Manipuri-English Machine Translation using Comparable Corpus

Lenin Laitonjam[1,2] lenin.lai@iitg.ac.in
Sanasam Ranbir Singh[1] ranbir@iitg.ac.in

[1]Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, 781039, India
[2]Department of Computer Science and Engineering, National Institute of Technology Mizoram, 796012, India

Abstract

Unsupervised Machine Translation (MT) model, which has the ability to perform MT without parallel sentences using comparable corpora, is becoming a promising approach for developing MT in low-resource languages. However, majority of the studies in unsupervised MT have considered resource-rich language pairs with similar linguistic characteristics. In this paper, we investigate the effectiveness of unsupervised MT models over a Manipuri-English comparable corpus. Manipuri is a low-resource language having different linguistic characteristics from that of English. This paper focuses on identifying challenges in building unsupervised MT models over the comparable corpus. From various experimental observations, it is evident that the development of MT over comparable corpus using unsupervised methods is feasible. Further, the paper also identifies future directions of developing effective MT for Manipuri-English language pair under unsupervised scenarios.

1 Introduction

The performances of standard data-driven MT systems rely heavily on parallel sentences. Unfortunately, parallel resources are not readily available for most low-resource languages and specialized domains, as their generation is a very costly and time-consuming task. Manipuri[1], is a language spoken in the north-eastern states of India that lacks readily available large parallel sentences. Recently developed unsupervised MT models, called Unsupervised Statistical Machine Translation (USMT) (Lample et al., 2018; Artetxe et al., 2018c) and Unsupervised Neural Machine Translation (UNMT) (Song et al., 2019; Conneau and Lample, 2019), achieved remarkable results without using any parallel sentences. The ability to learn translation features without using parallel data will boost the progress of low-resource MT studies.

Despite the reported successes, the capability of the unsupervised MT to an actual low-resource scenario is still in question. Majorities of the previous unsupervised MT-related studies (Lample et al., 2018; Artetxe et al., 2018c; Conneau and Lample, 2019;

[1]Meitei Mayek is another script used for writing Manipuri. However, in this study, we are considering Manipuri texts in Bengali.

Song et al., 2019) are for combinations of high resource languages like English, German, French, etc. for which conventional MT works well and where quality monolingual corpora are also available in abundance. Studies in (Marchisio et al., 2020; Leng et al., 2019) have also reported that USMT and UNMT performances usually vary based on the similarity/difference of the source and the target language characteristics like quantity and quality of bilingual corpus, language branch, alphabet, morphology, etc. Not only Manipuri lacks a large-quality monolingual corpus, but the language is also highly agglutinative. It belongs to the Tibeto-Burman language group (Singh and Bandyopadhyay, 2010) and has a very complex morphological structure that is very different from English (Choudhury et al., 2004). The previous study related to unsupervised Manipuri-English MT has only exploited UNMT models (Singh and Singh, 2020). However, when considering resource-scarce languages, statistical machine translation (SMT) generally outperforms neural machine translation (NMT) (Dowling et al., 2018).

Motivated by the above reason, investigating the performances of both the USMT and UNMT models on the distant language pair is meaningful and challenging. To the best of our knowledge, this study is the first attempt to investigate the performance of the USMT model on Manipuri language. Empirical evaluation of the previous models show that USMT model outperforms UNMT models for the language pair. Monoses (Artetxe et al., 2018c), a popular USMT model, achieve the best BLEU score, followed by the UNMT model proposed in (Artetxe et al., 2017). However, more advance UNMT models, MASS (Song et al., 2019) and XLM (Conneau and Lample, 2019), fails miserably. Although the preliminary experimental results are encouraging, we observe that the direct adaptation of unsupervised MT methods on the language pair is associated with many critical issues. This study also provides an in-depth analysis of the previous USMT and UNMT models and investigates their strengths and weaknesses on the language pair. Furthermore, we also propose approaches that further improve the translation performance by (1) suffix segmenting Manipuri texts to alleviate the data sparsity due to its agglutinating nature, (2) weakly supervising the cross-lingual embeddings generation using transliteration pairs, and (3) generating phrase-table using transliteration models.

The rest of the paper is organized as follows. Section 2 discusses the related works. Section 3.1 and 3.2 provides a detailed description of the USMT and UNMT models respectively. Section 4 describe the proposed approaches. Our experimental setups are presented in the section 5 followed by the results and discussion in section 6. Section 8 conclude the study.

2 Related Studies

The majority of the previous studies that try to overcome the parallel sentences dependency problem exploited the monolingual data to enhanced the MT system trained on a few hundred thousand parallel sentences (Wu and Wang, 2007; Sennrich et al., 2016; Edunov et al., 2018; Rubino et al., 2020). As a result, apart from a few (Singh and Bandyopadhyay, 2010; Sing and Bandyopadhyay, 2010; Singh, 2013; Singh and Bandyopadhyay, 2011), MT studies for low-resource Manipur-English language pair is still in their inception. There are only a few thousands publicly available Manipuri-English parallel sentences (Jha, 2012; Bansal et al., 2013; Haddow and Kirefu, 2020), which are not sufficient for statistically motivated approaches.

Unsupervised MT has recently attracted lots of attention because of its ability to learn MT features from abundantly available non-parallel corpora. Unsupervised MT is motivated by the successes of word translation models developed based on the

Unsupervised Cross-lingual Embedding (UCLE) (Conneau et al., 2017; Artetxe et al., 2018b). UCLE forms the core of the unsupervised MT frameworks and is used for initializing the MT model. In this study, we systematically investigate whether the unsupervised MT methods apply to the distant Manipuri-English pair. Unsupervised MT can be approached by following either the SMT or NMT techniques. USMT (Artetxe et al., 2018c) follows the modular design comprising several models, whereas the UNMT methods (Conneau and Lample, 2019; Song et al., 2019) focus on training an end-to-end model. Each approach has its merits and demerits. A detailed description of the models is presented in the subsequent sections.

In the case of Manipuri unsupervised MT, to the best of our knowledge, study in (Singh and Singh, 2020) is the only available literature. The authors developed a UNMT for the Manipuri-English pair based on transformer with a shared encoder and language-specific decoders. They enhance the model by using a denoising autoencoder followed by a back-translation process, similar to the settings presented in (Artetxe et al., 2017). The models are fine-tuned using a few parallel sentences as a development set. However, in our study, we do not use any parallel sentences as we want to assess the applicability of the fully unsupervised models on the language pair.

3 Unsupervised MT models

This paper considers the following state-of-the-art unsupervised MT models.

3.1 Unsupervised Statistical MT

The USMT follows the standard statistical MT (Koehn et al., 2007) formulation of a log-linear combination of several models such as translation model, re-ordering model, word or phrase penalty, language model, etc., but in an unsupervised fashion. We consider the popular, Monoses (Artetxe et al., 2018c), as our USMT model representative as the other USMT models like (Lample et al., 2018; Artetxe et al., 2019) is also based on similar concept. Monoses follows a step-by-step training procedure. Firstly, a mapping between the source and target language embeddings is obtained by aligning the monolingual phrase embeddings to a common space using the Vecmap (Artetxe et al., 2018b). Secondly, an initial phrase-table is induced by using the cosine similarity of each source embedding with the mapped target embeddings. After the initial phrase-table induction, a preliminary phrase-based SMT model (PBSMT) (Koehn et al., 2007) is built by combining the initial phrase table, distortion penalty, and language model. Next, the initial PBSMT is then tuned by utilizing synthetic parallel data obtained from a non-parallel development dataset. Finally, the fine-tuned USMT model undergoes several rounds of iterative back-translation.

3.2 Unsupervised Neural MT

UNMT generally follows three main training steps: 1) Initialization, 2) Denoising Auto-Encoder, and 3) Back translation. Initialisation step, unlike the USMT, initialised the model itself following the NMT paradigm. Denoising auto-encoder improves the UNMT performance by introducing noise during learning phase. Then, the unsupervised features are finally fine-tuned by using iterative back-translation process. Initialization generally dictates the overall performance of the UNMT systems. Subsequently, various methods for effectively initializing the model has been proposed. Earlier UNMT studies relies on UCLE (Lample et al., 2018; Artetxe et al., 2017) for initialization of the word embedding layer in the encoder and the decoder. Later, they are succeeded by cross-lingual masked language models (CMLM) (Conneau and Lample, 2019;

Song et al., 2019). The CMLM initialised the entire the encoder and decoder of the UNMT. XLM (Conneau and Lample, 2019), motivated by BERT (Devlin et al., 2018) like pre-training, initialized both for the encoder and decoder, and achieved the previous state-of-the-art results on German-English unsupervised MT. Recently, authors in (Song et al., 2019) proposed a novel unsupervised model called MASS (MAsked Sequence to Sequence pre-training) that pre-trained the both the encoder and decoder jointly, enhancing the XLM model where encoder and decoder are pre-trained separately. In this study, we consider the UCLE-based UNMT model proposed in (Artetxe et al., 2017) and the CMLM-based UNMT models (XLM and MASS). The models performance are investigated on the distant Manipuri-English language pair.

4 Proposed Approaches for Handling Low-resource Scenarios

Majority of the studies considers bilingual dictionary between the target language pairs to generate cross-lingual embeddings (Artetxe et al., 2018a). Under a low-resource scenario, we may assume unavailability of such external resources. Motivated by this, this study exploits transliteration pairs of named-entities in place of bilingual dictionaries. The transliteration of named-entities is obtained using method proposed in (Laitonjam et al., 2018).

4.1 Weakly-supervised Cross-lingual Embeddings using Transliteration Pairs

The UCLEs in the Monoses are obtained by exploiting the intra-lingual similarity distribution of individually trained source and target language embeddings (Artetxe et al., 2018b). However, we approach the problem as a weakly-supervised by using the transliteration of named-entities to obtain the initial mapping between the source and target language embeddings. More specifically, we first learn two transformation matrices using the transliteration of named-entities as a dictionary to align the source language and target language embeddings into a shared embedding space and then iteratively refining them using the self-learning method (Artetxe et al., 2018a).

4.2 Phrase-table Generation using Transliteration Models

We investigate three different methods for generating the phrase-table in Monoses. Specifically, we re-score the phrase-translation and lexical probabilities using transliteration models (TMs)[2]. TMs enable the USMT to consider phonetic similarities between the source phrase embedding ($\overline{s}$) and the mapped target phrase embedding ($\overline{t}$).

1. *Re-score Lexical Weights(RS-lex)*: In this method, we introduce transliteration weights in place of lexical weights. The transliteration weights enable the model to exploit phonetic similarities, and are estimated using the TMs, as follows:

$$tns(\overline{t}|\overline{s}) = \prod_i max(\epsilon, max_j CA(t_i, TM_{S\rightarrow T}(s_j)) \quad (1)$$

Here, $TM_{S\rightarrow T}(x)$ is the transliterated word of the source word x using the source-to-target transliteration model (TM), and $CA(x, y)$ represents the character accuracy ([0,1]) between the word x and y. ϵ is a constant fixed at 0.3 (Artetxe et al., 2018c).

2. *Re-score phrase translation probabilities (RS-phrase)*: In this case, we modify the phrase translation probabilities ϕ_{ph} itself by incorporating the transliteration weights $tns(\overline{t}|\overline{s})$ as follows:

[2]Transliteration model converts a word from a source language to a target language by keeping the source language phonetic aspects intact.

Table 1: Manipuri-English News Domain Comparable Corpora. *Vocab* stands for vocabulary and *Seg-vocab* means vocabulary size on the segmented dataset.

Language	Documents	Words	Vocab	Seg-vocab
English	13408	5.79M	80855	80855
Manipuri	13177	5.62M	277406	165998

$$\phi_{ph}(\bar{t}|\bar{s}) = \frac{exp(cos(\bar{s},\bar{t})/\tau)}{\sum_{\bar{t}'} exp(cos(\bar{s},\bar{t}')/\tau)} * tns(\bar{t}|\bar{s}) \quad (2)$$

3. *Re-score both the phrase translation probabilities and lexical weights (RS-phrase-lex)*: In this method, we use the equation 2 for estimating the ϕ_{ph} and equation 1 for estimating the lexical weights alternative, the transliteration weights.

5 Experimental Setup

5.1 Manipuri Suffix Segmenter

Manipuri is highly agglutinative. Several new words can be formed by merely attaching prefixes and suffixes to a single root, leading to data sparsity. To normalize the agglutinative nature, we use a simple yet effective Manipuri suffix segmenter based on the popular unsupervised GRAph-based Stemmer (GRAS) (Paik et al., 2011) that segments Manipuri words into roots and suffixes before training the MT models. For example, words like ইম্ফালগী (for Imphal), ইম্ফালদগী (from Imphal), ইম্ফালদা (to Imphal), etc. are normalise by separating suffixes গী , দা and দগী from the root ইম্ফাল.

5.2 Dataset Description

We use a domain-aligned[3] Manipuri-English comparable corpus generated from news articles published on two of Manipur's leading newspapers: *Sangai Express*[4] and *Poknapham*[5]. The newspaper publishes dual edition in English and Manipuri. The articles from Sangai Express are published between January 2018 to November 2018, while the articles from the Poknapham are published between March 2017 to June 2020. The lower-cased English texts are tokenized by using the Moses Tokenizer[6], while a simple whitespace tokenization scheme[7] is use for Manipuri texts. A detailed description of the training dataset is presented in table 1. All the models are evaluated on a news domain Manipuri-English MT evaluation dataset, consisting of 1006 parallel sentences. The evaluation dataset is manually created by native speakers.

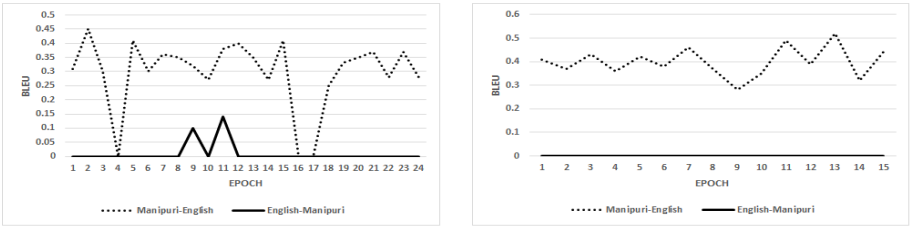
5.3 Transliteration Model Configurations

We consider the encoder-decoder based English-Manipuri transliteration model presented in the paper (Laitonjam et al., 2018) with attention mechanism (Bahdanau et al., 2015). The size of the hidden layer is fixed to 512 and embedding dimension to 256. The models are trained using the dataset presented in the study (Laitonjam et al., 2018). It consist of 4428 training transliteration pairs with 1000 development pairs.

[3]We consider the news domain.
[4]`https://www.thesangaiexpress.com/`
[5]`http://poknapham.in/`
[6]`https://github.com/moses-smt/mosesdecoder`
[7]Punctuation symbols are separated.

Table 2: Experimental results for preliminary experiments.

Methods	$En \rightarrow Mni$		$Mni \rightarrow En$	
	Non-segmented	Segmented	Non-segmented	Segmented
Conneau and Lample (2019) (XLM)	0	0.14	0	0.15
Song et al. (2019) (MASS)	0	0.18	0.44	0.23
Artetxe et al. (2017)	2.25	2.56	5.01	4.63
Artetxe et al. (2018c) (Monoses)	2.87	**3.13**	5.05	**6.37**



(a) Pre-training (b) Tuning

Figure 1: Training progress of the MASS on non-segmented dataset.

5.4 Unsupervised MT Configurations

For the UCLE-based UNMT model (Artetxe et al., 2017), we consider the original implementation[8] and default settings. We use the skip-gram model with ten negative samples to generate monolingual embeddings with size 300. Similarly, the hyperparameter of the XLM[9] and MASS[10] are set the same as in the studies (Conneau and Lample, 2019) and (Song et al., 2019) respectively. The embedding size is fixed to 1024. We jointly learn 60k sub-word units between source and target languages using BPE. However, unlike the studies (Song et al., 2019; Conneau and Lample, 2019) that uses multiple GPUs, we use only a single GPU with 12GB memory for training the model. In case of the USMT model, the Monoses[11], all model configuration settings are kept same as in the original work (Artetxe et al., 2018c).

6 Results and Discussion

Table 2 shows the translation results for our preliminary experiments. Here, *Segmented* represents the performance of the models on the segmented corpus. The segmentation is performed only on the Manipuri text using the segmenter presented in the section 5.1 to normalized the morphological infection issues of Manipuri language. Following the general practice, all the models are evaluated using BLEU scores (Papineni et al., 2002) as computed by the multi-bleu.perl[12] on the de-segmented outputs. It is evident from the experimental results that CMLM-based UNMT models (i.e., MASS and XLM) fail miserably for the language pair achieving less than 1% BLEU score on both the translation directions. Similar results were also previously reported in the study (Kim et al., 2020) for the distant English-Gujarati language pair. To further confirm CMLM-based UNMT models low performance, we evaluate the MASS at the end of each epoch during

[8] https://github.com/artetxem/undreamt
[9] https://github.com/facebookresearch/XLM
[10] https://github.com/microsoft/MASS
[11] https://github.com/artetxem/monoses
[12] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Table 3: Experimental results of proposed models on segmented dataset.

Methods	$En \rightarrow Mni$	$Mni \rightarrow En$
Monoses	3.13	6.37
Monoses + Weakly Supervised	**3.50**	6.59
Monoses + RS-lex	3.37	6.41
Monoses + RS-phrase	3.29	6.35
Monoses + RS-phrase-lex	3.47	**6.69**

Table 4: Some translation examples. The first three rows shows the reference sentences. The final three rows represent the predicted outputs of the references.

	English	***Manipuri***
Reference 1	a charge sheet has been raised	চার্জ সিট থাঙ্গৎখ্রে
Reference 2	then prime minister , dr manmohan singh personally flew down to manipur	মতমদুগী প্রাইম মিনিষ্টর দাক্টর মনমোহন সিংহ মণিপুর দা লাকখি
Reference 3	academic career of the students	মহৈরোয় শিংগী একাডমিক কেরিয়ার
	$Mni \rightarrow En$	$En \rightarrow Mni$
Predicted 1	the charge sheet filed	অমা চার্জ সিতবু থাদোকউ
Predicted 2	the then prime minister dr manmohan singh put in manipur	অমুক হন্না প্রাইম মিনিস্তর , দাঃ মনমোহন সিংহ উনরগা flew তৌরগদি মণিপুর
Predicted 3	students and their academic career	মহৈরোয় শিংনা মহৈ তম্বগী

training. Figure 1 (a) and (b) shows the progress of the model in terms of BLEU score during pre-training and fine-tuning on the non-segmented dataset. It is found that the model never gets going. Apart from the distant language pair issues, the small training corpus size may also aid to this poor BLEU score. In previous studies, CMLM-based UNMT models are generally trained on very large corpora (in term of billions of words). However, such resources are currently not available for Manipuri. On the other hand, the UCLE-based UNMT model and the Monoses performs relatively better than the XLM and MASS. Monoses obtains the best BLEU score of 3.13 for En→Mni (English-to-Manipuri) and a score of 6.37 for Mni→En (Manipuri-English) outperforming the UNMT systems. Further, on comparing the performance of each models on *segmented* and *non-segmented* corpora to investigate the effectiveness of the suffix segmenter. It is observed that the BLEU score for both the translation directions increases on the segmented dataset in almost all the cases, except for the UCLE-based UNMT model and the MASS in $Mni \rightarrow En$, as shown in the table 2. This clearly shows that the segmenting Manipuri text significantly reduces the data spareness due to morphological inflections and improves the overall performance.

Table 3 shows the results observed after incorporating the proposed approaches presented in the section 4 to enhance the USMT model. We compare its performance with the original Monoses over the segmented corpus. It is evident from the results that for all the cases, except the *Monoses with RS-phrase* for Mni→En direction, the proposed methods outperform the baseline. Weakly supervising the cross-lingual embedding generation on Monoses using the transliteration pairs obtained the best result with 3.50 BLEU for En→Mni, while Monoses with RS-phrase-lex achieved the best BLEU score of 6.64 for the Mni→En. This shows that the proposed methods are able to exploit the phonetic similarity between the language pair.

Table 5: Monoses with RS-phrase-lex N-gram precisions along with corresponding BLEU scores

	BLEU	P_1	P_2	P_3	P_4
$Mni \rightarrow En$	6.69	33.8	9.1	3.7	1.7
$En \rightarrow Mni$	3.47	23.5	4.6	1.7	0.8

6.1 Error Analysis

To gain further insights, we perform an error analysis of one of the best performing model (*Monoses with RS-phrase-lex*) on the language pair. Table 4 shows some of the translation examples of the model. It is observed that the proposed model can generate unigram translations quite accurately. For instance, unigram translation pair (students, মহৈরোয়), as shown in table 4 (*Reference 3*), is correctly predicted for both the translation directions, as shown in *predicted 3* of table 4. Similarly, multi-word pairs like (prime minister, প্রাইম মিনিস্তর) are also correctly predicted. However, in most of the cases, the models fail to handle higher multi-gram translations, thereby leading to overall low BLEU score. The difference in BLEU score and the corresponding modified n-gram precisions P_n (n = *1,2,3,4*) for the model can also be seen in the table 5. The n-gram precision scores significantly decreases with increase in n. For instance, the uni-gram precision for $Mni \rightarrow En$ MT is 33.8%. However, the corresponding BLEU score is 6.69% only. We believe that difference in word order between the language pair is a major contributing factor to such a massive difference between the BLEU and n-gram precisions. English follows a Subject-Verb-Object (SVO) order in contrast to the Manipuri SOV order. As a result, the unsupervised model fails to handle the word order differences. For instance, in the $Mni \rightarrow En$ translation example, the order of the words *students* and *academic career*, shown in the reference 3 of table 4, gets interchange and is wrongly predicted as shown in the corresponding translation (predicted 3).

7 Future Research Directions

It is observed from the above observation that there is a potential for developing MT system for Manipuri-English language pair using comparable corpora, and may be a way forward to counter the challenges of creating sentence level parallel corpora. However, for developing such a system, we would need effective multi-lingual embedding techniques to develop effective bilingual dictionary, phrase-table, language modelling for post processing sentence correction etc. Further, we would also need to take care the dynamic writing styles followed in Manipuri. For instance, (জনুৱারী, জনুৱারি, জানুৱারী and জানুৱারি) are acceptable writing forms of the word *January*. Such a variation is inevitable for comparable corpora while the text are pooled specially from different sources.

In addition, from the P_1 performance in Table 5, it also evident that the translation performance can be further enhanced using post processing correction using methods like language modelling, NMT hybridization on the USMT model (Artetxe et al., 2019; Marie and Fujita, 2020), etc.

8 Conclusion

We develop a MT system for low-resource distant Manipuri-English language pair without using parallel sentences. Our study reveals that a relatively cheaper domain-aligned comparable corpora befit potential replacement of expensive parallel sentences for the language pair MT task. We also compare a popular USMT model with state-of-the-art UNMT models and found that the modular design of the USMT model is better suited

for the language pair. Furthermore, this paper empirically shows that using a Manipuri suffix segmenter reduces the data sparseness issue due to the Manipuri text's agglutinative nature. Also, we found that weakly-supervising the USMT model using the transliteration pairs and transliteration models improves the translation performance. Though not with high performance, this work provides a stable MT baseline for the low-resource Manipuri-English language pair. We also offer several directions for future studies to encourage more research on this crucial problem.

References

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Artetxe, M., Labaka, G., and Agirre, E. (2018c). Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Bansal, A., Banerjee, E., and Jha, G. N. (2013). Corpora creation for indian language technologies–the ilci project. In *the sixth Proceedings of Language Technology Conference (LTC '13)*.

Choudhury, S. I., Singh, L. S., Borgohain, S., and Das, P. K. (2004). Morphological analyzer for manipuri: Design and implementation. In *Asian Applied Computing Conference*, pages 123–129. Springer.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dowling, M., Lynn, T., Poncelas, A., and Way, A. (2018). Smt versus nmt: Preliminary comparisons for irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Haddow, B. and Kirefu, F. (2020). Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Jha, G. N. (2012). The tdil program and the indian language corpora initiative. In *Language Resources and Evaluation Conference*.

Kim, Y., Graça, M., and Ney, H. (2020). When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Laitonjam, L., Singh, L. G., and Singh, S. R. (2018). Transliteration of english loanwords and named-entities to manipuri: Phoneme vs grapheme representation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 255–260. IEEE.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Leng, Y., Tan, X., Qin, T., Li, X.-Y., and Liu, T.-Y. (2019). Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Marie, B. and Fujita, A. (2020). Iterative training of unsupervised neural and statistical machine translation systems. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–21.

Paik, J. H., Mitra, M., Parui, S. K., and Järvelin, K. (2011). Gras: An effective and efficient stemming algorithm for information retrieval. *ACM Transactions on Information Systems (TOIS)*, 29(4):1–24.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rubino, R., Marie, B., Dabre, R., Fujita, A., Utiyama, M., and Sumita, E. (2020). Extremely low-resource neural machine translation for asian languages. *Machine Translation*, 34(4):347–382.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Sing, T. D. and Bandyopadhyay, S. (2010). Statistical machine translation of english–manipuri using morpho-syntactic and semantic information. *Proceedings of the Association for Machine Translation in the Americas (AMTA 2010)*.

Singh, S. M. and Singh, T. D. (2020). Unsupervised neural machine translation for english and manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78.

Singh, T. D. (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.

Singh, T. D. and Bandyopadhyay, S. (2010). 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.

Singh, T. D. and Bandyopadhyay, S. (2011). Integration of reduplicated multiword expressions and named entities in a phrase based statistical machine translation system. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1304–1312.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.