domain English to Indian natural languages parallel text corpora, such as English-Assamese, English-Bodo, English-Hindi, English-Manipuri, English-Nepali, and English-Urdu (http://tdil-dc.in/index.php). A list of some Indian educational institutions which have been constructed monolingual text corpora for Indian languages is shown in Table 1 [8, 14].

Table 1. Existing text corpora for Indian languages

Name of Educational Institutions	Name of Corpus
Indian Institute of Technology Guwahati, Assam	Assamese and Manipuri
Indian Institute of Technology Delhi, New Delhi	Hindi and Punjabi
Indian Institute of Technology Kanpur, Uttar Pradesh	Hindi and Nepali
Indian Institute of Technology Mumbai, Maharashtra	Marathi and Konkani
Indian Institute of Science, Bangalore, Karnataka	Kannada and Sanskrit
Gauhati University, Guwahati, Assam	Assamese
Indian Statistical Institute, Kolkata, West Bengal	Bengali
Maharaja Sayajirao University of Baroda, Gujarat	Gujarati
Utkal University, Bhubaneswar, Odisha	Oriya
Jawaharlal Nehru University, New Delhi	Sanskrit
Anna University, Chennai, Tamil Nadu	Tamil
University of Hyderabad, Hyderabad, Telangana	Telugu
Aligarh Muslim University, Aligarh, Uttar Pradesh	Urdu, Kashmiri, and Sindhi

3. CONSTRUCTION OF E-BPTC

Corpus construction is a very difficult and laborious task. A corpus is constructed according to the corpus constructor's objectives for a specific purpose using one or more natural language(s). The planning stages of a corpus construction are: type of corpus, type of domain, size of the corpus, and data collection. The process of parallel text corpus construction can be divided into three phases, namely translation, validation and sentence alignment [1].

A small parallel text corpus construction is relatively easier compared to large size specific domain parallel text corpus construction, which is difficult, time consuming and more expensive. In this section, General and Newspaper domains E-BPTC construction techniques are discussed. The main purpose of the two domains parallel corpora construction is to implement the English to Bodo SMT system. The architecture of the E-BPTC construction is shown in Figure 1.

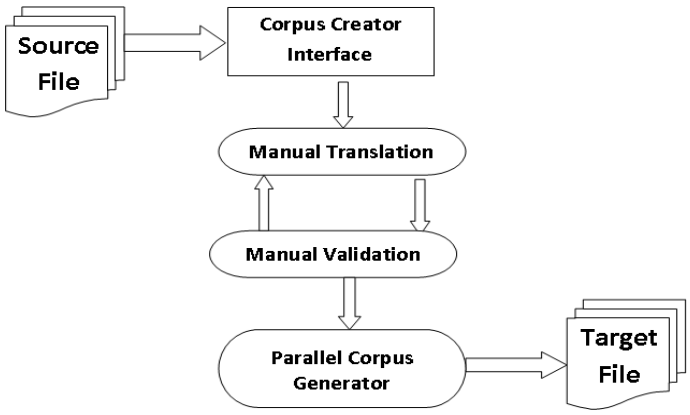

Figure1. Architecture of the E-BPTC construction

To construct the two domains E-BPTC, handwritten translated Bodo (target) sentences of the corresponding English (source) sentences have been collected from different sources. An E-BPTC creator tool has been designed for typing the texts of both the English and Bodo languages. The creator tool consists of hard and virtual keyboards for both the languages. The snapshot of the E-BPTC creator tool and the virtual keyboards for both the languages are shown in Figure 2 and 3 respectively.

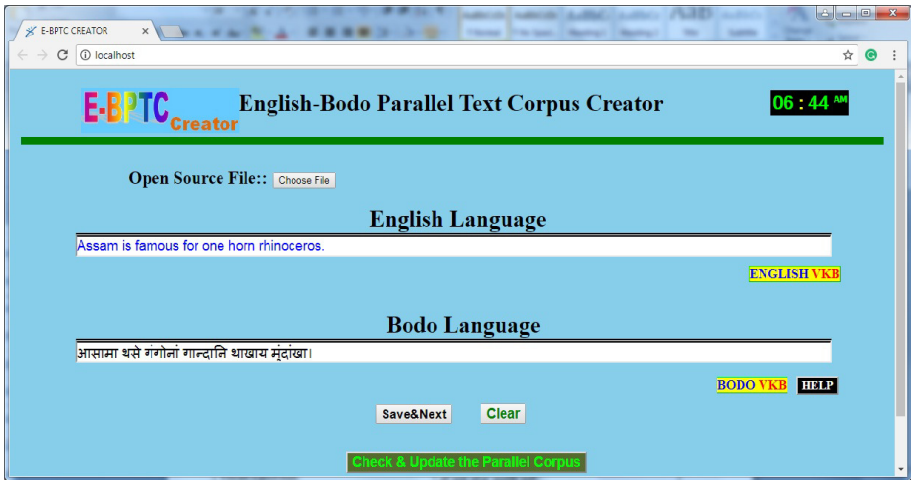

Figure 2. Snapshot of the E-BPTC creator tool

Figure 3. Snapshot of the virtual keyboards for English and Bodo languages

The E-BPTC creator tool has been designed primarily as an interface for writing the translated Bodo sentences of the corresponding English sentences and generating English-Bodo parallel text corpora. The process involved in the E-BPTC construction can be classified as follows:

- **Manual Translation:** The manual translation is the first step towards generating the English-Bodo parallel text corpus. To use this tool one need an experienced translator who depends on his/her own knowledge of the source and target languages to perform the translation. Though

the manual translation is a time consuming task, still we rely on manual translation due to the perfection of rare word/sentence translation. The chance of meaning transfer and fluency of target sentence using manual translation is higher compared to automated translation.

- **Manual Validation:** After manual translation, the translator validates the sentences with the help of linguistic persons, paper dictionary, electronic dictionary, etc. In this phase, the translator can do any kind of modification in the translated Bodo sentences. The translator also checks the correct spelling, fluency, and adequacy of the target sentences as well as the source sentences.

- **Parallel Corpus Generator:** Finally, the translator can generate the English-Bodo parallel text corpora which contain bilingual parallel sentences (In this case, English sentence and its translated Bodo sentences).

In this way, the General and Newspaper domains E-BPTC have been constructed to develop the English to Bodo SMT system. The constructed parallel corpora are briefly discussed below:

3.1 General Domain E-BPTC

The General domain E-BPTC has been constructed using English-Bodo parallel sentences which are commonly used in our daily life. The parallel sentences of the source and target languages have been collected from different sources, such as monolingual corpus, dictionaries, books, and the web. The General domain E-BPTC contains 6500 (six thousand five hundred) parallel sentences of each English and Bodo language which have been constructed using the E-BPTC creator tool. Some of the General domain English-Bodo parallel sentences are shown in Table 2.

Table 2. Sample of English-Bodo parallel sentences in the General domain E-BPTC

English Sentences	Bodo Sentences
Abdul Kalam was the best president of India.	आब्दुल कालामा भारतनि साबसिन हादोरगिरिमोन।
Assam is famous for one horn rhinoceros.	आसामा थसे गंगोनां गान्दानि थाखाय मुंदांखा।
I am feeling pleasure today.	आं दिनै गोजोनाय मोनदों।
Mahatma Gandhi was an honest person.	महत्मा गान्धीया सासे गियानि सुबुंमोन।
Oxygen gas is essential for our life.	अक्सिजेन गेसा जोंनि जिउनि थाखाय जोबोद गोनांथार।
Peacock is a very beautiful bird.	दाउराइया मोनसे जोबोद समायना दाउ।
Television can be used for entertainment.	गोजोनग्लायनो थाखाय टेलिभिसनखौ बाहायनो हागौ।
The kite is flying in the sky.	सिलाया अख्राडाव बिरगासिनो दङ।
You are a good man.	नों सासे मोजां मानसि नंगौ।
What is your mother tongue?	नोंनि बिमा रावा मा?

3.2 Newspaper Domain E-BPTC

The Newspaper domain E-BPTC has been constructed using English-Bodo parallel sentences which are completely related to news and the sentences are simple, important and generally happened news. The parallel sentences of the source and target languages have been collected from different sources, such as monolingual corpus, dictionaries, newspapers, and the web. The newspapers that have been generally considered for data collection purpose are: English newspapers (The Assam Tribune and The Times of India) and Bodo newspaper (Bodoland Sansri). The Newspaper domain E-BPTC contains 4000 (four thousand) parallel sentences of

each English and Bodo language which have been constructed using the E-BPTC creator tool. Some of the Newspaper domain English-Bodo parallel sentences are shown in Table 3.

Table 3. Sample of English-Bodo parallel sentences in the Newspaper domain E-BPTC

English Sentences	Bodo Sentences
Flood situation in Assam still grim.	आसामाव दैबानाया थासारिया दासिमबो गिथाव बाथाव।
Guwahati is served by regular Indian Airlines and Jet Airways flights from Kolkata and Delhi.	गुवाहाटीयाव जेब्लाबो इण्डियान एयारलायइन्स आरो जेट एयारवेजनि बिरखंजों दिल्ली आरो कलकातानिफ्राय हान्थामेला जायो।
One year old Sonowal government asked to take concrete steps for women security.	से-बोसोरारि बैसोनि सन'वाल सरकारखो हिनजाव फोरनि रैखाथिनि थाखाय गोख्रों थांखि लानो खावलायनाय जादों।
Police have apprehended three drugs peddlers from the Guwahati railway station today.	साथाम फेग्रा बेसाद फानग्रा फोरखो दिनै गुवाहाटी रेल स्टेसननिफ्राय पुलिसा हमो।
The Assam government has failed to fulfill the agreement.	आसाम सोरखारा गोरोबथा मावफुंनायाव फेलें जादोंमोन।
The prime minister Narendra Modi will come in Guwahati on Friday.	गाहाय मन्थ्रि नरेन्द्र म' दिया शुक्रबाराव गुवाहाटीयाव फैगोन।
The results of HSLC and Assam High Madrasa Examination 2017 will be declared on 30 May.	2017 नि HSLC आरो आसाम हाइ माद्रासा आनजादनि रिजाल्टखो मे दामनि 30 अक्ट' सानखालि फोसावनाय जागोन।
Today, Ajmer is a popular pilgrimage center for the Hindus as well as Muslims.	दिनै आजमेरा हिन्दु आरो मुसलमानफोरनि थाखाय मोनसे मुंदांखा गोथार दावबायग्रा थावनि।

4. IMPLEMENTATION OF ENGLISH TO BODO SMT

The English to Bodo SMT system has been developed using Phrase-Based SMT (PBSMT) approach with the help of General and Newspaper domains E-BPTC. The PBSMT approach is a more accurate and deeply used in SMT system nowadays. It has several advantages as compared to Word-Based SMT (WBSMT). The aim of it is to reduce the limitations of the WBSMT. In the PBSMT, each source and target sentences are divided into separate phrases before translation [4]. The word or phrase alignment between the sentences of the source and target languages normally follows certain patterns, which is very similar to WBSMT.

The following operations have been performed to develop the English-Bodo SMT system using PBSMT approach with the help of General and Newspaper domains E-BPTC.

- **Corpus Preparation:** The corpus preparation is a very important task to train the SMT system. The constructed General and Newspaper domains English-Bodo parallel text corpora have been first converted into UTF-8 text file format in Linux. The English and Bodo sentences have been separated from the two domains parallel text corpora and created two separate files for English and Bodo languages. After that, the following steps have been performed on the English and Bodo 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).

- **Language Model:** The Language Model (LM) is built to compute the probability of Bodo sentences (B), *i.e.* P(B). In this system, the toolkit KenLM is used to build the LM with the help of 3-gram technique. The LM is used to ensure the fluency of the translated Bodo sentences.

- **Translation Model:** The Translation Model (TM) is used to compute the probability of the English sentence E for a given Bodo sentence B, *i.e.* P(E|B). The toolkit Giza++ is used for

word or phrase alignment to develop the TM. The TM is used to ensure the adequacy of the translation result.

- **Decoder:** The decoder is used to find the maximum translation probability from the English language to the corresponding Bodo language. The decoder uses A* search method to find the best possible translation result [11]. The decoder determines the maximum 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

5. RESULT, EVALUATION, AND COMPARISON

The phrase-based English to Bodo SMT system has been trained with the General and Newspaper domains E-BPTC separately. The SMT system has been examined several times with various numbers of General and Newspaper domains English-Bodo parallel sentences separately and achieved various translation results. One thing has been noticed from the translation results that the quality of the translation results can be increased by increasing the sizes (number of sentences) and quality of the parallel corpora to train the SMT system. Lastly, the number of words and sentences which have been used in the two domains E-BPTC for training the SMT system are shown in Table 4.

Table 4. Number of words and sentences exist in the two domains E-BPTC

E-BPTC	**Languages**	**Words**	**Sentences**
General domain	English	52,778	6,500
	Bodo	41,920	
Newspaper domain	English	45,914	4,000
	Bodo	38,205	
Total	English	98,692	10,500
	Bodo	80,125	

Finally, the accuracy of the translation results has been evaluated using two evaluation techniques viz. manually and automatically. In the manually evaluation technique, a linguistic person Mr. Dwipen Baro (Assistant Professor, Department of Bodo, Dhamdhama Anchalik College, Nalbari, Assam) has been evaluated the SMT system manually in terms of level of Translation Accuracy (TA) i.e. adequacy and fluency. The levels of TA (in terms of percentage) of the General and Newspaper domains E-BPTC in the SMT system are shown in Table 5 and compared in Figure 4.

Table 5. Levels of TA of the two domains E-BPTC

Levels	Definitions	TA(%) of the E-BPTC	
		General domain	Newspaper domain
Perfect	The translated sentence is very good to understand.	45	43
Fair	The translated sentence is easy to understand, but need a minor correction.	35	36
Acceptable	The translated sentence is broken, but is understandable.	16	17
Nonsense	The translated sentence is not understandable.	4	4

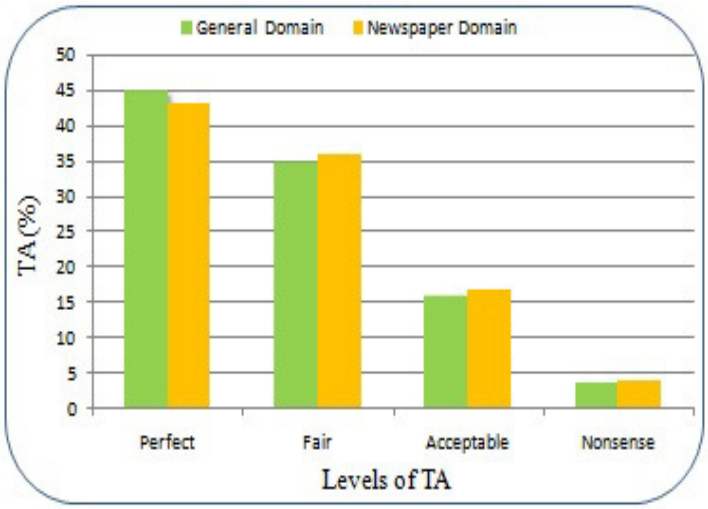

Figure 4. Comparison between the levels of TA of the two domains E-BPTC

In the automatic evaluation technique, BLEU (Bilingual Evaluation Understudy) technique has been used to evaluate the quality of the translation results. BLEU is the best and useful technique for automatic evaluation of any SMT system. The BLEU technique was developed by Kishore Papineni [2]. The BLEU scores of the General and Newspaper domains E-BPTC in the English to Bodo SMT system are shown in Table 6 and compared in Figure 5.

Table 6. BLEU scores of the two domains E-BPTC

Name of the Parallel Corpus	**BLEU Score**
General domain E-BPTC	38.25
Newspaper domain E-BPTC	33.08

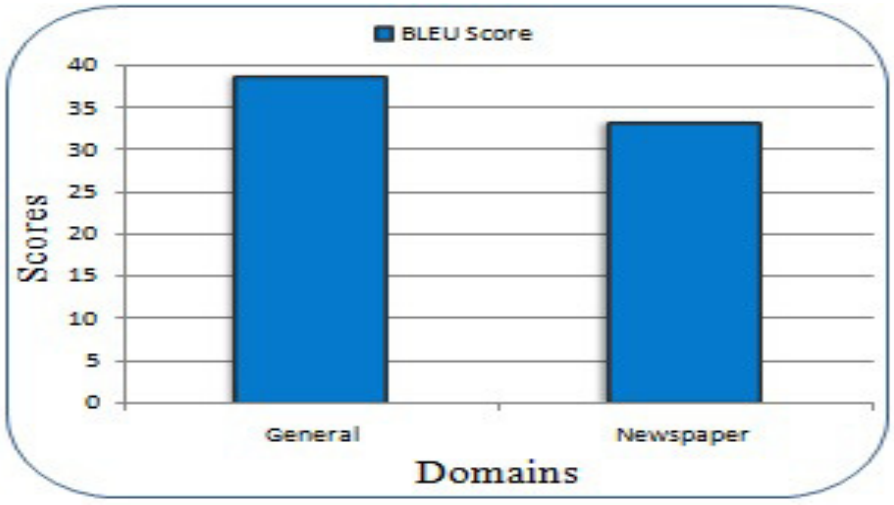

Figure 5. Comparison between the BLEU scores of the two domains E-BPTC

BLEU score can be increased by increasing the sizes of the parallel corpora. Good quality parallel corpora may also enhance the BLEU score.

6. CONCLUSION AND FUTURE RESEARCH WORK

Though corpus construction is a very difficult and laborious task, but it could be enhanced in language education, language technology, linguistic research, and NLP tasks. In this paper, the General and Newspaper domains E-BPTC have been constructed using E-BPTC creator tool and the English-Bodo SMT system has been developed using PBSMT approach with the help of the General and Newspaper domains parallel corpora separately. A total number of 10,500 (ten thousand five hundred) English-Bodo parallel sentences in the two domains parallel text corpora have been constructed and prepared to train the English to Bodo SMT system. Finally, the SMT system has been tested with various numbers of parallel sentences of the two domains parallel corpora separately and achieved different translation results. Although the sizes of the two domains parallel corpora are not large, still relatively good translation results have been achieved in the system. The General and Newspaper domains E-BPTC may be useful as language resources for research scholars and other people who are interested in language learning, develop language technology, and linguistic research. Since North-East is a multilingual region and not much work has been done on corpus construction as well as machine translation for Bodo language. Hence, it can be expected that the English to Bodo SMT system would be immensely helpful for students, tourists, and other people of India, especially of North-East region of India.

The research work can be extended by adding a greater number of parallel sentences in both the General and Newspaper domains E-BPTC for better translation results. The transliteration module can be added in the SMT system to improve the quality of the translation results. The research work can also be extended by developing English⇔Bodo MT systems using Neural Machine Translation approach with the help of multi-domain English-Bodo parallel text corpora.

REFERENCES

[1] Diptesh Kanojia, Manish Shrivastava, Raj Dabre, and Pushpak Bhattacharyya (2014). PaCMan: Parallel Corpus Management Workbench. *In Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, 162–166.

[2] Hans Uszkoreit (2007). Survey of Machine Translation Evaluation. *EuroMatrix Project*, Saarland University, Germany, 1-80.

[3] Jyotismita Talukdar, Chandan Sarma, and Prof. P.H. Talukdar (2012). Automatic Syllabification Rules for Bodo Language. *International Journal of Computational Engineering Research*, 2(6):110-114.

[4] Kathiravan, P., Makila, S., Prasanna, H., and Vimala, P. (2016). Over View- The Machine Translation in NLP. *International Journal for Science and Advance Research in Technology*, 2(7):19-25.

[5] Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 1837-1842

[6] Nadja Nesselhauf (2005). Corpus Linguistics: A Practical Introduction.

[7] Niladri Sekhar Dash (2010). Corpus Linguistics: A General Introduction. *Central Institute of Indian Languages*, Mysore.

[8] Niladri Sekhar Dash (2004). Language Corpora: Present Indian Need. *In the Proceedings of the SCALLA 2004 Working Conference*.

[9] Paul Baker, Andrew Hardie, Tony McEnery, and B.D. Jayaram (2003). Constructing Corpora of South Asian Languages. *UK EPSRC-Project*, Department of Linguistics, Lancaster University and Central Institute of Indian Languages, Mysore, India,71-80.

[10] Philipp Koehn (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit X*.

[11] Philipp Koehn (2016). MOSES (User Manual and Code Guide). *Statistical machine translation system*, University of Edinburgh, UK.

[12] Saiful Islam (2016). An English to Assamese, Bengali and Hindi Multilingual E-Dictionary. *International Journal of Current Engineering and Scientific Research,* 3(9):74-80.

[13] Saiful Islam, Maibam Indika Devi, and Bipul Syam Purkayastha (2017). A Study on Various Applications of NLP Developed for North-East Languages. *International Journal on Computer Science and Engineering*, 9(6):368-378.

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

[15] Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Ch. Deka, and Anup Kr. Barman (2012). A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges. *In the proceedings of the 10th Workshop on Asian Language Resources,* COLING, Mumbai, 21–28.

[16] Thoudam Doren Singh (2012). Building Parallel Corpora for SMT System: A Case Study of English-Manipuri. *International Journal of Computer Applications*, 52(14): 47-51.

AUTHORS

Saiful Islam is working as PhD research scholar under Prof. B. S. Purkayastha in the Department of Computer Science, Assam University, Silchar, India. He has completed MCA from Tezpur University, Tezpur and MPhil in Computer Science from Assam University, Silchar, Assam, India.

Abhijit Paul is working as PhD research scholar under Prof. B. S. Purkayastha in the Department of Computer Science, Assam University, Silchar, India. He has completed MCA from Assam Engineering College, Guwahati and MPhil in Computer Science from Assam University, Silchar, Assam, India.

Bipul Syam Purkayastha is currently working as Professor in the Department of Computer Science, Assam University, Silchar, India. He has completed PhD in Computer Science from North-Eastern Hill University, Shillong, Meghalaya, India. He has published many papers in the International Journals and Conferences. His major research area is Natural Language Processing (NLP).

Ismail Hussain is currently working as Assistant Professor in the Department of Bodo, Gauhati University, Guwahati, India. He was also working as Assistant Professor in the Department of Bodo, Bodoland University, Kokrajhar, India. He has completed PhD in Bodo language from Gauhati University, Guwahati, Assam, India.