# English to Bodo Statistical Machine Translation System Using Multi-domain Parallel Corpora

**Saiful Islam**
Dept. of Computer Science
Assam University
Silchar, India

**Ismail Hussain**
Dept. of Bodo
Gauhati University
Guwahati, India

**Bipul Syam Purkayastha**
Dept. of Computer Science
Assam University
Silchar, India

## Abstract

Parallel corpus is a primary resource for most of the applications of Natural Language Processing (NLP) like Machine Translation (MT) and CLIR. Statistical Machine Translation (SMT) is a highly successful and popular approach of MT that can produce high-quality translation results using a huge amount of bilingual parallel corpus. This paper primarily focuses on the development of English to Bodo SMT system using multi-domain English-Bodo Parallel Text Corpus (E-BPTC). The SMT system has been developed using the Phrase-based SMT approach for the different domains of E-BPTC, namely Tourism, News, Health, General and Agriculture. The translation accuracy of the different domains of E-BPTC in the SMT has been evaluated using the Manual and Automatic evaluation techniques.

## 1 Introduction

Machine translation is the most important application of NLP that translates texts from one natural language to another automatically and quickly. Nowadays, MT is a very challenging research task globally in the field of NLP. It is a very difficult task due to some challenges in natural languages like word order and word ambiguity. The approaches of MT can be classified into different categories (Antony, 2013; Islam & Purkayastha, 2018) as shown in Figure 1.

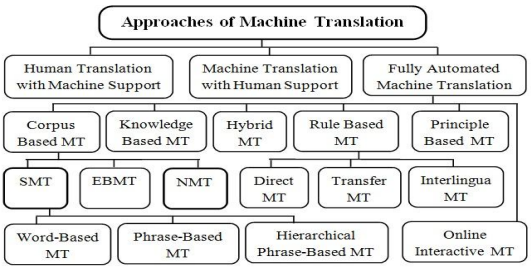

Figure 1: Approaches of MT

At present, the most popular and state-of-the-art approaches of MT are SMT and NMT (Neural Machine Translation). The SMT has gained tremendous potential globally in the research community as well as in the commercial sectors. In 1949, Warren Weaver introduced the first concepts of SMT approach (Kathiravan et al., 2016). The SMT is the best technique of MT for reducing word ambiguity problems in the natural languages. It requires less linguistic knowledge and can reduce human efforts (Koehn, 2009). It is classified as Word-based SMT, Phrase-based SMT and Hierarchical Phrase-based SMT.

Bodo is one of the major spoken languages of North-East India (Islam et al., 2017). It is a recognized language of India and is an official language of Bodoland Territorial Council (Assam). The Bodo language is written using Devanagari script. It is a low resource language and the word order of it is Subject+ Object + Verb.

English is an International human language and is primarily spoken by the people of many countries, such as Australia, United Kingdom and the United States. English is an associate official language of India (Islam and Purkayastha, 2018). The English language is written using Latin script. It is a high resource language and the word order of it is Subject + Verb + Object.

## 2 Related Work

A large number of SMT systems and parallel corpora have been developed and constructed for popular natural languages as well as for Indian languages. Some of them are discussed below.

### 2.1 Parallel Corpus

Lots of parallel corpus has been built globally for popular natural languages. Some of the parallel corpora are briefly discussed as follows:

**Bible corpus:** The corpus was constructed from the translation of the Bible. It is a multilingual



*Proc. of ICON-2018, Patiala, India. December 2018* ©*2018 NLPAI*, pages 75–81

parallel text corpus and contains 100 natural languages. The corpus is freely available online[1].

**English-Kazakh parallel corpus:** The corpus was constructed at Al-Farabi Kazakh National University, Kazakhstan by (Kuandykova et al., 2014) for the English⇔Kazakh SMT system.

**Europarl corpus:** The corpus was constructed at the University of Edinburgh, Scotland, UK by (Koehn, 2005) for the SMT system. The corpus is freely available online[2].

**OPUS Corpus:** The OPUS is a multilingual parallel corpus that contains 60 different languages. The corpus is freely available online[3].

**UM corpus:** The corpus was constructed at the University of Macau, China by (Tian et at., 2014) for the SMT system. It is a multi-domain English-Chinese parallel text corpus.

Some of the parallel corpora which have been constructed for English and Indian natural languages are discussed as follows:

**TDIL corpus:** The TDIL[4] (Technology Development for Indian Languages) programme has constructed different domains of parallel corpora. Some of the parallel corpora exist in this corpus are English-Assamese, English-Bodo, English-Hindi, Hindi-Punjabi and Hindi-Urdu.

**EMILLE/CIIL Corpus:** The EMILLE (Enabling Minority Language Engineering)/CILL (Central Institute of Indian Languages) corpus was constructed jointly at Lancaster University and CIIL, Mysore, India. Some of the parallel corpora exist in this corpus are English-Hindi, English-Bengali and English-Urdu (Baker et al., 2003)**.**

**English-Punjabi parallel corpus:** The corpus was constructed at Punjabi University, Patiala, India by (Jindal et al., 2017) for SMT system.

**English-Manipuri parallel corpus:** The corpus was constructed at CDAC Mumbai, India by (Singh, 2012) for SMT system.

### 2.2 SMT System

A large number of SMT systems have been developed globally for popular languages. Some of the SMT systems are discussed as follows:

**English to Arabic SMT system:** The system was developed at MIT, USA by (Badr et al., 2008) using the Phrase-based SMT (PBSMT) technique and Moses. The BLEU score was 28.9.

**English to Spanish SMT system:** The system was developed at the University of California, Berkeley by (Nakov, 2008) using the PBSMT technique. The BLEU score was 21.92.

**English to Vietnamese SMT system:** The system was developed at the University of Ulsan, Ulsan, Korea by (Phuoc et al., 2016) using the PBSMT technique and Moses. The BLEU score was 32.30**.**

**English⇔Welsh SMT system:** The system was developed using the Phrase-based SMT approach by (Jones & Eisele, 2006) at Saarland University, Germany. The BLEU scores of the English to Welsh and Welsh to English SMT systems were 36.16 and 42.22 respectively.

**French to English SMT system:** The system was developed at Carnegie Mellon University, Pittsburgh, USA by (Hanneman et al., 2009) using the PBSMT approach and Moses.

Lots of SMT system has been developed for English and Indian natural languages. Some of the SMT systems are discussed as follows:

**Assamese to English SMT system:** The system was developed at Gauhati University, Guwahati, India by (Baruah et al., 2014) using the PBSMT approach and Moses. The BLEU score was 9.72.

**English to Punjabi SMT system:** The system was developed at the I. K. Gujral Punjab Technical University, Punjab, India by (Jindal et al., 2018) using the PBSMT technique.

**Hindi⇔English SMT system:** The system was developed at IIT Bombay, India by (Dungarwal et al., 2014) using the PBSMT technique.

**Manipuri⇔English SMT system:** The system was developed at Jadavpur University, Kolkata, India by (Singh & Bandyopadhyay, 2010) using the Phrase-based SMT technique and Moses.

## 3 Corpus Construction and Collection

The different domains of English-Bodo parallel corpus which have been used to develop the English to Bodo SMT system are discussed below.

### 3.1 Construction of Parallel Corpus

[1]http://christos-c.com/bible
[2]https://en.wikipedia.org/wiki/Europarl_Corpus
[3]http://opus.nlpl.eu
[4]http://tdil-dc.in/index.php?lang=en

The construction of parallel corpus is a very laborious and difficult task. A GUI based E-BPTC (English-Bodo parallel Text Corpus) creator tool has been designed for constructing English-Bodo parallel text corpus. The tool has been designed primarily for typing the handwritten translated sentences of Bodo language of the corresponding given sentences of English language. In this tool, Unicode-based a Bodo hard keyboard and a Bodo soft keyboard have been designed for typing the texts of the Bodo language. The Bodo hard keyboard is used through the English hard keyboard. The General and News domains of English-Bodo parallel text corpus have been built using the E-BPTC creator tool. The constructed parallel corpora are discussed as follows:

**General domain E-BPTC:** The corpus has been constructed using the E-BPTC creator tool. The screenshot of the creator tool for building the General domain E-BPTC is shown in Figure 2.

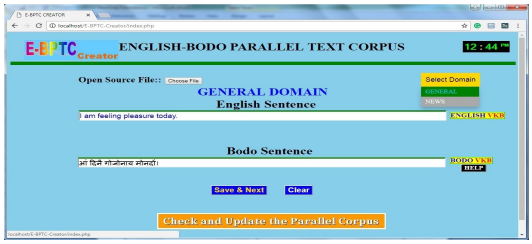


Figure 2: Screenshot of the E-BPTC creator tool for constructing the General domain E-BPTC

The corpus contains English-Bodo parallel sentences which are generally used in our daily life like communication, meeting, teaching and interview purposes. The English sentences and their corresponding translated handwritten Bodo sentences have been collected from the various sources, such as dictionaries, books, corpora and the web. Total 6,500 English-Bodo parallel sentences have been constructed in this corpus.

**News domain E-BPTC:** The corpus has been constructed using the E-BPTC creator tool. The screenshot of the creator tool for building the News domain E-BPTC is shown in Figure 3.

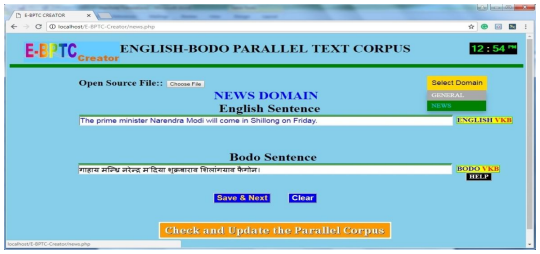


Figure 3: Screenshot of the E-BPTC creator tool for building the News domain E-BPTC

The corpus contains English-Bodo parallel sentences and the sentences have been collected mainly from the Educational news, General news, Political news and Sports news. The English sentences and their corresponding translated handwritten Bodo sentences have been collected from the different sources, such as English and Bodo Newspapers, corpora and the web. Total 4000 English-Bodo parallel sentences have been constructed in this corpus.

### 3.2 Collection of Parallel Corpus

The different domains of English-Bodo parallel corpus which have been collected from the TDIL programme are described as follows:

**Agriculture domain E-BPTC:** The corpus contains English-Bodo parallel sentences. Total 4000 English-Bodo parallel sentences have been prepared in this corpus.

**Health Domain E-BPTC:** The corpus contains English-Bodo parallel sentences. Total 12,300 English-Bodo parallel sentences have been prepared in this corpus.

**Tourism Domain E-BPTC:** The corpus contains English-Bodo parallel sentences. Total 9,200 English-Bodo parallel sentences have been prepared in this corpus.

## 4. English to Bodo SMT System

The English to Bodo SMT system has been developed using the Phrase-based SMT technique for the different domains of English-Bodo parallel text corpus, namely Tourism, News, Health, General and Agriculture. The PBSMT is a perfect and widely used approach of MT nowadays. It needs a huge amount of bilingual parallel aligned corpus for the best translation results

The overall architecture of the English-Bodo Phrase-based SMT system is shown in Figure 4.

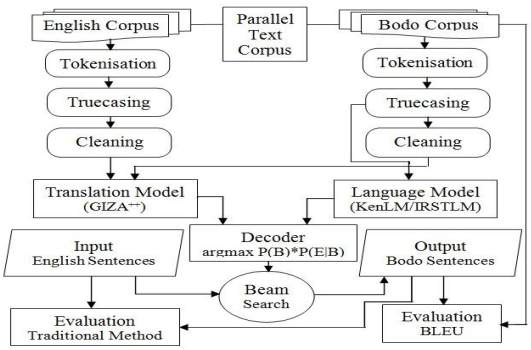


Figure 4: Overall architecture of the SMT system

The following operations have been performed for the different domains of E-BPTC to train the English to Bodo SMT system.

### 4.1 Corpus Preparation

Corpus pre-processing is the most essential task to prepare a bilingual parallel corpus for training the SMT system using Moses (Koehn, 2016). The following steps have been performed for the different domains of E-BPTC to build the different statistical models.

Step 1: Tokenization has been performed for the parallel corpus to insert space between the words and punctuation.

Step 2: True casing has been performed for the corpus to convert the first word of each sentence to their most probable casing.

Step 3: Cleaning has been performed to remove the long sentences, empty sentences, extra spaces and misaligned sentences from both the English and Bodo corpora.

### 4.2 Language Model

The Language Model (LM) is an important component of the SMT. It is built to measure the fluency of the sentences of the target language. In this system, LM has been built for the different domains of Bodo corpus using the toolkit KenLM. The LM has computed the probability of the sentences of Bodo language P(B) using the 3-gram modelling technique. It has computed the probability of a Bodo sentence as the probability of particular words P(w) using the Markov Chain Rule (Koehn, 2009) as shown in Eq. 1.

$$P(B)=P(w_1,w_2,w_3,.....,w_n)$$
$$=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)P(w_4|w_1w_2w_3).....$$
$$.....P(w_n|w_1w_2...w_{n-1}) \quad (1)$$

Where $w_1$, $w_2$, $w_3$,……., $w_n$ are the words of the Bodo language.

The formula for calculating tri-gram probabilities $P(w_n|w_{n-2}w_{n-1})$ of the sentences of target language is shown in Eq. 2.

$$P(w_n|w_{n-2}w_{n-1})= \frac{\text{Count}(w_{n-2}w_{n-1}w_n)}{\text{Count}(w_{n-2}w_{n-1})} \quad (2)$$

Where Count $(w_{n-2}w_{n-1}w_n)$ indicates the number of occurrences of the sequence $w_{n-2}w_{n-1}w_n$ in the corpus.

### 4.3 Translation Model

Translation Model (TM) is the most important component of the SMT. It is used to confirm the adequacy of the translation results in the SMT system. The TM has calculated the probabilities of the English sentence (E) and the Bodo sentence (B) based on the behaviour of the sentences, i.e. P(E|B). It can be calculated as the sum over all probabilities of all possible alignments (A) between the words or phrases in two sentences of the English and Bodo languages (Brunning, 2010) as shown in Eq. 3.

$$P(E|B) = \sum_{A} P(E, A|B) \quad (3)$$

The GIZA++ toolkit has been used in the SMT system for word or phrase alignment and to build the translation model for the different domains of E-BPTC. In the TM, a phrase translation table is created and the table ensures that the English phrases and the Bodo phrases are good translations of each other. An example of the word (or phrase) alignment in the English to Bodo Phrase-based translation model is shown in Figure 5.

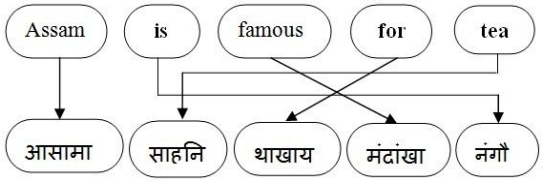


Figure 5: Word alignment between English and Bodo sentences

### 4.4 Decoder

The decoder is an essential component of SMT. It can find out the maximum translation probability using the output results of the LM and TM as shown in Eq. 4. The decoder uses a Beam search technique to find the best possible translation results (Koehn, 2004).

$$P(E, B) = \text{argmax } P(B) * P(E|B) \quad (4)$$

Where P(B) is the output result of the LM and P (E|B) is the output result of the TM.

## 5 Experimental Results

The English to Bodo SMT system has been tested several times using the different numbers of English-Bodo parallel sentences for the different domains of E-BPTC, namely Tourism, News, Health, General and Agriculture. It has achieved good translation results in the SMT system using the General domain E-BPTC. The English-Bodo parallel sentences which have been used to train, tune and test the SMT system for the different domains of E-BPTC are shown in Table 1.

| Multi-domain E-BPTC | Languages | Training | | Tuning | Testing |
|---|---|---|---|---|---|
| | | Words | English-Bodo parallel sentences | | |
| Agriculture | English | 85,901 | 4,000 | 500 | 3,000 |
| | Bodo | 64,935 | | | |
| General | English | 52,778 | 6,500 | 600 | 5,000 |
| | Bodo | 41,920 | | | |
| Health | English | 224,077 | 12,300 | 1000 | 10,000 |
| | Bodo | 172,723 | | | |
| News | English | 45,914 | 4,000 | 500 | 3,000 |
| | Bodo | 38,205 | | | |
| Tourism | English | 199,570 | 9,200 | 1000 | 8,000 |
| | Bodo | 165, 214 | | | |

Table 1: No. of sentences used for training, tuning and testing the SMT system

## 6 Evaluation

The Translation Accuracy (TA) of the different domains of English-Bodo parallel corpus in the English to Bodo SMT system has been evaluated using the Manual or Human and Automatic or Machine evaluation techniques.

### 6.1 Manual Evaluation Technique

The translation accuracy of the different domains of E-BPTC in the SMT system has been evaluated by a linguistic person Dr. Ismail Hussain, Assistant Professor, Dept. of Bodo, Gauhati University, Guwahati, India. He has evaluated the TA (in terms of percentage) based on the adequacy and fluency of the input English sentences and their corresponding translated or output Bodo sentences. The levels of TA of the different domains of parallel corpus in the SMT system are shown in Table 2.

| Levels | Definitions | TA (%) of the multi-domain parallel corpora | | | | |
|---|---|---|---|---|---|---|
| | | Agriculture | General | Health | News | Tourism |
| Perfect | A | 35 | 45 | 40 | 40 | 38 |
| Fair | B | 32 | 36 | 34 | 37 | 35 |
| Acceptable | C | 25 | 14 | 20 | 18 | 20 |
| Nonsense | D | 8 | 5 | 6 | 5 | 7 |

Table 2: Levels of TA of the multi-domain E-BPTC (Approx)

In the above table, the definition A means Translated Bodo sentences are very good to understand, B means Translated Bodo sentences are easy to understand but need a minor correction, C means Translated Bodo sentences are broken but are understandable, and D means Translated sentences are not understandable.

### 6.2 Automatic Evaluation Technique

The translation accuracy of the different domains of E-BPTC in the SMT system has been evaluated using the BLEU (Bilingual Evaluation Understudy) technique. The BLEU is a popular automatic and language independent evaluation technique. It can evaluate the best translation accuracy in an SMT system. The BLEU score is computed based on the average of matching n-grams between a proposed or candidate translation (in this case, Machine translated Bodo corpus) and a reference or human translation (in this case, human translated Bodo corpus). The BLEU score seems to correspond well with the human judgment based on the fluency and accuracy (Uszkoreit, 2007). The BLEU scores of the different domains of E-BPTC are shown in Table 3.

| Multi-domain E-BPTC | Translation | Bodo Sentences | BLEU Scores |
|---|---|---|---|
| Agriculture | Reference | 3000 | 29.25 |
| | Candidate | | |
| General | Reference | 5000 | 38.12 |
| | Candidate | | |
| Health | Reference | 10000 | 36.05 |
| | Candidate | | |
| News | Reference | 3000 | 32.96 |
| | Candidate | | |
| Tourism | Reference | 8000 | 35.40 |
| | Candidate | | |

Table 3: BLEU scores of the different domains of E-BPTC

## 7 Conclusion and Future Work

In this paper, the English to Bodo SMT system has been developed using the Phrase-based SMT technique for the different domains of English-Bodo parallel corpus, namely Tourism, News, Health, General and Agriculture. A GUI based E-BPTC creator tool has been developed for building the General and News domains of English-Bodo parallel text corpus. The translation accuracy of the different domains of E-BPTC has been evaluated using the Manual evaluation and BLEU techniques. The General domain E-BPTC has produced good translation results in the English to Bodo SMT system.

The SMT system can be extended by adding more number of good quality parallel sentences in the different domains of English-Bodo parallel corpus to achieve the best translation results. The accuracy of the translation results of the different domains of E-BPTC can be enhanced using the Machine transliteration technique in the SMT system. The research work can be explored by developing bidirectional English⇔Bodo MT system using the NMT approach for the different domains of English-Bodo parallel corpus like Agriculture, General, Health, News and Tourism.

## References


Antony P. J. 2013. Machine Translation Approaches and Survey for Indian Languages. *Computational Linguistics and Chinese Language Processing (CLCLP)*, 18(1):47-78.

Ayana Kuandykova, Amandyk Kartbayev, and Tannur Kaldybekov. 2014. English-Kazakh Parallel Corpus for Statistical Machine Translation. *International Journal on Natural Language Computing*, 3(3): 65-72.

Dafydd Jones and Andreas Eisele. 2006. Phrase-based Statistical Machine Translation between English and Welsh. *In the proceedings of the 5th SALTMIL workshop on Minority Languages*, pp.75-77.

Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, and Alon Lavie. 2009. An Improved Statistical Transfer System for French-English Machine Translation. *In the Proceedings of the 4th Workshop on Statistical Machine Translation*, ACL, Athens, Greece, pp.140-144.

Hans Uszkoreit. 2007. Survey of Machine Translation Evaluation. *EuroMatrix Project*, Saarland University, Germany, pp.1-80.

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for English-to-Arabic Statistical Machine Translation. *In the Proceedings of Association for Computational Linguistics-08: HLT*, USA, pp.153-156.

James Brunning. 2010. Alignment Models Algorithms for Statistical Machine Translation (PhD Thesis). Cambridge University, UK.

Kalyanee K. Baruah, Pranjal Das, Abdul Hannan, Shikhar kr. Sarma. 2014. Assamese-English Bilingual Machine Translation. *International Journal on Natural Language Computing (IJNLC)*, 3(3): 73-82.

Kathiravan, P., Makila, S., Prasanna, H., and Vimala, P. 2016. Overview- The Machine Translation in NLP. *International Journal for Science and Advanced Research in Technology*, 2(7):19-25.

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 the proceedings of the 9th International Conference on Language Resources and Evaluation.* European Language Resources Association (ELRA), pp.1837-1842.

Nguyen Quang Phuoc, Yingxiu Quan, and Cheol-Young Ock. 2016. Building a Bidirectional English-Vietnamese Statistical Machine Translation System by using Moses. *International Journal of Computer and Electrical Engineering*, 8(2):161-168.

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, pp.71-80.

Philipp Koehn. 2004. Pharaoh: A Beam Search Decoder for Statistical Machine Translation. *In the proceedings of the 6th Conference of the Association for Machine Translation in the Americas*(AMTA), Springer.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *In the proceedings of the MT Summit X.*

Philipp Koehn. 2009. Statistical Machine Translation (e-Book). *Cambridge University Press*, New York, pp.1-447.

Philipp Koehn. 2016. MOSES (User Manual). Statistical Machine Translation (SMT) system, *University of Edinburgh*, UK.

Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2014. The IIT Bombay Hindi⇔English Translation System at WMT 2014. *In the proceedings of the Ninth Workshop on Statistical Machine Translation*, ACL, Maryland, USA, pp.90-96.

Preslav Nakov. 2008. Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing. *In the proceedings of the 3rd Workshop on Statistical Machine Translation*, Columbus, USA, pp.147-150.

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 (IJCSE)*, 9(6):368-378.

Saiful Islam and Bipul Syam Purkayastha. 2018. English to Bodo Phrase-Based Statistical Machine Translation. *In the proceedings of the 10th International Conference on Advanced Computing and Communication Technologies (ICACCT-2017). Advances in Intelligent*

*Systems and Computing*, Springer, Singapore, vol. 562: 207-217.

Shishpal Jindal, Vishal Goyal, and Jaskarn Singh Bhullar. 2017. Building English-Punjabi Parallel Corpus for Machine Translation. *International Journal of Computer Applications*, 180(8):26-29.

Shishpal Jindal, Vishal Goyal, and Jaskarn Singh Bhullar. 2018. English to Punjabi statistical machine translation using Moses (Corpus-Based). *Journal of Statistics and Management Systems*, 21(4):553-560.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations. *In the proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation*, COLING-2010, Beijing, pp.83-91.

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