English to Manipuri and Mizo Post-Editing Effort and its Impact on Low Resource Machine Translation

Loitongbam Sanayai Meetei[1], **Thoudam Doren Singh**[1], **Sivaji Bandyopadhyay**[1],
Mihaela Vela[2], and **Josef van Genabith**[2,3]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
[2]Dept. of Language Science and Technology, Saarland University, Saarbrücken, Germany
[3]DFKI, Saarbrücken, Germany
{loisanayai,thoudam.doren,sivaji.cse.ju}@gmail.com
m.vela@mx.uni-saarland.de, josef.van_genabith@dfki.de

Abstract

We present the first study on the post-editing (PE) effort required to build a parallel dataset for English-Manipuri and English-Mizo, in the context of a project on creating data for machine translation (MT). English source text from a local daily newspaper are machine translated into Manipuri and Mizo using PBSMT systems built in-house. A Computer Assisted Translation (CAT) tool is used to record the time, keystroke and other indicators to measure PE effort in terms of temporal and technical effort. A positive correlation between the technical effort and the number of function words is seen for English-Manipuri and English-Mizo but a negative correlation between the technical effort and the number of noun words for English-Mizo. However, average time spent per token in PE English-Mizo text is negatively correlated with the temporal effort. The main reason for these results are due to (i) English and Mizo using the same script, while Manipuri uses a different script and (ii) the agglutinative nature of Manipuri. Further, we check the impact of training a MT system in an incremental approach, by including the post-edited dataset as additional training data. The result shows an increase in HBLEU of up to 4.6 for English-Manipuri.

1 Introduction

In our increasingly globalized world, communication plays a vital role and with it, demand for translation between different languages is on the rise. Despite much progress, machine translation (MT) on its own may not always be sufficient to meet the demand. MT output may sometimes be erroneous and needs to be checked and corrected. The use of translation technology such as MT systems, translation memories (TM) and CAT tools can boost translation productivity (Koehn, 2009; Plitt and Masselot, 2010). However, limited numbers of professional translators for a language pair can be a major challenge, especially for low resource languages.

Raw MT output is not always exempt from errors. Often post-editing MT output (where a human translator reviews and where required corrects MT output) is the most productive approach to translation. PE effort is the amount of effort required to generate a reasonable target text from MT output. Following Krings (2001) PE effort can be subdivided into **temporal effort**, **technical effort** and **cognitive effort**. Temporal effort represents the overall time taken to complete a PE task. Technical effort can be measured tracking keyboard and mouse interactions, including insertion, deletion, mouse movement, etc. Cognitive effort (considered the most difficult to measure) involves mental effort such as reading and understanding the text, identifying errors, and the decision making process towards correcting errors.

To date, PE research has mainly concentrated on a few well-studied languages. In this work, the same source data in English are machine translated to low resource languages to carry out a PE task. The dataset used in the experiment consists of news articles collected from a local daily newspaper, Imphal Free Press[1] in Manipur, a Northeastern state of India with a population of around 3 million[2] and a geographic size of 22,327 sq. km. The collected news corpus is originally in English and then machine translated into three

[1] https://ifp.co.in/

[2] http://censusindia.gov.in

Proceedings of the 17th International Conference on Natural Language Processing, pages 50–59
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

English	Manipuri	Mizo	Hindi
SVO	SOV	OSV	SOV
Roman	Bengali	Roman	Devanagari

Table 1: Typological Word Order and Script of Languages in the Study.

Acronyms: O = Object, S = Subject, V = Verb.

target languages Manipuri, Mizo and Hindi resulting in three parallel datasets. The basic word order of the languages involved in our experiment along with their scripts are listed in Table 1.

We conduct a study on the PE effort required to produce reasonable target text in English-Manipuri and English-Mizo. As there is no commercially available machine translation system for these two languages, for comparison we also studied English-Hindi PE effort on commercial MT output on the same dataset. Two levels of PE are generally distinguished: light and full. For our experiments, we instruct our post-editors to carry out light PE to achieve the desired level of output quality. With various PE effort indicators computed using the data captured from the CAT tool, we carry out an experiment to measure the PE effort for English-Manipuri, English-Mizo and English-Hindi MT systems. Lexical words and function words are observed to have a different impact on the PE effort on the MT output for the three language pairs. We also carry out an experiment to test the impact of training a machine translation system for the low resource language pairs English-Manipuri and English-Mizo in an incremental manner, i.e. by adding the PEed dataset to the original MT training data. The rest of the paper is structured as follows: Section 2 reviews previous research, Section 3 desrcibes the PE task and the human PEers, Section 4 presents our approach and system set up. Section 5 details our findings. Section 6 summarizes our main results and avenues for further research.

2 Related Work

Early studies on the correlation between PE effort and various aspects of PE include O'Brien (2005). O'Brien studied the temporal, technical and cognitive effort involved in PE by analyzing keyboard-data using Translog and Choice Network Analysis (CNA). Several studies investigated bi-lingual PE and monolingual PE. In bilingual PE (Zampieri and Vela, 2014) post editors have access to the source text, while in monolingual PE (Nitzke, 2016) MT output is edited without the source text. Zampieri and Vela (2014) studied the use of TMs generated by MT output and their effect on human translation. The authors reported a significant increase in translation speed while using the TM as compared to translating from the scratch.

Similarly, Toral et al. (2018) show that post-editing an MT dataset involves less effort than translating from scratch. Post-editing MT output increases the productivity of the translators. Zaretskaya et al. (2016) examine various types of MT errors and the challenges they present for PE. Burchardt et al. (2013) compile a corpus consisting of English to German translation generated by different types of MT systems. The dataset is then annotated for translation errors using the MQM error typology, with only one error in each sentence. As the dataset is already annotated, the post-editors could skip the effort of identifying the errors and concentrate only on the highlighted error text segment in the PE process. Focusing on how PE effort changes with the different types of MT errors, the authors reported a weak correlation between PE time and PE effort. The authors also report that no direct dependency was found between the temporal and technical PE effort. Investigating the various types of PE operations for French to English and English to Spanish translation outputs, Popovic et al. (2014) reported lexical edits as the main factor in PE time.

Koponen et al. (2012), study the cognitive effort of post-editing MT output based on measuring PE time and HTER (Snover et al., 2006). HTER (Human-targeted Translation Edit Rate), is an automatic metric that computes the minimum number of edits required to change MT output into the post-edited version. The authors reported that the absolute PE time increases with the number of printable keystrokes and sentence length while seconds per word remain relatively constant. Despite the fact that HTER captures the difference between the final translation and raw

MT, it does not disclose much of the time and keystroke effort required to produce the final result. A similar study is also reported by Moorkens et al. (2015) where the human (or H-) variants of the reference based similarity measure such as BLEU (Papineni et al., 2002) is used to analyze PE effort.

Singh and Bandyopadhyay (2010a); Singh (2013) focus on MT for English to Manipuri, Pathak et al. (2019) on English to Mizo and Singh et al. (2017); Meetei et al. (2019b) on English to Hindi, using different MT approaches. But, to date there is no report on PE effort required to turn raw MT output for these target languages into useful translations. To address this gap in the literature, our paper investigates different aspects that impact PE effort and time spend to generate a reasonable target text from MT into Manipuri, Mizo and Hindi.

3 Description of the PE Task

Two post-editors who are native speakers of the target languages and also proficient with the source language are employed for each of the language pairs to carry out the PE task. For English-Manipuri and English-Hindi, the post-editors are undergraduate students of Computer Science and Engineering and for English-Mizo, the post-editors are postgraduate students of Science. When PEing machine translated text, it is important to clearly define what level of output quality is to be achieved. Generally two PE levels are distinguished: light or complete. In our work, the post-editors are asked to carry out light PE with the following instructions: 1) Using the maximum possible amount of raw MT text in the output of PE. 2) Ensure no addition or omission of source content. 3) Restructuring output, where the meaning is inaccurate.

4 Methodology and Experimental Design

We use an English language corpus collected from a local daily newspaper as the source text. We normalize the data in a pre-processing step. The normalized text is then machine translated into three different target languages using different MT systems. After post-editing a sample dataset of the machine translated text using a CAT tool, we study the data collected from the CAT tool to analyze PE effort and the time required to generate a reasonable target text. A pictorial representation of our experimental design is shown in Figure 1. The remainder of this section details individual steps in our approach.

	Sentences	Tokens
Total Dataset, D	64976	1688440
Sample Dataset, D_{PE}	200	5500

Table 2: Statistics of our collected dataset and data partitioning.

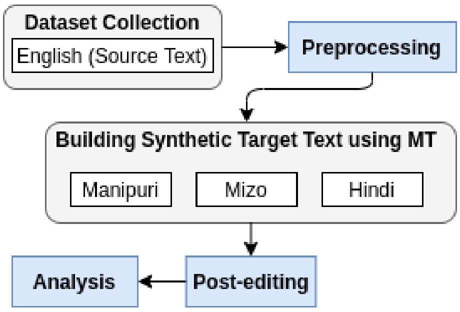

Figure 1: Experimental design.

4.1 Data Collection

The dataset used in our experiment is collected from a local daily newspaper based in Manipur, Imphal Free Press[3]. The news articles are in English. The complete dataset consists of 3770 news articles from the period July 2011 to October 2019 comprising 64976 sentences. We randomly select 200 sentences (D_{PE}) for our PE experiment. The statistics of the dataset and data partitioning are shown in Table 2. The dataset is collected using a web-scrapper built in-house.

4.2 Pre-processing

Data collected from the web is not free from noise. The pre-processing step includes removal of non-ascii special characters. Each of the news articles in our dataset is split into sentences using the Moses tokenizer (Koehn et al., 2007).

4.3 Building Machine Translated Target Text

We build a machine translated dataset using MT systems resulting in three language pairs,

[3] https://ifp.co.in/

	Sentences	Tokens	Types
en-mn	18070	*en*:390141; *mn*:358947	*en*:27891; *mn*:54611
$mono_{mn}$	131755	2798317	270998
en-mz	7500	*en*:86353; *mz*:87511	*en*:4301; *mz*:6151
$mono_{mz}$	1005675	29338218	312062

Table 3: Dataset for PBSMT systems.
en : English, *mn* : Manipuri, *mz* : Mizo.

namely, English-Manipuri, English-Mizo and English-Hindi.

4.3.1 English to Manipuri and Mizo MT

Manipuri and Mizo are the lingua francas of Manipur and Mizoram, two neighbouring north-eastern states of India. Both Manipuri and Mizo are low resource languages. Limited availability of data in Manipuri and Mizo is one of the main reasons that hamper the development of NLP systems for the language. The training datasets used for training the MT system for the languages are shown in Table 3. On the same English-Manipuri training dataset, we first examine the performance of MT systems trained with Phrase Based Statistical Machine Translation, PBSMT (Koehn et al., 2003) and the RNN-based NMT with attention mechanism (Bahdanau et al., 2014). The trained MT systems are evaluated on a held-out test dataset of 900 sentences. The result shows a BLEU score of 6.45, (34.7/9.6/3.5/1.5) on the PBSMT system while the NMT system achieved a BLEU score of 0.00, (11.8/0.3/0.0/0.0). For this reason, we use PBSMT systems for both English-Manipuri and English-Mizo MT systems as our parallel NMT results are substantially worse in these low-resource scenarios. To build language models for the target languages, we used the dataset in (Singh and Bandyopadhyay, 2010b) and (Meetei et al., 2019a) for Manipuri and Mizo respectively. mgiza[4] is used to generate the phrase table and srilm[5] to build the language model.

4.3.2 English to Hindi MT

In order to translate the English dataset to Hindi, we use Google Translate which is a Neural Machine Translation (NMT) system.

4.4 Post-editing

To investigate PE effort, we randomly select a subset of 200 sentences from the original English data and automatically translate it into the three target languages. We create a translation memory (TM) for each of the language pairs to prepare the source and MT output data for use with a CAT tool. The resulting TMs are uploaded in a commercial CAT tool[6].

4.4.1 PE effort indicators

During the post-editing process using the CAT tool, we record post-editing logs capturing Seconds per Word, Time to Edit and Post-editing Effort for each sentence. We measure:

1. *Post Editing Time (PET)*: Total time taken to edit a sentence in the target language.

2. *Post Editing Effort[6] (PEE)*: Post-editing effort expended on the machine translated output to produce the desired target text per sentence. *PEE* is computed based on edit distance measured in words obtained using a heavily customized version of the Levenshtein distance algorithm (Levenshtein, 1966).

3. *Seconds per Word (SpW)*: The *PET* spent by the translator to post-edit divided by the number of tokens of the post-edited translation.

4. *Total number of tokens (TT)*: Total number of tokens per sentence in the source language.

5. *Noun Words (NN)*: The word content that can be used to refer to a named entity, quality or action.

6. *Lexical Words (LW)*: Lexical words per sentence in the source language. Lexical words are the essential building blocks of a language's vocabulary. Lexical words are nouns, adjectives, verbs, and adverbs.

[4] https://github.com/moses-smt/mgiza

[5] http://www.speech.sri.com/projects/srilm/

[6] https://www.matecat.com

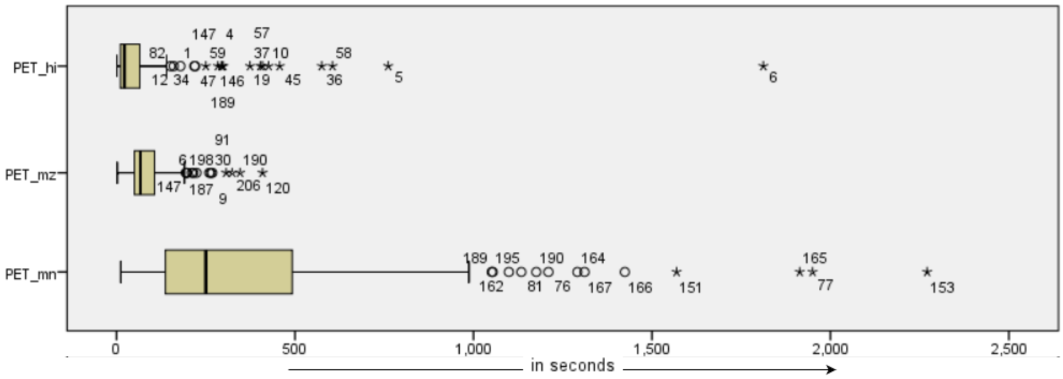

Figure 2: Distribution of Post-Editing Time (*PET*) for Manipuri, Mizo and Hindi

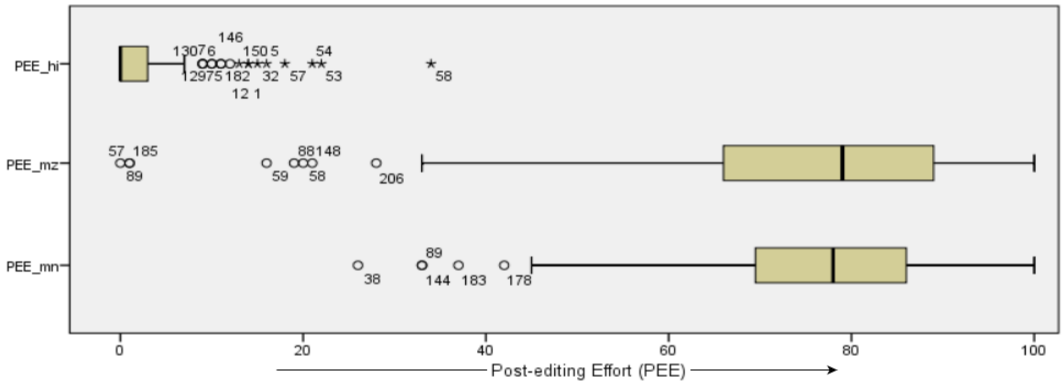

Figure 3: Distribution of Post-Editing Effort (*PEE*) for Manipuri, Mizo and Hindi

7. *Function words (FT)*: Function words per sentence in the source language. Function words are those words which are more grammatical in nature, such as articles, prepositions, etc. Here, $FT = TT\text{-}LW$.

Temporal effort is measured by the overall time taken *PET* while *PEE* represents the technical effort.

4.4.2 Descriptive Statistics and Correlation

We use mean, standard deviation as well as box plots to capture descriptive statistics of our datasets and results. To investigate whether variables co-vary we measure the correlation coefficient r with value between -1 to 1. A positive correlation shows the degree to which variables increase or decrease in parallel, while a negative correlation indicates that one variable increases as the other decreases.

5 Results and Discussion

To measure PE effort, PE logs are collected from the CAT tool after post-editing the raw MT output (in Manipuri, Mizo and Hindi) of the sample dataset, D_{PE}. We compare the general distribution of *PET* and *PEE* for each of the 200 sentences in the language pairs in the form of box plots. Mean and standard deviations for the rest of the source PE effort indicators (*TT, NN, LW, FT*) and the target PE effort indicators (*PET, PEE, SpW*) are computed and we investigate correlations between indicator variables.

5.1 Statistics and Correlation

Figure 2 and 3 show the distribution of post editing time (*PET*) and post editing effort (*PEE*) for the language pairs investigated. The box plots show:

- the minimum value, maximum value, first, second, third and fourth quartile of the experimental measures. The thick line represents the median.
- *Outliers*: these are the values that lie beyond the whiskers of a box plot. Outliers are marked by circles or asterisks along with their observation number. A circle represents an outlier (a value that appears to be outside of what is expected

for the observations), while asterisks represent extreme outliers (a value which is far away from what is expected).

Figure 2 shows that the *PET* of Manipuri (PET_mn) deviates far from the *PET* of Mizo (PET_mz) and Hindi (PET_hi). A likely cause of this is the massive amount of post-editing required in the output from English-Manipuri MT system combined with time spend on typing the Bengali script on the keyboard. Bengali scripts are used mostly by the news reporters while for daily communication, Roman scripts are used. To verify our findings, the English-Manipuri post-editors typed a set of randomly selected 50 English-Manipuri parallel sentences from the training corpus to measure the typing speed of Roman and Bengali scripts. 3001 seconds are spend on typing the English text consisting of 1153 tokens resulting in an average typing speed of 2.6 seconds per token. While for the Manipuri text with 1148 tokens, 5610 seconds are spend on typing, resulting in an average typing speed of 4.8 seconds per token. This led our post-editors to spend more time while post-editing a large portion of the translated text. The post-editing effort (*PEE*) for the three languages are shown in Figure 3. While the PEE for Hindi (PEE_hi) is small, the *PEE* of Manipuri (PEE_mn) and Mizo (PEE_mz) are very large with a maximum value of 100. This shows that massive effort is required in the post-editing task of Manipuri and Mizo resulting from low performance of the current state of the art of English-Manipuri and English-Mizo MT systems.

Mean and Standard Deviation (*SD*) of the indicators in the source language (English) of the dataset D_{PE} are shown in Table 4. To identify the lexical words (*LW*), we POS-tag the data using the Stanford Log-linear Part-Of-Speech Tagger[7].

Table 5 summarizes the the descriptive statistics, mean and SD of the indicators (*PET, PEE, SpW*) for our Manipuri, Mizo and Hindi experiments. Compared to Mizo and Hindi, *SpW* for Manipuri is longer. The main reason for this is the large portion of text post-edited and because of the difficulty in writing Bengali characters by the post-editors. The result in Table 5 shows a mean value ≈ 76.35 and ≈ 75.02 in the *PEE* for English-Manipuri and English-Mizo dataset respectively. Much of this is due to the current state of the art of English-Manipuri and English-Mizo MT systems. We note that significant effort is required to improve the English-Manipuri and English-Mizo MT system which requires a *PEE* > 55 in all the cases.

	NN	*LW*	*FT*	*TT*
Mean	9.64	15.39	9.16	24.56
SD	6.06	7.51	5.28	11.85

Table 4: Descriptive Statistics of the source text of dataset D_{PE}.

Acronyms: *NN*: Nouns, *LW*: Lexical Words, *FT*: Function words, *TT*: Total tokens

	Mean	*SD*
PEE_{mn}	**76.35**	13.45
PEE_{mz}	75.02	19.22
PEE_{hi}	2.29	4.64
PET_{mn}	**400.78**	378.93
PET_{mz}	91.00	63.75
PET_{hi}	72.48	168.14
SpW_{mn}	**16.14**	13.55
SpW_{mz}	3.51	2.57
SpW_{hi}	2.93	6.14

Table 5: Descriptive Statistics of target text.

Subscripts- $_{mn}$: Manipuri, $_{mz}$: Mizo, $_{hi}$: Hindi.

5.2 Correlation

In our experiment, the correlation between PE effort indicators is computed using Pearson's correlation coefficient to determine whether there is a potential dependency between them. The correlations among the indicators for the three language pairs involved in our experiment are shown in Table 6.

A Pearson's r data analysis shows a significant positive correlation between PEE_{mn} and *FT* (p<0.05, *r*= .16) and also between PEE_{mz} and *FT* (p<0.01, *r*= .27). The result also shows a significant (p<0.01) negative correlation between PEE_{mz} and *NN*. The main reason for the above result is because Mizo uses the same script as the source text and the agglutinative nature of the Manipuri text. In Manipuri function words such as articles (a, the), prepositions (on, at, in), etc. are suf-

[7] https://nlp.stanford.edu

	NN	LW	FT	TT
PEE_{mn}	0.041	0.088	0.162*	0.128
PEE_{mz}	-0.300†	-0.066	0.274†	0.080
PEE_{hi}	0.049	-0.065	-0.047	-0.062
PET_{mn}	0.344†	0.407†	0.405†	0.438†
PET_{mz}	0.620†	0.646†	0.452†	0.611†
PET_{hi}	0.234†	0.213†	0.256†	0.249†
SpW_{mn}	-0.038	-0.067	-0.043	-0.062
SpW_{mz}	-0.100	-0.202†	-0.322†	-0.271†
SpW_{hi}	-0.040	-0.106	-0.105	-0.114

Table 6: Correlation of source text and target text indicators. ***Note:*** † significant at 0.01 level of significance. * significant at 0.05 level of significance.

HBLEU	1-g	2-g	3-g	4-g	Average
MT_{mn}	33.8	10.0	4.0	2.0	7.16
MT_{mz}	34.8	10.7	5.5	4.0	9.48
MT_{hi}	96.5	94.0	91.5	89.2	92.78

Table 7: Evaluation against post-edited dataset D_{PE}.

fixed to the noun words in most of the cases, resulting in the formation of a new word.

The PET for all the language pairs involved is observed to be positively correlated with all the source text indicators.

In terms of seconds per word, SpW, only the English-Mizo pair is significantly negatively correlated with LW, FT and TT. With an increase in the number of tokens in the source text, the average time spent per token decreases. A likely cause is the use of same script.

In addition to the computation of correlations between indicator variables, we also calculated automatic MT evaluation (4-gram HBLEU) scores between the raw MT outputs and their post-edited versions of dataset D_{PE} for each language pair as shown in Table 7.

5.3 A Control Experiment

As, to the best of our knowledge, this is the first paper to report on PE research on Manipuri, Mizo and Hindi, it is not clear how the results obtained compare with previous research on well-resourced languages. Furthermore, as the PE for Manipuri, Mizo and Hindi did not involve professional translators, but students who are native speakers of our target languages with excellent command of English, we conducted a control PE experiment with German as target language and professional translator trainees at Saarland University. Our data consists of the same dataset selected for the Manipuri, Mizo and Hindi experiments, translated into German by DeepL[8], and PEed by seven Translation Study MA students with native German from the English to German translation track of the degree. We used the same CAT tool as in our Manipuri, Mizo and Hindi experiments and collected the same set of measurements.

For English-German we measure mean values of 48.94 in PET and 7.14 in PEE, compared to 72.48 and 2.30 for Hindi (see Figures 2 and 3). As both German and Hindi are well supported languages (both Google Translate and DeepL are some of the strongest performing systems for the EN-Hi and EN-DE language pairs), this provides additional support that the Hindi PE results we report are reliable and properly indicative of the task. Further, and in turn, this “anchoring” of the Hindi PE results through the German PE results, supports our belief that the large gap between the Hindi and with that of Manipuri and Mizo results observed in our experiments is also reliable, and can be traced to the fact that Manipuri and Mizo are much less well supported by language technologies and data (here machine translation) than Hindi or German.

5.4 Training English to Manipuri and Mizo MT systems on additional PEed data

Further, in an effort to improve the English-Manipuri and English-Mizo MT systems, we train the PBSMT systems for the language pairs in an incremental approach. We use the PEed dataset of (D_{PE}) for each language pair as additional training data to retrain our PB-SMT systems ($MT\text{-}I_{mn}$ and $MT\text{-}I_{mz}$). We further increase the additional training data of English-Manipuri [$D_{PEed\text{-}2} = 200$ (D_{PE}) + 656] to retrain English-Manipuri PBSMT system ($MT\text{-}I2_{mn}$) but could not acquire the same for English-Mizo due to the lack of post-editors. In order to check the quality improvement in the translated text, we compare the

[8] https://www.deepl.com/

	Sentences	Tokens	Unique tokens
$D_{PEed\text{-}2}$	856	20309	4884
D_{Ev}	50	434	293

Table 8: Dataset to retrain and evaluate MT systems. [D_{Ev}: Evaluation Dataset]

HBLEU	1-g	2-g	3-g	4-g	Average
MT_{mn}	21.9	4.6	0.6	0.3	2.14
$MT\text{-}I_{mn}$	22.6	5.0	0.9	0.4	2.45
$MT\text{-}I2_{mn}$	30.0	9.1	3.8	2.0	**6.78**
MT_{mz}	47.6	20.3	9.0	3.4	11.83
$MT\text{-}I_{mz}$	49.2	21.6	9.5	3.7	**12.64**

Table 9: Evaluation for English-Manipuri ($_{mn}$) and English-Mizo ($_{mz}$) MT systems on D_{Ev}.

	Sentence
Source	my stint as dc of tamenglong has been professionally and personally satisfying : armstrong pame.
MT_{mn}	ঐগী ওইনা stint মীৎয়েং থম্বা অমনি ওফ tamenglong অসি professionally লৈ অমসুং personally armstrong pame satisfying :
$MT\text{-}I_{mn}$	ঐগী ওইনা stint দিসি ওফ tamenglong অসি professionally লৈ অমসুং personally armstrong pame মফমনি ।
$MT\text{-}I2_{mn}$	ঐগী stint তমেংলোং ডিষ্ট্রিক্টকি ডিসি ওইনা অসি লৈ অমসুং ইশাগি ওইনা মফমনি : অরমসত্রোং পামে
Reference	ঐনা মতম খরা তমেংলোংগী ডিসি ওইবসি শিনফমগী ওইনা অমসুং ইশাগী ওইনা অপেনবা ফাওই : অরমসত্রোং পামে
MT_{mz}	my stint , dc te chuan tamenglong bana professionally leh personally satisfying : armstrong pame
$MT\text{-}I_{mz}$	my stint dc te an nei a , tamenglong professionally leh personally satisfying : armstrong pame
Reference	tamenglong dc ka nih chhung hian hnathawh dan leh mimal tak pawhin hlawkna tam tak ka hmu : armstrong pame

Table 10: Sample Output of PBSMT systems.

HBLEU scores of the retrained PBSMT systems and the original PBSMT systems (MT_{mn} and MT_{mz}). Table 8 and 9 summarize the dataset used to evaluate the MT systems and their evaluations in terms of *HBLEU* score. Table 9 shows that the retrained MT systems gives clearly better results than original MT systems with an increase in *HBLEU* score of up to 4.6. Sample outputs from the MT systems are shown in Table 10.

6 Conclusions

Using log-information gathered from our CAT tool, an analysis of the PE effort and PE time is carried out for three target languages: Manipuri, Mizo and Hindi with English as the source language. To our knowledge, this is the first PE analysis conducted on English-Manipuri, English-Mizo and English-Hindi.

Our analysis shows that current state of the art in commercially available MT for English-Hindi requires small PE effort and PE time. While MT systems for low resource languages such as Manipuri and Mizo are under development, MT training data for the languages is very scarce. Using a PBSMT system built in-house, a study on the PE effort and PE time is carried out for English-Manipuri and English-Mizo. Our findings show that, compared to English-Manipuri and English-Mizo, *PEE* is low for English-Hindi. By contrast, for English-Manipuri and English-Mizo, the problems in MT output are far more serious requiring heavy PE effort. Interestingly, while there is a significant correlation between *PEE* and

FT for the language pair English-Manipuri and English-Mizo, there is a significant negative correlation between *PEE* and *NN* for the English-Mizo language pair. The *PEE* for English-Mizo decreases with the increase in noun words in the source text, which might be because of Mizo sharing the same script as the source language. Also, a significant negative correlation is observed between *SpW* and *TT* for English-Mizo. This suggests that with the increase in the number of tokens in source text, the average time taken per word decreases for English-Mizo. We identify MT quality as well as script and ease of typing script as a factor in PE effort for languages like Manipuri and Mizo.

We also made a first attempt to address the scarcity of a parallel training data of English-Manipuri and English-Mizo MT by training the MT systems in an incremental manner using additional data created by the PE experiment. The result indicates an improvement of up to 4.6 in *HBLEU* for English-Manipuri.

7 Acknowledgments

This work is supported by Scheme for Promotion of Academic and Research Collaboration (SPARC) Project Code: P995 of No: SPARC/2018-2019/119/SL(IN) under MHRD, Govt of India.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Aljoscha Burchardt, Arle Lommel, and Maja Popovic. 2013. Tq error corpus. Technical report, Technical Report Deliverable D 1.2. 1, QT Launchpad Project.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *Proceedings of WPTP*, pages 11–20.

Hans P Krings. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019a. Extraction and identification of manipuri and mizo texts from scene and document images. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 405–414. Springer.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019b. Wat2019: English-hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

Joss Moorkens, Sharon O'brien, Igor AL Da Silva, Norma B de Lima Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29(3-4):267–284.

Jean Nitzke. 2016. Monolingual post-editing: An exploratory study on research behaviour and target text quality. *Eyetracking and applied linguistics*, 2:83–108.

Sharon O'Brien. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine translation*, 19(1):37–58.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2019. English–mizo machine translation using neural and statistical approaches. *Neural Computing and Applications*, 31(11):7615–7631.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics*, 93(1):7–16.

Maja Popovic, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th annual conference of the european association for machine translation*, pages 191–198. European Association for Machine Translation Dubrovnik, Croatia.

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2017. Comparing recurrent and convolutional architectures for english-hindi neural machine translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 167–170.

Thoudam Doren Singh. 2013. Taste of two different flavours: Which manipuri script works better for english-manipuri language pair smt systems? In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 11–18.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010a. Manipuri-english bidirectional statistical machine translation systems using morphology and dependency relations. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. Web based manipuri corpus for multiword ner and reduplicated mwes identification using svm. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, pages 35–42.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Marcos Zampieri and Mihaela Vela. 2014. Quantifying the influence of mt output in the translators' performance: A case study in technical translation. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 93–98.

Anna Zaretskaya, Mihaela Vela, Gloria Corpas Pastor, and Miriam Seghiri. 2016. Measuring post-editing time and effort for different types of machine translation errors.