

Addressing the Issue of Unavailability of Parallel Corpus Incorporating Monolingual Corpus on PBSMT System for English-Manipuri Translation

Amika Achom¹, Partha Pakray² and Alexander Gelbukh³

¹ National Institute of Technology, Mizoram, Aizawl, India,
achomamika01@gmail.com

² National Institute of Technology, Mizoram, Aizawl, India,
parthapakray@gmail.com

³ Centro de Investigación en Computación (CIC)
of the Instituto Politécnico Nacional (IPN), Mexico, Mexico
gelbukh@cic.ipn.mx

Abstract This research paper work establishes an important concept of improving Phrase based Statistical Machine Translation System incorporating monolingual corpus on the target side of the English to Manipuri translation language pair. However, there has been no work that focuses on translating one of the Indian Minority Tibeton-Burman Manipuri language pair. This Phrase based Statistical Machine Translation system has been developed using the Moses open-source toolkit and evaluated carefully using various automatic and human evaluation techniques. PBSMT achieves a BLEU Score of 10.15 as compared to the baseline PBSMT of BLEU Score 9.89 using the same training, tuning, and testing datasets. This research paper work addresses the issue of limited availability of parallel text corpora (English-Manipuri pair).

Keywords: Phrase Based Statistical Machine Translation, English to Manipuri Parallel Corpora, Automatic-Bilingual Evaluation UnderStudy Score (BLEU), Human Evaluation-Fluency, Adequacy, Overall Rating

1 Introduction

Machine Translation (MT) is the automatic translation from one source language to any another target language. MT is one of the most dominant emerging fields of today's world. MT is a key to success to any new services. This has promoted the development of many new models and resulted in the development of an open source Statistical Machine Translation (SMT) system, Moses,⁴ which is used in various institutes, research project and so on.

According to the latest Census Survey of India⁵, it is reported that, 1.5 million people speak Manipuri language or Meiteilon. This arises the need of

⁴ <http://www.statmt.org/moses/?n=Development.GetStarted>

⁵ <https://en.wikipedia.org/wiki/Languages-of-India>

MT technology to translate any low-level Manipuri language. There are different approaches to MT. They are undermentioned:

Rule-based Machine Translation (RBMT) is one of the earliest approaches of MT. In RBMT, human have to develop and handle rules using different grammatical conventions, lexicon [1]. One of the advantages of RBMT is that it is very simple. There are different approaches of RBMT, namely: Transfer-based RBMT, Interlingua-based RBMT, and Dictionary-based RBMT. One of the limitation of RBMT is that human need to create rules for every analysis and generation stage that proves to be quite tedious and cumbersome.

Thus, the failure of rule-based approaches led to the dominant application of corpus-based Machine Translation. It may be due to the increasing availability of machine readable text. There are different approaches of corpus-based Machine Translation namely:

1. Example-based Machine Translation
2. Statistical Machine Translation
3. Phrase based Statistical Machine Translation
4. Tree-based Statistical Machine Translation
5. Neural based Machine Translation

Example-based Machine Translation (EBMT) is one of the SMT approaches that is based on the analogy. The bilingual corpus act as its main source of knowledge base [1]. Given a new test sentence, it is translated using examples or analogy from the knowledge base. The translated sentences are then stored back into the knowledge base. This saves the effort of translating on every new test sentences. One of the limitation to this approaches is that if there is any unmatched test sentences, then it need to be regenerated from the scratch. It cannot use the concept of close phrases or neighboring word to predict the translation of unmatched words [1, 10].

SMT is a data-driven or corpus-based approach. The idea of SMT comes from the information theory. It translates any document according to the probability distribution. The probability distribution applies the Bayes' Theorem.

$$P(m|e)=P(e|m) * P(m) \quad (1)$$

$P(e|m)$ in Equation (1) gives the probability that the source string is the translation of the target language. This component is called the translation model in SMT. $P(m)$ gives the probability of getting the target string. This particular component is called the language model in SMT.

In phrase-based model of SMT, phrases are considered to be the atomic units of translation. It translates a sequences of words (or phrases) instead of word by word. The use of the phrases as a unit of translation allows the model to learn from local reordering i.e., local word orders and local agreements, idiomatic collocations, deletion and insertion by memorization that are sensitive to the contexts. The PBSMT segments the source sentences into words or phrases with uniform distribution. It then, translates each of the English phrases into Manipuri phrases according to phrase translation model estimated from the training data. Phrase

based model uses the joint probability distribution $P(m,e)$. $P(m)$ and $p(e|m)$ are mapped into feature of log-linear model. Finally, PBSMT needs to reorder the translated output in order to improve the fluency. Phrases longer than 3 words do not improve the quality of translations and do not follow the reordering concept has been proved in the experimental research findings of Koehn, Och, and Marcu in 2003.

The remainder section of the paper work is organized as follows: Section 2 describes the existing and past works on SMT and PBSMT. Section 3 explains the system architecture and methodology of the improved PBSMT system with a detailed explanation on the working of PBSMT decoder. Section 5 explains the evaluation and experimentation performs on PBSMT system. Section 6 describes the analysis result and the relative comparison between the baseline PBSMT system and the improved PBSMT system. Section 7 concludes the paper work and points the direction towards the future research works.

2 Related Works

The rapid advancement and steady progress in corpus-based machine translation reported the increasing growth of research works in the field of SMT. However, large corpora do not exist for many language pairs especially for low-level Asian languages. Thereby, we proceed our research work by making the best use of the existing parallel corpora for English to Manipuri language pair translation.

Large parallel text corpora could be mined from the web for the low density language pair has been shown in [8].

The development of phrase based SMT system trained on Europarl data has been shown in [4]. The developed system integrates the news data corpora with the translation model and language model. It results in the significant improvement of different feature weights.

A recent work on SMT incorporating syntactic and morphological processing has also been reported in [7]. The translation system has been developed for English to Hindi language pair. The developed paper work demonstrates that the baseline PBSMT system has been improved by incorporating syntactic and morphological processing with the addition of suffixes on the target side of the translation language pair.

A research paper on Manipuri-English bidirectional SMT system has also been reported. This paper work is based on factored model approach of SMT. The paper work has shown the improvement of SMT system using domain specific parallel corpora and the incorporation of morphological and dependency relations in [10].

A Manipuri to English MT system based on examples has also been reported in [10].

Incorporation of Morpho-syntactic information between the inflected form of the lemmas improves the translation quality has been proposed in [5].

The use of pivot language (English) to bridge the source and target languages in the translation process address the issue of scarcity of data resources has been proved in [11].

The unavailability of parallel corpus for many low density language is one of a big hindrances to the progress of research works. In [9], it has shown that the integration of the inflectional and derivational morphemes for Manipuri language using factor-model based SMT, solves the issue of limited availability of parallel corpora for English to Manipuri language pair translation.

Thus, the main focus of this research paper work is to build and improve the baseline PBSMT system. This requires an enormous amount of parallel corpus. But, we are limited to work with only the available text corpora. Our paper work tries to address all these issues by incorporating monolingual corpus on the target side of Manipuri language.

3 System Architecture of Improved PBSMT

There are various open source toolkits for MT system. They are Joshua, cdec, Apertium, Docent, Phrasal and so on. We use Moses toolkit for our research work [4].

This section of the paper will discuss on how we trained the system using PBSMT model. The remainder part of this section will explain the working of Phrase-based decoder and its underlying concept on how we perform the translation from English to Manipuri. The rest part of this section will discussed on how to incorporate monolingual corpus on the target side of the language pair during language model.

3.1 Architecture of PBSMT based on SMT-Moses

SMT based Moses system is independent of any language pair. There are two main components in SMT Moses system namely: training pipeline and decoder. The training pipeline is a collection of tools and utilities (mainly written in perl and in C++). The training pipeline can accept any raw data (parallel corpora and monolingual corpus) and can train the translation model.

An important part of the translation model is language model. It is a statistical model built using monolingual corpus on the target side of the translation language pair. In our research work, we incorporate Manipuri monolingual corpus while building the language model. Moses include KenLM language model creation program Implz. There are some freely available toolkits for language model creation namely: IRSTLM, SRILM, RandLM, OXLM, NPLM, Bilingual N-gram LM (OSM), Dependency Language Model (RDLM) and so on. In SMT, language model toolkit performs two important task: training and querying. In our work, we train the language model with KenLM, it produce an arpa file and query with blm file. The core of the translation model is the phrase translation table, that is learned from the parallel corpora. In PBSMT, training start with the alignment of the word. For this, SMT Moses use the GIZA++. Once the

words are aligned, phrases pair of any length that are consistent with the aligned words could be extracted from the phrase translation table. The final step in the creation of translation model is tuning where different statistical models are weighted against each other to produce the best possible translation. Moses use the Minimum Error Rate Training (MERT) algorithm for tuning task [4]. SMT

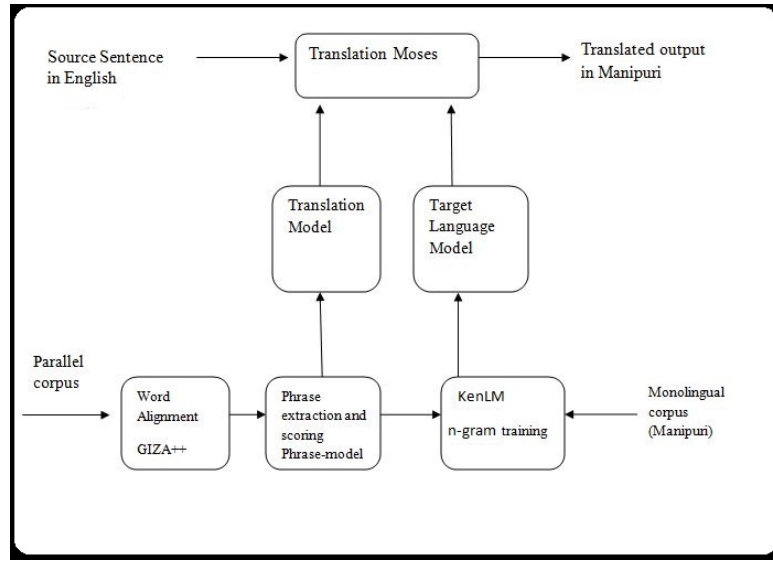


Fig. 1: System architecture of PBSMT system

Moses system includes decoder Pharaoh which is one of a freely available tools for research purposes. Decoding in SMT Moses is a beam search process over all the possible segmentation of the phrases; translation probabilities for each phrases; and reordering cost. The decoder component in Moses tries to find the highest scoring sentence in the target language from the phrase translation table corresponding to a given source sentence. The decoder can rank the translated candidate sentences [3, 4]. The basic architecture of PBSMT system is shown in Figure 1. The methodology and control flow of the PBSMT system using Moses toolkit are explained as below:

1. Corpus collection and splitting the corpus for training, tuning and testing
2. Preprocessing the collected corpus

3. Language model training by incorporating monolingual corpus
4. Training the English-Manipuri translation system
5. Tuning the model and working of Phrase-based decoder
6. Running and testing PBSMT system
7. Calculating BLEU Score of PBSMT

3.2 Corpus Collection and Splitting the Corpus for Training, Tuning and Testing

This is one of the foremost tasks for any system training. We collect different parallel and monolingual text corpora from different news portal⁶ and from TDIL⁷ program, which is a program initiated by the Ministry of Electronics and Information Technology(MiETY), Government of India. TDIL program allows the creation and accession of multilingual resources. The collected corpus is then splitted into different sets for our system training. We collected 9565 parallel training sentences or Bitexts corpora for English to Manipuri language pair, tuning or the validation dataset consists of 1000 parallel text sentences, testing dataset consists of 100 sentences and finally, we used monolingual corpus of 1,39,411 sentences for training the language model. The information on the size and the number of sentences that we use exclusively for our system training are given in Table 1 and Table 2.

Table 1: English corpus description

Sl no	Type	Training Data	Tuning Data	Test Data	Monolingual Corpus
1	Sentences	9565	1000	100	1,39,411
2	Size	1.3 MB	141 KB	12.4 KB	26.1 MB

Table 2: Manipuri monolingual corpus description

Sl no	Type	Training Data	Tuning data	Test Data	Monolingual Corpus
1	Sentences	9565	1000	100	1,39,411
2	Size	3.9 MB	404.6 KB	36.4 KB	26.1 MB

⁶ <http://e-pao.net/>

⁷ <http://ildc.in/Manipuri/Mnindex.aspx>

3.3 Preprocessing

We need to preprocess the data before training the system. For this, we tokenize, truecase and clean the data. During tokenization, a space is inserted in between the words and punctuation. MT system need to be trained on lowercase. During recasing, the initial words in each sentence is converted to their most probable lowercase character. Finally, we cleaned the data. This is done to remove all the long, misaligned and the empty sentences. Then, we limit the length of the sentences in each lines to 80 words.

3.4 Training the Language Model Incorporating Monolingual corpus

Training the language model is one of foremost steps in any translation system. We have very limited parallel data. In our work, we are trying to make the best use of the available limited resources. This is achieved through the incorporation of Manipuri monolingual corpus into PBSMT system during language model training. For this, we modify the system by tuning and tweaking some of the parameters, that are used during the language model training. We used the KenLM language modeling toolkit for interpolating the monolingual corpus. It is very fast and consumes very less memory. It is compiled by default and distributed along with Moses toolkit. It use the `lmplz` program to estimate language models with modified Kneser-ney smoothing. The end result of language model training is an `.arpa` file [4]. We further query the language model by converting the `.arpa` file format into `.blm` format.

3.5 Training the English-Manipuri Translation System

We train the improved PBSMT System. Training algorithm (`train-model.perl`) is used by PBSMT system. The underlying control flow for training the PBSMT system are undermentioned:

3.5.1 Running GIZA++ for word alignment

GIZA++ is a freely available toolkit for aligning the word. The aligned words are extracted from the intersection of bidirectional runs of GIZA++ and some additional alignments point from the union of the two runs. To establish a proper alignment of the words, different heuristics approaches may be applied in [4]. Moses used the default parameter `grow-diag-final` parameter to run GIZA++ for aligning the word. The aligned words or phrases generated by the directional runs of GIZA++ for the translation of English to Manipuri language pair translation are shown in Table 3. The “+++” entries in the Table 3, indicates the intersection of the words or phrases during the bidirectional run of GIZA++.

Table 3: Alignment of English-Manipuri phrases using GIZA++

Sl no	মসিগী (0)	girl(1)	অসি(2)	য়াম ্ না(3)	beautiful(4)
this(0)	+++				
girl(1)		+++			
is(2)			+++	+++	
very(3)			+++	+++	
beautiful(4)					+++

3.5.2 Creating Lexical Translation Table

From the Phrase translation table, we can easily estimate the maximum likelihood probability $p(m/e)$ and inverse likelihood of $p(e/m)$ using the Noisy channel model of Bayes' theorem.

3.5.3 Extract Phrase and Score Phrase

In this, all the extracted phrases are stored into one big file. We process one phrase at a time and then compute $p(e/m)$, for Manipuri phrase m . Similarly, $p(m/e)$ is also estimated, for every English phrases e . Inorder to compute the scoring function, we use the lexical weighting function, word penalty, and phrase penalty.

3.5.4 Building Reordering Model

We used the distance-based reordering model. It is used to assign a cost linear to the reordered distance in the target output during language model training. For example, skipping of two word is assigned a cost higher than skipping one word. Another advanced model known as lexicalized reordering model can also be used to reorder the target output.

3.5.5 Building Generation Model

The model is built on the target side of the parallel corpus. We build the generation model for our improved PBSMT by incorporating monolingual corpus. The candidate translation C^s from the monolingual corpus of the target language can be expressed in Equation (2)

$$\log((P_{LM}(C^s)) = \sum_{j=1}^{C^s} \log P(C_j^s | C_1^s, C_2^s, C_3^s, \dots, C_{j-1}^s) \quad (2)$$

3.5.6 Creating Moses Configuration file

After successfully training the model, a moses.ini configuration file is obtained. This file is used to tune the parameters further in the following subprocess.

3.6 Tuning the Model

This is one of the most time consuming tasks of the translation process. There are various tuning algorithms that are available namely: MERT, PRO, MIRA, Batch MIRA. We used the MERT algorithm (Minimum Error Rate Training) for our system training. The main idea behind tuning is to learn the weights of different discriminative model and produce the best possible translations in [4].

3.7 Mathematics of Phrase-based Model and Working of PBSMT Decoder

Equation (3) represents the basic PBSMT equation:

$$m_{best} = \arg \max_m P(m|e) = \arg \max_m [P(e|m)P_{LM}(m)] \quad (3)$$

Translation with the highest score is given by m_{best} . $P(m|e)$ and $P_{LM}(m)$ are translation model and language model respectively.

$$P(\bar{e}_i|\bar{m}_i) = P(\bar{e}_1, \bar{e}_2, \bar{e}_3, \bar{e}_4, \bar{e}_5, \bar{e}_6, \dots, \bar{e}_I|\bar{m}_1, \bar{m}_2, \bar{m}_3, \bar{m}_4, \bar{m}_5, \bar{m}_6, \dots, \bar{m}_I) \quad (4)$$

$$= \prod_{i=1}^I \phi(\bar{e}_i|\bar{m}_i) * d(start_i - end_i - 1 - 1) \quad (5)$$

The LHS indicates the probability of a sequence of I phrase in the sentence m, given I phrases in sentence e. ϕ is the phrase translation probability and $d(start_i - end_i - 1 - 1)$ gives the distortion probability or the reordered distance induced by the target phrases in PBSMT system in [2].

The various control flows of Phrase-based decoder are undermentioned :

1. It extract the n-gram segments or phrases from the input source sentence.
2. It matches the extracted phrases from the phrase table.
3. It retrieved the aligned phrase along with their translation probabilities.
4. It compute the score of each hypothesis.
5. It compute the language model probabilities and distortion probabilities.
6. Lastly, future word penalty can be added so that translation do not get too long or too short. This factor indicates the difficulty in translating.

The phrase model comprises of two important components. They are Translation Phrase table from where the decoder consults how to translate the phrase and moses.ini configuration file obtained successfully after running the tuning stage which gives the updated trained weights. PBSMT decoder uses the beam search algorithm to prune the hypothesis and generate the N-best possible candidate translation [2, 4]. Finally, we reorder the distance using distortion model. Let us take an example to understand the decoding process using PBSMT.

Example 1

Input: This(0) girl(1) is(2) very(3) beautiful.(4)

Output: (PBSMT Decoder + Monolingual corpus) : মসিগী |0-0| girl |1-1| অসি য়াম
 ্ না |2-3| beautiful. |4-4|

From this example, it can be concluded that word or phrases in the translated output are generated from the corresponding input words or phrases.

Example 2

Input: Their(0) countries(1) wares(2) is(3) many(4) years(5) now.(6)

Output: (PBSMT Decoder + Monolingual corpus) : মথোয়গী |0-0| ্ লাজাদগী |2-2|
 লৈবাকশিং |1-1| অসি |3-3| চহি কয়ামরুম |4-5| now. |6-6|

This example shows that the word “countries” moved forward one step ahead in the translated output and “wares” move backward one step. The phrase “many years” is translated to the phrase “চহি কয়ামরুম” in the translated output sentence.

Finally, we perform the product of the below parameters to score the PBSMT model. They are undermentioned:

1. $P_{LM}(\text{মসিগী নুপীমচা})$ and $P_{LM}(\text{নুপীমচা অসি})$
2. Translation Probability: $P(\text{মসিগী নুপীমচা}|\text{this girl})$ and $P(\text{অসি য়াম ্ না ফজৈ}|\text{is very beautiful})$
3. Distortion Probability or the reordering distance.

Mathematically, the cost of translation in PBSMT can be expressed using Equation (6):

$$P(m|e) = \phi(e|m)^{weight_{phi}} * LM^{weight_{LM}} * L(e, m)^{weight_d} * w(e)^{weight_{phi}} \quad (6)$$

Accordingly, the phrase based decoder will pick the best possible translation according to the probability score and translate the corresponding source sentence to the target sentence.

3.8 Running and Testing the PBSMT System

After successfully building the PBSMT system, we can further run the PBSMT system and perform the testing operation on some new test sentences. Table 4 highlights the translated output sentences generated after successfully running and testing the phrase-based decoder.

3.9 Calculating BLEU Score

To calculate the BLEU Score, we used the automatic software script (multi-bleu.perl). It is observed that the improved PBSMT achieved a BLEU score of 10.15 incorporating monolingual corpus as compared to the baseline PBSMT of BLEU score 9.89 only. This shows a significant improvement. Thus, we can improve the baseline PBSMT system by incorporating monolingual corpus. This paper work addresses the issue of scarcity of parallel text corpora.

Table 4: Running and Testing PBSMT Decoder

Sl no	Input test sentence	PBSMT translated output
1	You will see every-thing in the night	নহাক ্ না 0-0 উবা 1-2 পুম 3-4 ্ 5-5 night. 6-6
2	I am at home in evening.	অযুক পুং 1-1 । 0-0 ্ 2-2 evening. 5-5 লৈবা 3-4
3	What did you do last night?	What 0-0 ্ না লৈরম 1-1 ্ বা 3-3 নহাক ্ না 2-2 অরোইবা 4-4 night? 5-5
4	Have you seen me?	নহাক ্ না 1-1 me? 3-3 ্ না 0-0 উবা 2-2
5	Ram is a good person.	রাম 0-0 অসি য়াম ্ না 1-2 অফবা 3-3 person. 4-4

4 PBSMT System Evaluation and Experimentation

Evaluating the translated output generated from the PBSMT system is a very important concern of SMT Technology. SMT relies on two very important evaluation metrics. They are automatic metrics and human evaluation.

4.1 Automatic Evaluation

Automatic evaluation is useful in situation when we have limited amounts of parallel text corpora, less candidate test sentences and reference translations. In general, every automatic metrics should be able to correlate easily with the human evaluation. There are various automatic technique for evaluating the MT system. They are namely: BLEU, NIST, Word Error Rate, TER, GTM, METEOR and LEPOR. Among all these, we used the Bilingual Evaluation UnderStudy (BLEU) Score for our experiment.

BLEU Score is a text based metric. It is used for evaluating the quality of the translated output generated by the MT system. The main goal of BLEU Score evaluation is to compare and count the number of matches between the n-gram of the reference translation with the n-gram of the candidate translated sentence. So more the number of matches in the n-gram, closer the candidate translation is to the human translation in the citation [6]. In SMT system, calculation of BLEU score uses the sentence Brevity Penalty (BP) multiplicative factor, hyp_{len} (candidate translation sentence length) and ref_{len} (reference translation sentence length) in [6]. We assume c to be the length of the candidate translation and r be the length of the reference corpus. Equation (7) shows the computation of Brevity Penalty(BP) factor.

$$BP = \begin{cases} 1 & \text{when } c > r \\ e^{(1-r/c)} & \text{when } c \leq r \end{cases} \quad (7)$$

Using Equation (7) in Equation (8), we can calculate the BLEU Score

$$BLEU = BP * \exp(\sum_{n=1}^N W_n \log P_n) \quad (8)$$

4.2 Human Evaluation

Evaluation of the PBSMT System by humans is also equally important. Although the human evaluation may be prone to error but this cannot be neglected. This is a necessary task in order to establish a valid consistency with the obtained BLEU Score. In this evaluation, we used the adequacy-fluency strategy for the human evaluation. Table 5 show the analysis of different output generated by PBSMT and the way to assign rating to different test sentences.

4.2.1 Adequacy-Fluency

The adequacy factor indicates how much of the information from the original reference translation is expressed in the candidate translation. Accordingly, the evaluators assign score or rating to the test sentences. The adequacy score is also called as “fidelity” in SMT system.

Fluency indicates how natural a translated sentence sounds to the native speaker of the target Manipuri language. It involves the grammatical correction and the choice in the proper order of the word in [6].

Let us consider the following English-Manipuri translated sentence :

Example 1

Input Sentence: Tomba has a house.

Reference Sentence: Tomba য়ুম অমা লৈ

Output Sentence: অসি house অমা লৈ

The PBSMT system cannot recognize the name of a person. So, Tomba is translated to “অসি”, “house” is translated to “house”. The translated sentence do not convey the appropriate meaning. It cannot identify the agent of the action “who bought the house?”. Thus, it is not an adequate translation.

However, the word order and syntactic structure of the target grammar is maintained in the final translated output. Therefore, we can conclude that it is a fluent translation.

Example 2

Input Sentence: It is a very nice scenery.

Reference Sentence: মসি অসি য়াম্মা ফজবা দ্ৰিশা অমনী

Output Sentence: মসি ্ অসি য়াম ্ না ফজবা scenery.

The PBSMT system cannot translate the word “scenery” here. But, much of the meaning is convey by the translated sentence in this particular example. Thus,

Table 5: Analysis on the PBSMT Translated Output and the Way to Assign Rating

Sl no	Reference Sentence	English Sentence	PBSMT	Rating
1	ঐখ্যোগী অহাওবা দোসা	Our delectable dosas	ঐখ্যোগী অহাওবা দোসা	VERY GOOD
2	মসিগী অশংবা টুরিজম প্রো-জেক	This green tourism project	মসিগী অশংবা টুরিজম project	GOOD
3	দমপুং ফাবা অকুপ্লা মরোল-শীং অমসুং এপ্লিকেসন ফো-মেটশীংগীদমস্তা	For the complete details and application formats	তুরিষ ্ মপুং ্ অকুপ ্ পা অমসুং ্ লিকেসন থাখিবা formats	WORST
4	ভরন অসি হৌজিককী তামিল নাদুগী মফম ২২ দা লৈ	Bhawan is present in 22 places in Tamil Nadu	ভরন অসি হৌজিক ্ কী ২২ দা মফম তামিল নাদু	AVERAGE
5	য়ান্না কুইনা মতম মতমদুদা	For a long time,	মসি য়াম ্ না কুইনা মতম মতমদুদা,	GOOD
6	উচেক মতুনা অমসুং বারগী শউনগী লৈমায় অসিনা মসি-গী হোটেল অসিগী মওং-মতৌ হেনা ফজহলি	Feather and a bar with a leather floor makes this lavish hotel a design beautiful.	উচেক মতুনা অমসুং বার ্ শউনগী লৈমায় অসিনা মসিগী য়াম ্ না হোটেল মওং-মতৌ ফজবগী.	GOOD
7	কনাপুয়া অমনা নুমিদাংরাইর-মগী সমুদ্রা নুঙাইথোকলিঙৈ-দা	When one is enjoying an evening by the seaside.	মতম অসি অসিদি নুমিদাংরা-ইরম অমা ্ সমুদ ্র .	AVERAGE
8	তাকপা মনি-মুক তা লৈবা মসিগী য়ুমশারোলগী রায়েল অসি লর্ড কোন্মেমেরা সুইতে;	The jewel in its architectural crown is the lord Connemera suite;	তাকপা মনি-মুক ্ তা লৈবা মসিগী য়ুমশারোলগী রায়েল ্ অসি lord Connemera suite;	AVERAGE
9	ঐখ্যোগী অহাওবা মোরোক.	Our delectable pepper.	ঐখ্যোগী [0-0] অহাওবা [1-1] মোরোক. [2-2]	VERY GOOD
10	নহাক্সা নুমিদাংরাইরমদা লে-ায়না উবা পুন্মক ফংনি	You will see everything in the night.	নহাক ্ না [0-0] উবা [1-2] পুম [3-4] ্ [5-5] night. [6-6] [6-6]	WORST

it is an adequate translation.

In addition, the syntactic structure and the word order is also preserved by the translated sentence. Therefore, the example considered above is both an adequate and a fluent translation.

Example 3

Input Sentence: This is hotel and pickles.

Reference Sentence: মসি হোটেল অমসুং আছারনী

Output Sentence: মসি |0-1| হোটেল |2-2| অমসুং |3-3| pickles. |4-4|

The PBSMT system cannot translate the word “pickles” here. But, much of the meaning is convey by the translated sentence in this particular example. Thus, it is an adequate translation.

In addition, the syntactic structure and the word order is also preserved by the translated sentence. It is also an easily readable and a fluent translated sentence. Therefore, the example considered above is also both an adequate and a fluent translation.

Example 4

Input Sentence: Konar Mess named after a community in Tamil.

Reference Sentence: কোনর মেস অসি তমীলগী খুগং অমগী মমিং লৌদুনা থোনখিবনি

Output Sentence: Konar Mess মমিং লৌদুনা থোনখিবনি ্ গোয়না তামিল .

The PBSMT system cannot recognize the word “Konar Mess”. The translated sentence is very hard to understand and correlate properly. Moreover, the translated output sentence is totally disorder. Therefore, the example considered above is not an adequate and a fluent translation.

4.3 PBSMT System Error Analysis

Analyzing the PBSMT system error is one of the main subsequent task of MT system experimentation and evaluation. We carefully examined the various types of errors generated by the PBSMT system and they are briefly undermentioned:

4.3.1 Word Order Divergence

The order of the word in English sentence is subject+verb+object. Manipuri grammar follows the word order subject+object+verb. The example discussed below show the divergence in the syntactic structure or the order of the word.

Consider the example discussed below:

Example 1

Input Sentence: Limited menu.

Translated Output Sentence: মেনু অকক ্ নবা

In this, input sentence limited menu, the subject of the sentence is implicit (It). Menu refers to the object. Thus, it follows the subject+verb+object word order. The corresponding translated output sentence in Manipuri is: “মেনু অকক ্ নবা”. Therefore, it can be inferred that, the word order in Manipuri sentence is subject+object+verb.

Thus, we can infer from this example that English and Manipuri language differs significantly in their word order.

Let us consider another example below:

Example 2

Input Sentence: He is the best.

Translated output Sentence: মহাক ্ না |0-0| best. |3-3| অসি |1-2|

Considering the translated sentence: “he” is subject and it is translated to “মহাক ্ না” in Manipuri, which is a subject, “best” is object and “is” is verb and it is translated to “অসি” in Manipuri. Thus, we can conclude that word order in English language follows the subject+verb + object and the Manipuri language follows subject+object+verb word order.

4.3.2 Unable to Handle Consonant Conjunct

Manipuri language is rich in morphology and it is highly agglutinative in nature. We observe that PBSMT system cannot handle the conjunction of two consonants. It generate the consonant separately with a character ্ in between.

Let us consider an example given below:

Example 1

Input Sentence: Ram is a very good person.

Output Sentence: Ram অসি য়াম্মা অফবা

Here, the word “য়াম্মা” (yaamna) that represent the conjunction of two consonant ম and ন cannot be represented by the PBSMT system. The two different consonant character ম and ন is separated by an UNK character ্ in the translated output. Instead it should be clustered together to form a single glyph ম্ন in the translated output.

Therefore, this example establishes the fact that, PBSMT system cannot handle the conjunction of two consonants.

Consider another example given below:

Example 2

Input Sentence: You have amazing of lemon.

Output Sentence: নহাক |0-0| ং |1-1| ং |3-3| চম ্প ্প রা . |4-4| অঙ্কপা |2-2|

In this example, the word “চম্প” or “champra” or “lemon” represents the conjunct of two consonants প and র. PBSMT system cannot handle the formation of glyph. It cannot generate the word “চম্প”. PBSMT generates the character প and র separately instead of, combining these two character together to form a single character conjunct ষ্প in the final translated output sentence. Therefore, PBSMT system cannot handle the conjunction of consonant. This need to be solved.

Let us consider another example in different scenario:

Example 3

Input Sentence: Limited menu.

Output Sentence: মেনু. অকক ্ নবা

In the example discussed above, the PBSMT system cannot handle the conjunction of two similar consonants ক and ক to produce a glyph “ক্ক”. PBSMT cannot translate the word “অক্কনবা (limited)”. This is also one of the limitations of PBSMT system. It need to be handled.

4.3.3 Unable to Handle Case Markers and Suffixation

The PBSMT system cannot handle the suffixation of suffix or case markers on the target side of the translated output sentences. It cannot identify the agent or the doers of the action in the final translated sentence. For example, the suffixes -ni (- নী) , -dbni (- দবনী) , -ri (-রী) , re (- রে), (-oi/-ওই) and so on, cannot be suffixed by the system to construct a perfect translation of the input sentences.

Let us consider another example.

Example 1

Input Sentence: He like double malai.

Output Sentence: মহাক ্ না |0-0| দবল |2-2| malai. |3-3| ং অমগুম ্ না |1-1|

The translated output sentence do not clarify who like “double malai”. Therefore, the PBSMT system cannot identify the agent of the action in the translated sentence.

Let us consider another example.

Example 2

Input Sentence: This is hotel and pickles.

Output Sentence: মসি |0-1| হোটেল |2-2| অমসুং |3-3| আছার. |4-4|

The output translated sentence would be a perfect and a complete sentence with 100 % adequacy and 100 % fluency, if the PBSMT system were able to affix the suffix-ni (-নী) to the translated sentence. Thus, we can conclude that PBSMT system cannot handle the affixation of suffix to the translated output sentences.

4.3.4 Simple and Compound Sentence

The study of Statistical Machine Translation from English to Manipuri Language pair translation infers that simple sentence are translated without error as compared to the complex sentence.

The translation of simple sentence given below illustrates the above concept.

Example 1

Input Simple Sentence: Last edited

Output Simple Sentence: অকোনবা শেষদোক

Each of the word in the source sentence is translated to their corresponding word in the target language. Therefore, PBSMT system can translate the simple sentence quite easily.

Example 2

Input compound Sentence: We enjoy hot chicken and mutton food.

Output compound Sentence: ঐখোয়না |0-0| নুংঙাইনা |1-1| অশাবা |2-2| য়েন |3-3| অমসুং য়াও |4-5| food. |6-6|

Each of the word in the source sentence is translated to their corresponding word in the target language. Only the meaning of the word “food” cannot be translated by the PBSMT system. Therefore, PBSMT system can translate the compound sentence to a great extent.

Example 3

Input complex Sentence: Eat straight from the plantain leaf with a wide range of pickles to accompany.

Output complex Sentence: চাবগী ওইবা লমকোই খোঙচত ্ ্ লাদা মনাদা অমা মখল য়াম ্ লবা আচাৰশিং ্ টেনিয়াৰিং ্ নবা

We can see the difference from the above considered two example. PBSMT system translate the complex or the long sentence with the the character ্ in between the alphabet or the character. We can see there are 5 ্ character in the translated output. It is quite difficult to convey the meaning of the translation in the complex sentence even. Therefore, PBSMT cannot handle the complex sentence correctly.

4.3.5 Unable to Translate Some Words

The study of Statistical Machine Translation from English to Manipuri language pair shows that some words are not able to translate by the PBSMT System. The corresponding target words need to be transliterated further from the lexicon. The translation given below illustrates the above concept.

Example 1

Input Sentence: What rubbish!

Output Sentence: কঁৰি rubbish!

Here the words “rubbish” cannot be translated even though it is a simple sentence.

Example 2

Input Sentence: This girl is very beautiful.

Output Sentence: মসিগী girl অসি য়াম ্ না beautiful.

Here the words “beautiful” and “girl” cannot be translated by the PBSMT System.

Example 3

Input Simple Sentence: First aid kit and manual

Output Simple Sentence: অহানবা হীদাকশিং অমসুং manual

Here the word “manual” cannot be translated by the PBSMT System.

5 Result Analysis and System Comparison

1. Analysis on the Basis of Inter-Rater Reliability or Agreement:

In this experimentation, we evaluate and compare the performance of two different system. The first system S1 is the baseline PBSMT system and system S2 is the improved PBSMT system build incorporating monolingual corpus. We calculate the Percentage(%) Agreement/Inter-Rater Reliability from the rating assigned by different evaluators. We represent the calculated Inter-Rater Reliability along the Y-axis of the graph and the agreement between the evaluator along the X-axis. This is depicted clearly in the figure. From Figure 2, it is observed that the Percentage(%) Agreement/Inter-Rater Reliability between different evaluators converges more in System 2 which is indicated by the UPPER line for the improved PBSMT system. It has been observed that there is much more convergence in the opinion of different evaluators while evaluating the improved PBSMT system as compared to the baseline PBSMT.

2. Analysis on the Basis of BLEU Score:

In the next analysis, we used the automatic metrics BLEU Score to eval-

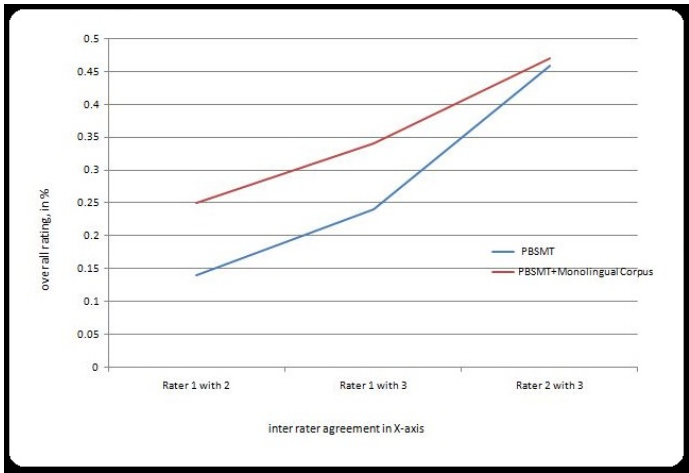


Fig. 2: System Performance Comparison based on Evaluator Inter-Rater Reliability

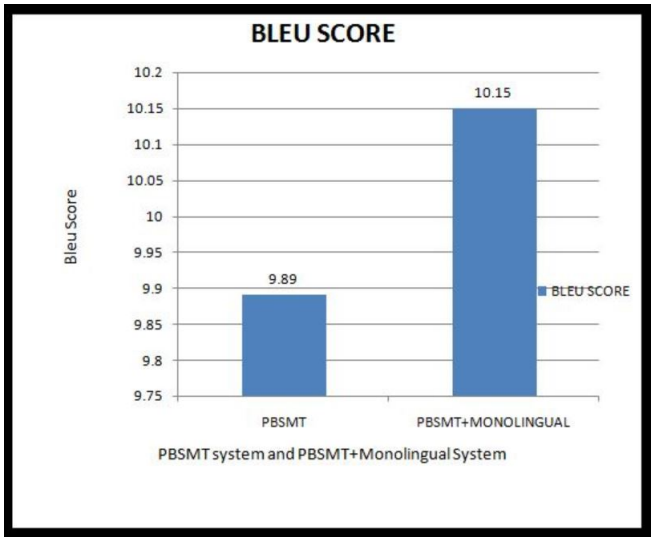


Fig. 3: System Performance Comparison based on BLEU Score

Table 6: Mean of the Overall Rating for Adequacy and Fluency

Sl no	System	Evaluator	Adequacy	Fluency
1	PBSMT Baseline	Evaluator1	1.73	1.56
2	PBSMT Baseline	Evaluator2	2.01	1.99
3	PBSMT Baseline	Evaluator3	2.48	2.32
4	PBSMT+Monolingual	Evaluator1	1.86	1.89
5	PBSMT+Monolingual	Evaluator2	2.13	2.07
6	PBSMT+Monolingual	Evaluator3	2.52	2.46

uate the performance of two different system. Moses system uses the script multi-bleu.perl to generate the BLEU Score of the developed two different systems. It can be observed from Figure 3, that the improved PBSMT system achieved a BLEU Score of 10.15 in comparison to the baseline PBSMT system of BLEU Score 9.89 using the same training, tuning and testing datasets. The difference in the improved BLEU Score by a factor of 26% shows that a marked significant achievement by the System S2 as compared to System S1. From these research findings, we can conclude that more the BLEU score the more is the convergence to the Inter-Rater Agreement among the evaluators.

3. Analysis on PBSMT Translated Output:

In the last part of analysis section, we have conducted some comparison with respect to fluency on different output translated by the baseline PBSMT system and the improved (PBSMT+Monolingual corpus) system. The different parameters has been evaluated from various feedback posted by different evaluators during the evaluation session.

Let us consider an example below to illustrate the above concept:

Input Sentence : What rubbish!

Output sentence (System S2 PBSMT+monolingual corpus) : কব্রি (What) rubbish!

Output Sentence (System S1 Baseline PBSMT) : মসিগী (This) rubbish!

It can be seen clearly from the example that system S2 improves the fluency and readability of the translated output sentence as compared to the output sentence translated by the baseline PBSMT system S1. The native speakers find the translated word “কব্রি” to be more appropriate and suitable in this particular context as compared to the translated word “মসিগী”, which is translated by the system S1.

Thus, the system S2 (PBSMT+monolingual corpus) outperforms the baseline PBSMT System. From this observation, it can be concluded that system S2 improves the fluency and readability of the translated output. Table 6

shows the mean of the overall rating of adequacy and fluency rated by the evaluators.

6 Conclusion and Future Works

The study on this research paper work clearly indicates that the interpolation of monolingual corpus on the target side of the English to Manipuri language pair translation significantly improves the BLEU Score and also the fluency of the translated output sentences. The research work conducted on SMT using Phrase-based approach could be more improved with more training data sets. However, we perform PBSMT system training by making the best use of the limited available resources. This paper work addresses the issue of the unavailability of the parallel text corpora. This is accomplished through the incorporation of Manipuri monolingual corpus. We interpolated monolingual corpus during the language model training. It achieves an improved BLEU Score of 10.15 as compared to the baseline PBSMT system of 9.89 only. Besides these, from the feedback and suggestion posted by different evaluators, the fluency and quality of the translated output improves drastically. This is one of the major contribution to our research findings on PBSMT system. Another point worthwhile to be mentioned, output of the translated output and even the BLEU Score could be enhanced further. This would have been achieved through the incorporation of more morphological information and case markers on the target side of the translation language pair in the citation [9]. We could still focus on the post editing task to handle the system error and after the correction of these translated output sentences, we could feed the correct sentences during retraining the system. Thereby, PBSMT system could be trained to learn more correctly and able to predict accurately on the new test sentences.

Acknowledgments I would like to express my deepest appreciation to the Technology Development for Indian Languages (TDIL) Programme, initiated by the Ministry of Electronics and Information Technology, Govt. of India for sharing the valuable parallel corpus on English to Manipuri Language pair and the monolingual corpus in Manipuri Language for this research paper. Furthermore, I would like to extend my heart full gratitude to the Department of Computer Science and Engineering, National Institute of Technology, Mizoram for providing me the required financial assistance and the laboratory facilities for conducting out the full experimental research works on this research paper.

References

1. Antony, P.: Machine translation approaches and survey for Indian languages. *International journal of computational linguistics and Chinese language processing* 18(1), 47–78 (2013)
2. Dave, S., Parikh, J., Bhattacharyya, P.: Interlingua-based English–Hindi machine translation and language divergence. *Machine Translation* 16(4), 251–304 (2001)

3. Hoang, H., Koehn, P.: Design of the mooses decoder for statistical machine translation. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing. pp. 58–65. Association for Computational Linguistics (2008)
4. Koehn, P.: Machine Translation System User Manual and Code Guide (2011)
5. Nießen, S., Ney, H.: Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics* 30(2), 181–204 (2004)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
7. Ramanathan, A., Hegde, J., Shah, R.M., Bhattacharyya, P., Sasikumar, M.: Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In: IJCNLP. pp. 513–520 (2008)
8. Resnik, P., Smith, N.A.: The web as a parallel corpus. *Computational Linguistics* 29(3), 349–380 (2003)
9. Singh, T.D.: Addressing some Issues of Data Sparsity towards Improving English-Manipuri SMT using Morphological Information. *Monolingual Machine Translation* p. 46
10. Singh, T.D., Bandyopadhyay, S.: Manipuri-English Example Based Machine Translation System. *International Journal of Computational Linguistics and Applications (IJCLA)*, ISSN pp. 0976–0962 (2010)
11. Utiyama, M., Isahara, H.: A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In: HLT-NAACL. pp. 484–491 (2007)