International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 14, No. 2, pp. 20 – 30, 2017

IMPLEMENTATION OF ENGLISH TO BODO MACHINE TRANSLATION SYSTEM USING SMT APPROACH

SAIFUL ISLAM*

Department of Computer Science, Assam University, Silchar,
PIN-788011, Assam, India
sislam.mca@gmail.com

BIPUL SYAM PURKAYASTHA
Department of Computer Science, Assam University, Silchar,
PIN-788011, Assam, India
bipul_sh@hotmail.com

Statistical Machine Translation (SMT) is a highly successful technique in Machine Translation (MT) system and is deeply used by many commercial systems like Google translate, Bing translate, and so on. At present, the demand of machine translation has greatly increased in India as well as all over the world due to the necessity for communication amongst human. Bodo language is one of the popular natural languages of North-East India and also recognized language of India. Even then the computerized information of Bodo language is very low. Thus, we want to expand the computerized information of Bodo language. The primary objective of the proposed system is to develop English to Bodo MT system using General domain English-Bodo parallel text corpora. The proposed system is implemented using SMT approach and Moses. We have achieved relatively good translation result and the accuracy of the translation result is evaluated using two evaluation techniques in our system.

Keywords: Bodo language; English language; Machine translation; Moses; SMT.

1. Introduction

Machine translation is a process which can translate text or speech from a source natural language (SNL) to target natural language (TNL) using computers automatically. The first computer based application related to natural language was the machine translation. The first concept of machine translation was started by the philosopher René Descartes in the seventeenth century [Antony (2013)]. Generally, machine translation occurs between two particular natural languages and it may be either unidirectional or bi-directional [Uszkoreit (2007)]. Machine translation is a very difficult task due to some problems with it like word order, word sense ambiguity, idioms, and preposition or post-position. The main benefits of MT are a huge amount of text can be translated from one natural language to another language without the help of human translators, can reduce expenditure and lessen human efforts [Islam *et al.* (2017)]. Nowadays, MT is a very challenging research task in the field of Computational Linguistics and Natural Language Processing (NLP) in India as well as all around the world.

There are many approaches of machine translation system. At present, the most frequently used approaches of MT system are Rule Based MT, Statistical MT, Example Based MT and Hybrid MT [Islam *et al.* (2017)]. The different approaches of machine translation system are shown in Fig.1.

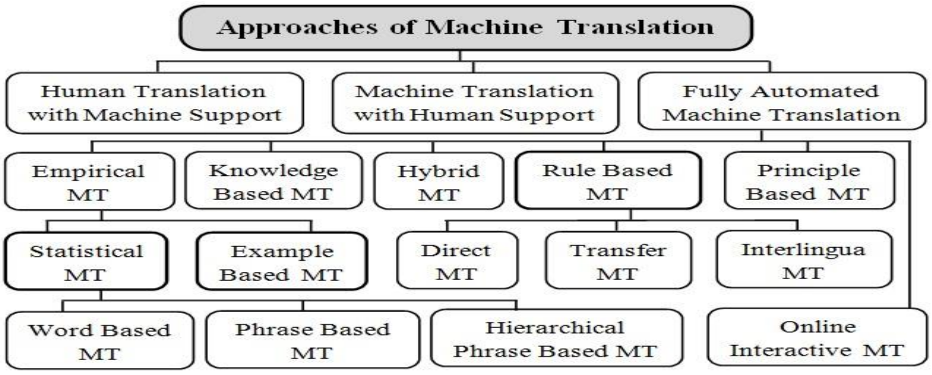


Fig. 1. Different approaches of MT.

1.1. *Natural language*

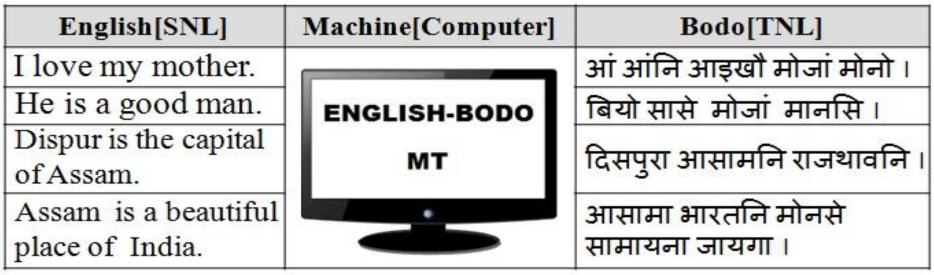
Language is an essential aspect of all human beings for communication. The languages which are used for human communication are called natural or human languages. In this section, two natural languages are briefly discussed as follows:

Bodo language is also pronounced as Boro language. Bodo is one of the famous natural languages of North-East India. It is mainly spoken by the people of North-East India and Nepal [Talukdar *et al.* (2012)]. The Bodo language is also known as Mech and is the fundamental language of Bodo people. It is the official language of Assam (Bodoland Territorial Council) and one of the recognized languages of India. The Bodo language is highly used by the maximum population of Kokrajhar, Chirang, Baksa, and Udalguri districts of Assam. This language is also used by some population of Cooch Behar, Alipurduar and Jalpaiguri districts of West Bengal. Devanagari script (Hindi script) is used to write the Bodo language and word order in this language is SOV (Subject +Object+Verb).

The English language was the first spoken language in England and now it is a global lingua franca [Islam (2016)]. English is spoken mainly by the population of Australia, Canada, Ireland, New Zealand, United Kingdom and the United States. It is an official language of sixty sovereign states and third most common native language in the world. The English language was introduced in India during the rule of the East India Company in 1830. In 1951, the Constitution of India declared Hindi as the primary official language and English as the associate official language of India. Now, it is the third most spoken language in India. Latin script is used to write the English language and word order in this language is SVO (Subject +Verb+Object).

1.2. *English to Bodo machine translation*

Machine translation is a very important and one of the major applications of NLP. Many MT research works have been developed and some are going on for Indian natural languages. Bodo is one of the natural languages of India. However, it has not sufficient corpus and no MT system is available for Bodo language. Therefore, we want to expand the computerized information (or corpus) for Bodo language and to develop English to Bodo MT system using a huge amount of General domain English-Bodo parallel text corpora, Phrase-Based SMT approach and Moses that it can produce high quality translation result from English to Bodo language. Some examples of sentences in English to Bodo MT system are shown in Fig.2.



English[SNL]	Machine[Computer]	Bodo[TNL]
I love my mother.	ENGLISH-BODO MT	आं आंनि आइखौ मोजां मोनो ।
He is a good man.		बियो सासे मोजां मानसि ।
Dispur is the capital of Assam.		दिसपुरा आसामनि राजथावनि ।
Assam is a beautiful place of India.		आसामा भारतनि मोनसे सामायना जायगा ।

Fig. 2. Examples of sentences in English to Bodo MT system.

2. Related Work

In this section, the prior works of MT system using SMT approach developed in the world and in India are briefly discussed.

A lot of machine translation research work has been developed by many institutions/organizations in many countries using the SMT approach on natural languages. Nowadays, the SMT approach has become very popular and mainly focuses on many MT works. The first idea of SMT approach was suggested by Warren Weaver in 1949 [Hutchins (1995)]. The first word based SMT system was developed by the researchers at IBM. They also developed the Candide project for French and English languages using SMT approach in 1988 [Kathiravan *et al*. (2016)]. The EuroMatrix project was begun between all the European Union languages using SMT approach in 2006 [Uszkoreit (2007)]. The Aachen University, Edinburgh University, and Southern California University are the main places for MT works using the SMT approach for natural languages. Recently, the Phrase-Based SMT approach is a successful technique and deeply used by many MT researchers. The Phrase-Based French to English Statistical Machine Translation was developed by Philipp Koehn using Moses at Edinburgh University [Brunning (2010); Koehn (2009)]. The English to Spanish Statistical Machine Translation was developed by Preslav Nakov at University of California [Nakov (2008)]. The English to Urdu Hierarchical Phrase Based SMT system was developed by Nadeem

Khan and his colleagues in Pakistan [Khan *et al.* (2013)]. The Google translate (2006) and Bing translate (2009) are developed by Google and Microsoft respectively, using the SMT approach to translate text between the various natural languages [George (2013)].

A large number of MT research works have been developed in India also using the SMT approach. Several organizations like Centre for Development of Advanced Computing (C-DAC), Technology Development for Indian Languages (TDIL), Ministry of Communications and Information Technology (MCIT), and educational institutions have developed many MT system using the SMT approach for Indian natural languages [Islam *et al.* (2017)]. A small number of machine translation projects like ANUVAADAK (IIT Bombay), E-ILMT (Consortium of Nine Institutions, 2006), and Shakti (2003) were developed using the SMT approach in India [Godase and Govilkar (2015); Antony (2013)]. Some examples of MT research works which are developed using SMT approach are mentioned below:

- Telugu to English Phrase Based Statistical Machine Translation System was developed by G. Lakshmikanth and B. Dhana Lakshmi, 2016 [Lakshmikanth and Lakshmi (2016)].
- English to Dogri Translation System using MOSES was developed by Avinash Singh, Asmeet Kour and Shubhnandan S. Jamwal, 2016 [Singh *et al.* (2016)].
- English to Malayalam Statistical Machine Translation System was developed by Aneena George, Adi Shankara College of Engineering and Technology, 2013 [George (2013)].
- Assamese to English Bilingual Machine Translation was developed by Kalyanee Kanchan Baruah, Pranjal Das, Abdul Hannan and Shikhar Kr. Sarma, Gauhati University, 2014 [Baruah *et al.* (2014)].
- English to Kannada Statistical Machine Translation system was developed by P.J. Antony, P. Unnikrishnan and K.P. Soman, 2010 [Antony (2013)].

3. Implementation of English to Bodo MT System

In this section, the approach, corpus preparation, and other steps are discussed to develop the English to Bodo MT system. The Phrase-Based Statistical Machine Translation (PBSMT) approach, Moses, and General domain English-Bodo parallel text corpora are used to implement the system.

3.1. *Statistical machine translation*

The statistical machine translation comes under Empirical or Corpus based machine translation which needs a very large amount of parallel text corpora in both the source and target languages to achieve high quality translation result. Essentially, this approach uses computing power to build sophisticated data models to translate text from one source natural language into target language. The SMT approach offers the best solution for ambiguity problems in natural languages than other MT approaches. It is language independent and disambiguates the sense automatically with the use of large quantities of parallel corpora. The advantages of SMT approach are easy to build and maintain, less requirement of linguistic knowledge earns knowledge from a corpus, reduces human efforts and time-saving [Koehn (2009)]. There are three categories of SMT approach, namely Word-Based SMT, Phrased-Based SMT and Hierarchal Phrased-Based SMT. The SMT approach contains three main components which are described below:

- Language Model (LM): The LM computes the probability of the target language (Bodo language) B, *i.e.* P(B).
- Translation Model (TM): The TM helps to compute the probabilities of the source language sentence E (English) for a given target language sentence B (Bodo), *i.e.* P(E|B).
- Decoder: The decoder maximizes the translation probability using the product of LM and TM probabilities, *i.e.* argmaxP(B)*P(E|B).

The architecture of English to Bodo machine translation system is shown in Fig. 3.

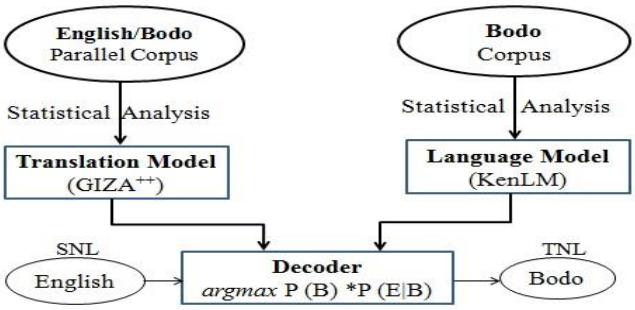


Fig. 3. Architecture of English to Bodo MT system.

3.2. *Phrase-based statistical machine translation*

A phrase is a collection of two or more words that stands together as a single unit. The Phrase-Based SMT approach is a more accurate and highly used in the SMT system nowadays. The PBSMT is the extended form of the Word-Based Statistical Machine Translation (WBSMT) and it has many advantages than WBSMT. The PBSMT approach allows the translation of non-compositional phrases and can handle many to many translations. Phrase translations are learned from data in an unsupervised way. In phrase based translation, each sentence of the source and target languages are fragmented into different phrases before the translation. In PBSMT, a word alignment follows certain patterns in both the source and target sentences which are almost similar to WBSMT [Brunning (2010); Koehn (2009)].

In the PBSMT approach, the following steps are performed to develop the system using SMT toolkit Moses and Perl language.

3.2.1. *Corpus construction and preparation*

Corpus is a collection of huge amount of texts in digital format of a particular natural language. We have constructed General domain English-Bodo parallel text corpus to train the proposed system. The General domain corpus means, the corpus contains the sentences which are commonly used in our daily life. An example of one parallel sentence in English-Bodo parallel corpus is as: Today is very hot (an English sentence) - दिनै जोबोद गुदुं (Bodo sentence). The parallel text corpus is constructed with 6000 (six thousand) parallel sentences of each English and Bodo language in the proposed system.

To train the English to Bodo MT system, two text files are prepared in UTF-8 format for English and Bodo corpus separately and the following pre-processing steps are performed for both the English and Bodo corpora.

- Tokenization: It is done to insert space between words and punctuation in both the corpora.
- True Casing: It is done to convert the first words of each sentence to their most probable casing for both the tokenized corpora.
- Cleaning: It is done for removing the long sentences, empty sentences and extra spaces from both the corpora.

3.2.2. *Language model*

The language model is an essential part of any SMT system. The LM is used to ensure the fluency of the translated sentences. In this system, the LM is built for Bodo corpus using the LM toolkit KenLM. The KenLM is inbuilt in Moses. The LM calculates the probability of sentences of Bodo language P(B) using the n-gram modeling technique. It decomposes the probability of a target sentence (Bodo sentence) as the probability of particular words P(w) using Markov Chain Rule [Brunning (2010); 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 words of Bodo language.

The n-gram technique uses the last n-1 words to compute the probability of the next word. The language model probability of a sentence is the product of the probabilities of all words in the sentence. In n-gram model, the size N=1, 2, 3,...., n are represented as uni-gram, bi-gram, tri-gram, ….., n-gram respectively. The n-gram probabilities can be computed in a straightforward manner $P(w_n|w_{n-2}w_{n-1})$ from the Bodo corpus. In the proposed system, we have used tri-gram model. The formula for calculating tri-gram probabilities (maximum likelihood) of sentences from the corpus is shown in Eq. (2).

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

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

Suppose, we want to find the probability of a sentence like राजुवा आसाम भुमफरायसालिनि सासे लेङाइ बिबुंगिरि from the given General domain Bodo text corpus using tri-gram (3-gram) language model.

The probability of the sentence is calculated by simply multiplying the tri-gram probabilities together which are found in the proposed system as shown as below:

P(<s> राजुवा आसाम भुमफरायसालिनि सासे लेङाइ बिबुंगिरि </s>)
=P(राजुवा | <s><s>) P(आसाम | राजुवा <s>) P(भुमफरायसालिनि | राजुवा आसाम) P(सासे | आसाम भुमफरायसालिनि) P(लेङाइ | भुमफरायसालिनि सासे) P(बिबुंगिरि | सासे लेङाइ) P(</s>) | लेङाइ बिबुंगिरि) P(<s> | बिबुंगिरि </s>)
=0.204 x 0.520 x 0.079 x 0.430 x 0.095 x 0.127 x 0.061 x 0.006
=0.000000015

Where, <s> and </s> are used to represent start and end symbol to every sentence and treated these as additional words in the corpus.

3.2.3. *Translation model*

The translation model is an essential component of any SMT system. The TM is used to ensure the adequacy of the translation result. In this system, it computes the probability of the source sentence (E) for a given target sentence (B), *i.e.* P (E|B), where E is the monolingual phrase or sentence of English corpus and B is the monolingual phrase or sentence of Bodo corpus. The TM calculates the probabilities of sentences by depending on the behavior of the sentences in the corpus. The translation model can be computed as the sum over all probabilities of all possible alignments (A) between two sentences of E and B [Lakshmikanth and Lakshmi (2016)] as shown in Eq. (3).

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

To train the translation model, the most necessary step is word (or phrase) alignment. An alignment is a many to many relationship between the words of a source sentence (E) and its corresponding translation in the target sentence (B). The TM toolkit, Giza++ is used for word alignment in the translation model. Since, the computation of TM probabilities is not possible at the sentence level, therefore, the sentence is broken down into small units of words or phrases and their probabilities are calculated [Lakshmikanth and Lakshmi (2016)]. A word (or phrase) alignment example of English to Bodo Phrase-Based translation model is shown in Fig. 4.

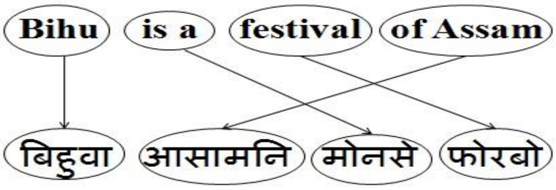


Fig. 4. Alignment example of English to Bodo Phrase-Based TM

3.2.4. *Decoder*

The decoder is an essential component of any SMT approach. The Moses decoder is used to find the maximum translation probability from the source language to the corresponding target language. The performance of the translation directly depends on the decoder in any SMT system. The Moses decoder decodes a source sentence into target translated sentence using LM and TM. The output results obtained from the LM and TM are fed into the decoder and finally, the decoder will find out the maximum translation probability in the proposed system using the following Eq. (4).

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

The decoder takes the text of English language as input and generates the text of Bodo language as output. The decoder uses A* search based on heuristic search method to find the best possible translation [Koehn (2016)]. The A* search is an efficient method to find the best possible translation in any SMT system than beam search and greedy search approaches [Och (2001)].

4. Result

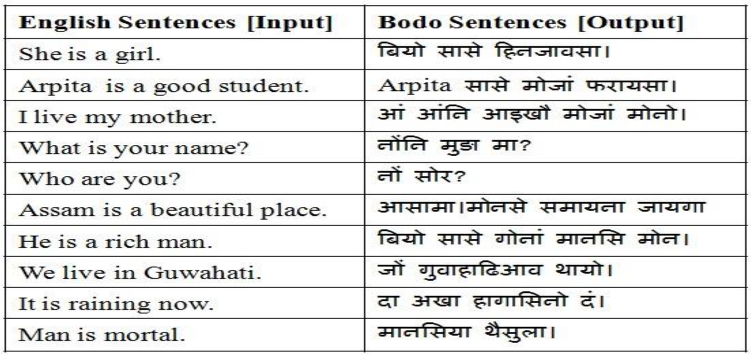
To get the translation result, the following command is used to execute the Moses decoder in the English to Bodo MT system.

```
  ~/mosesdecoder/bin/moses –f  ~/mert-work/moses.ini
<~/corpus/input.general.eng-bod.en > output.general.eng-bod.bd
```

Where, input.general.eng-bod.en is an input file of English text and output.general.eng-bod.bd is an output or translated file of Bodo text.

The English to Bodo MT system is examined several times with various numbers of General domain parallel sentences of English and Bodo languages and we have got various translation results. It has been observed that if we increase the size (number of sentences) of the given parallel corpora to train the system, then the quality of the translation result is also enhanced. Finally, we have used General domain English-Bodo parallel text corpora with 6000 (six thousand) sentences of each language to train the system. Examples of ten English-Bodo parallel sentences which are found as translation results in our system are shown in Table 1.

Table 1. English to Bodo translation result.



English Sentences [Input]	Bodo Sentences [Output]
She is a girl.	बियो सासे हिनजावसा।
Arpita is a good student.	Arpita सासे मोजां फरायसा।
I live my mother.	आं आंनि आइखौ मोजां मोनो।
What is your name?	नोंनि मुङा मा?
Who are you?	नों सोर?
Assam is a beautiful place.	आसामा।मोनसे समायना जायगा
He is a rich man.	बियो सासे गोनां मानसि मोन।
We live in Guwahati.	जों गुवाहाढिआव थायो।
It is raining now.	दा अखा हागासिनो दं।
Man is mortal.	मानसिया थैसुला।

5. Evaluation

In the proposed system, the accuracy of the translation result is evaluated in two methods which are briefly discussed below:

5.1. *Manual evaluation*

In the manual evaluation, we have taken ten English-Bodo parallel sentences to evaluate the accuracy of the translation which are found as translation results in our system as shown in the above Table 1. The translation accuracy is evaluated by a linguistic person Dr. Ismail Hussain, Assistant Professor, Department of Bodo, Bodoland University, Kokrajhar, Assam. He has evaluated the levels of translation accuracy (adequacy and fluency) from the given ten input and output sentences as shown in Table 2.

Table 2: Levels of translation accuracy (adequacy and fluency).

Levels	Definition	Number of sentences
Perfect	The translated sentence is very good to understand.	7
Fair	The translated sentence is easy to understand, but need a minor correction.	2
Acceptable	The translated sentence is broken, but is understandable.	1
Nonsense	The translated sentence is not understandable.	0

5.2. *Automatic evaluation*

In the automatic evaluation, BLEU (Bilingual Evaluation Understudy) technique is used to evaluate the quality of the translation result in the system. BLEU is an appropriate and a very useful method for automatic evaluation of any SMT system. It is developed by Kishore Papineni and his colleagues in 2001 [Koehn (2016); Uszkoreit (2007)]. It is based on the average of matching n-grams between a proposed translation and a reference translation and it seems to correspond well with human judgments on adequacy and fluency. The BLEU technique is inbuilt in Moses. The following command is used to find the BLEU score in the proposed system:

```
~/mosesdecoder/scripts/generic/multi-bleu.perl –lc      ~/corpus/training/general.eng-
bod.true.bd     < ~/working/output.general.eng-bod.bd
```

Where, the Bodo corpus “general.eng-bod.true.bd” is human or reference translation and “output.general.eng-bod.bd” is machine generated output or candidate translation.

To calculate the BLEU score, it has to count the number of n-grams in the candidate translation that have a match in the corresponding reference translations. The words of a candidate translation that match with a word in the reference translation are counted and then divided by the number of words in the candidate translation [Uszkoreit (2007)]. We have achieved 49.08 BLEU score in the proposed system. It has been observed that if the size of the given parallel corpus is increased to train the system, then the BLEU score would be relatively improved. A higher BLEU score denotes better translation.

6. Conclusion

Statistical machine translation approach is a very good solution for automatic translation of enormous text from one source natural language into another natural language. The main purpose of the proposed system is to implement English to Bodo MT system using a huge amount of General domain English-Bodo parallel text corpora that it can produce high quality and accurate translation result. To fulfill the purpose, the PBSMT approach, Moses, KenLM, N-gram technique, GIZA++, and BLEU technique have been used in the system. The proposed system has been examined with various sizes of General domain English-Bodo parallel text corpora and achieved different translation results. It has been observed that if the corpus size is large, then the accuracy of the translation will be good. We have achieved relatively good translation result using only 6000 (six thousand) parallel sentences of each English and Bodo language in the system. Since, the computerized information of Bodo language is very low. Therefore, it can be hoped that the proposed system would be helpful for students, research scholars and basically for Bodo people as well as other people of India and abroad.

References

Antony, P. J. (2013): Machine translation approaches and survey for Indian languages. Computational Linguistics and Chinese Language Processing, 18(1), pp. 47-78.

Baruah, K. K.; Das, P.; Hannan, A.; Sarma, S. K. (2014): Assamese-Englısh bılıngual machıne translatıon. International Journal on Natural Language Computing, 3(3), pp. 73-82.

Brunning, J. (2010): Alignment models and algorithms for statistical machine translation (Thesis). Cambridge University, UK.

George, A. (2013): English to Malayalam statistical machine translation system. International Journal of Engineering Research & Technology, 2(7), pp. 640-647.

Godase, A.; Govilkar, S. (2015): Machine translation development for Indian languages and its approaches. International Journal on Natural Language Computing, 4(2), pp. 55-74.

Hutchins, W. J. (1995): Machine translation: History of research and applications. University of East Anglia, UK.

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

Islam, S.; Devi, M.I.; Purkayastha, B.S. (2017): A study on various applications of NLP developed for North-East languages. International Journal on Computer Science and Engineering, 9(6), pp. 386-378.

Kathiravan, P.; Makila, S.; Prasanna, H.; Vimala, P. (2016): Over view- the machine translation in NLP. International Journal for Science and Research in Technology, 2(7), pp. 19-25.

Khan, N.; Anwar, W.; Bajwa, U. I.; Durrani, N. (2013): English to Urdu hierarchical phrase-based statistical machine translation. International Joint Conference on Natural Language Processing, pp. 72-76.

Koehn, P. (2009): Statistical machine translation (Book). Cambridge University Press, New York.

Koehn, P. (2016): MOSES (User Manual and Code Guide). Statistical machine translation system, University of Edinburgh, UK.

Lakshmikanth, G.; Lakshmi, B. D. (2016): An approach for Telugu to English Phrase-Based Statistical machine translation system International Journal of Magazine of Engineering, Technology, Management and Research Applications, 5(5), pp. 617-627.

Nakov, P. (2008): Improving English-Spanish statistical machine translation: Experiments in domain adaptation, Sentence paraphrasing, Tokenization, and Recasing. Proceedings of the third Workshop on statistical machine translation, pp. 147-150, USA.

Och, F. J.; Ueffing, N.; Ney, H. (2001): An efficient A* search algorithm for statistical machine translation. Computer Science Department, RWTH Aachen University of Technology, Germany, pp. 1-8.

Singh, A.; Kour, A.; Jamwal, S.S. (2016): English to Dogri translation system using MOSES. Circulation in Computer Science, 1(1), pp. 45-49.

Talukdar, J.; Sarma, C.; Talukdar, P.H. (2012): Automatic syllabification rules for Bodo language. International Journal of Computational Engineering Research, 2(6), pp. 110-114.

Uszkoreit, H. (2007): Survey of machine translation evaluation. EuroMatrix Project, Germany, pp. 1-80.