Unsupervised Neural Machine Translation for English and Manipuri

Salam Michael Singh and Thoudam Doren Singh

Department of Computer Science & Engg. National Institute of Technology, Silchar 788010, India {salammichaelcse, thoudam.doren}@gmail.com

Abstract

Availability of bitext dataset has been a key challenge in the conventional machine translation system which requires surplus amount of parallel data. In this work, we devise an unsupervised neural machine translation (UNMT) system consisting of a transformer based shared encoder and language specific decoders using denoising autoencoder and back-translation with an additional Manipuri side multiple test reference. We report our work on low resource setting for English (*en*) - Manipuri (*mni*) language pair and attain a BLEU score of 3.1 for $en \rightarrow mni$ and 2.7 for $mni \rightarrow en$ respectively. Subjective evaluation on translated output gives encouraging findings.

1 Introduction

Machine Translation had been dominated by the statistical methods, notably the phrase-based (Koehn et al., 2003; Och, 2003). But they suffered from a rigid structure since they have multiple modules (Brown et al., 1990, 1993) which are tuned independently. In other words, SMT lacked an end-to-end learning mechanism. For a quite long period, SMT dominated the MT systems, but with an RNN based sequence-to-sequence model Sutskever et al. (2014); Cho et al. (2014) marked the beginning of the NMT era. But, these primitive neural based MT systems choked when the input sentences get longer as the input sentence are squeezed into a fixed length vector. Fortunately, with the advent of attention (Bahdanau et al., 2014; Luong et al., 2015), sub-word tokenization (Sennrich et al., 2016b) and transformers (Vaswani et al., 2017), NMT outperformed SMT in various machine translation tasks. However, when the parallel corpus is scarce, the NMT fails to produce good translations, and performing much poorer than the phrase-based systems. Building parallel corpus is a costly task and specifically for

the low resource languages where bi-text is nonexistent. But, monolingual data is easily available even for the low resource languages and some have utilised it to augment parallel data (Sennrich et al., 2016a) with a little bi-text data or translation systems using monolingual data only (Lample et al., 2018a; Artetxe et al., 2018b; Ren et al., 2019). Recent works using monolingual data only show a positive direction in machine translation tasks. Although, these systems do not outperform a strong supervised system, they could be treated as a strong baseline system which should be the lower bound for any supervised system.

1.1 Motivation and Challenges

The machine translation (MT) systems have become very effective in recent times, however with the condition of huge parallel data availability. On the other hand, the MT task for many low-resource language is yet to be addressed. Likewise, Manipuri is a low-resource Indian language belonging to Tibeto-Burman language family where readily available English-Manipuri parallel data is close to non-existent. As the manual parallel data acquisition is both challenging and resource intensive task, while the monolingual data is comparatively easier to acquire and it is hence intuitive to look upon the techniques which exploits the monolingual data to the fullest.

Another challenge is that Manipuri language is highly agglutinitive and morphologically rich which causes linguistic diversity and variations in word forms thus penalizes the automatic n-gram matching metrics like BLEU. Lack of grammatical gender, agglutinative verb morphology, extensive suffix with more limited prefixation and Subject Object Verb (SOV) order are some of the linguistic features. Manipuri language uses Bengali¹ scripts

¹http://unicode.org/charts/PDF/U0980.pdf

and Meetei mayek² in written form. In this work, we will focus on the Bengali script.

1.2 Contributions

In order to tackle the above challenges i.e lack of parallel data and linguistic diversities, we make the following contributions: (1) We report a preliminary unsupervised MT task for English-Manipuri language pair using monolingual data only to tackle the parallel data scarcity and explore the effectiveness/non-effectiveness for the same. (2) We develop multiple references for the Manipuri side test data, specifically we built an additional test reference apart from the one extracted from the training corpus to tackle the linguistic diversity as it increases the n-gram overlapping probability. (3) We find that a cross-lingual mapping of embeddings performs better than a pretrained cross-lingual language model as an initialization step for this distant language pair (English-Manipuri).

The remaining of this paper is organized as follows. Section 2 explores the related work. Section 3 then describes the framework of our approach. The experimental settings are discussed in Section 4, while Section 5 discusses the obtained results and its analysis. Section 6 concludes the paper.

2 Related Works

To the best of our knowledge, we are the first to report the unsupervised neural machine translation for Manipuri language. However, there are reports of supervised Statistical based approach, notably by (Singh and Bandyopadhyay, 2010a; Singh, 2013), where the authors made an imperative study over the effects of the morpho-syntactic information and dependency relation in a Statistical Machine Translation setting. In another work, Singh and Bandyopadhyay (2011) showed that the Phrase Based Statistical Machine Translation system improves by incorporating linguistic features such as, named entities and reduplicated multiword expressions. However, the MT task for Manipuri is still in its inception, considering it being a low resource language. Low resource has always been a hurdle in the MT task. Many, have mended their hands to overcome this bottleneck. Sennrich et al. (2016a) leveraged the NMT system by using monolingual data to create a synthetic parallel corpus using backtranslation. This synthetic data greatly improved the MT systems, however it consisted mostly noises which is necessary to be postprocessed, making the task not so feasible. Zoph et al. (2016) devised a transfer learning mechanism by sharing model parameters from a high resource language pair (parent model) to lower resource language pair (child model) which significantly improved the BLEU (Papineni et al., 2002) score. This kind of parent-child model has been used in other works (Nguyen and Chiang, 2017; Kocmi and Bojar, 2018) where they used a shared vocabulary of subword units. Kocmi and Bojar (2018) further showed that transfer learning can be simplified where the parent model is trained until convergence and switching the low resource language pair as the training data while keeping the training parameters unchanged. Parameter sharing has also been explored in previous works such as Firat et al. (2016) by using a single shared attention mechanism with multiple encoder-decoder to devise a multi-way multilingual NMT system. Further, Johnson et al. (2017) devised another multilingual approach through parameter sharing using a single shared encoder-decoder model enabling zero-shot translation, where the model learned translation between multiple languages and could even translate unseen language pairs. However, these models require some form of parallel data which has led the dynamics to shift towards exploiting the monolingual data, which is comparatively abundant than the bi-text data. Furthermore, a noteworthy attempt is reported by He et al. (2016), where they used an auto-encoding task to ensure the translated sentence can be translated back to the original sentence using reinforcement learning technique. Similarly, Cheng et al. (2016) also used this auto-encoding task upon the monolingual data. Although, these models seemed promising, it still needed decent amount of parallel data for a warm-start. Machine translation task can be reduced to a deciphering task (Ravi and Knight, 2011; Pourdamghani and Knight, 2017) from a monolingual data using noisy channel model where the source language is treated as ciphertext generation, however these settings were mostly confined to short sentences and related language pairs.

All these works induced a promising starting point towards exploiting the monolingual data, but these primitive models were unable to stand

²http://unicode.org/charts/PDF/UABC0. pdf

against a supervised setting with abundant parallel data. Fortunately, it was the concurrent work of Lample et al. (2018a) and Artetxe et al. (2018b) which lifted the unsupervised MT on par with a supervised setting. Their approach first learns a linear transformation of the word embeddings of the two languages in an unsupervised manner which are trained independently and map this linear transformation into a shared space using adversarial training (Conneau et al., 2017) or through self learning (Artetxe et al., 2017, 2018a). A shared encoder for both the languages is initialized using the resulting cross-lingual embeddings. The model is trained using denoising auto-encoding, backtranslation and adversarial training³ iteratively giving rise to translation models of increasing quality. There are reports of unsupervised MT using PB-SMT(Phrase-based Statistical Machine Translation) where Lample et al. (2018b) used the backtranslated synthetic data to feed into the NMT system. Furthermore, there are reports of using the synthetic data generated from an unsupervised SMT to initialize an NMT system from scratch in an iterative way (Marie and Fujita, 2018; Ren et al., 2019). Recently, a pretrained cross-lingual language model (Lample and Conneau, 2019; Song et al., 2019) is used to pretrain a shared encoder and finetune using iterative backtranslation.

3 The Framework

In this section, we discuss the framework of our unsupervised neural machine translation system.

Denoising Autoecoder: The goal of the unsupervised MT setup is to be able to reconstruct an original sentence using a decoder from an input sentence which is encoded using a shared encoder. The reconstruction is carried out such that the encoder should learn the latent representations (embeddings of both the languages) in a language independent manner. Meanwhile, the decoder should be able to transfer this latent representations into their corresponding languages. Since, the systems lacks any constraint, it fails to perform a meaningful reconstruction mechanism, as the system sticks in a trivial copying task. This blindly copying results into failure of capturing any real and useful structure from the data. Denoising autoencoder (Vincent et al., 2008) is used in addition with noise (random word swaps, ran-



Figure 1: System Overview: The system first maps the monolingual vector embeddings into a common crosslingual embedding space. The shared encoder and the two decoders are transformer units. For each iterations, we first denoise the two languages in batches and back-translate from L1 to L2 and L2 to L1. L1 is *en* when L2 is *mni* and vice-versa.

dom word drops) to address this issue by constraining the system to learn latent representations.

Backtranslation: The denoising objective covers only a single language at one time which partially fulfills the final translation task. To achieve a full translation system, we need some sort of synthetic parallel data which is obtained through backtranslation (Sennrich et al., 2016a) since our setting is constraint to monolingual data alone. The system translates from one language to another in the inference mode using greedy decoding and then optimises the discrepancy between the actual and the synthetic translation. The backtranslation is carried out batch by batch for each models in an iterative manner producing an improved synthetic parallel data after each iteration.

Unsupervised Neural Machine Translation: The system first optimises the encoding loss objective of a noisy version of the source language (L1) through denoising autoencoder using a shared transformer based encoder and reconstruction using the (L1) decoder. Subsequently, this reconstructed version of L1 is backtranslated to the target language (L2) using the L2 decoder to create a synthetic parallel data and iteratively optimizes the objective to predict the original sentence from this synthetic data. This iterative process is executed alternatively for L1 and L2. A brief overview of our framework is shown in Figure 1.

4 Experimentation

We perform the translation task for both the $en \leftrightarrow mni$ sides. We use the same data for both

³Adversarial training is adopted by (Lample et al., 2018a,b)

tasks and compare our model with XLM (Lample and Conneau, 2019) and Artetxe et al. (2018b). First, we discuss the dataset description and its preprocessing followed by the baseline and the proposed framework.

4.1 Dataset Description and Preprocessing

Our unsupervised setting exploits the comparable English and Manipuri monolingual data accompanied with parallel data for development and evaluation purpose. Both monolingual data (Singh and Bandyopadhyay, 2010b) are crawled⁴ and comparable in news domain. For the validation and testing, we used the Technology Development for Indian Languages (TDIL)⁵ dataset. We took 1000 sentences for the development set and 500 sentences for the test set which fall under different domain.

4.1.1 Data Preprocessing

The preprocessing consists of normalization, sentence splitting and tokenization. The mosesdecoder⁶ toolkit scripts are used for the English side while IndicNLP⁷ library is used for the Manipuri counterpart. Further, the English monolingual data consisted multiple instances of hyphen separated words instead of a single continuous word as illustrated in Table 1 since the corpus was scraped from columnar news format where words are broken into hyphen separated subwords in order to maintain page layout which we removed. Finally,

HypSep Word	HypRem Word
Mani-pur	Manipur
Samaj-wadi	Samajwadi
irrespon-sible	irrespossible

Table 1: Instances of Hyphen separated (**HypSep**) words and their respective hyphen removed (**HypRem**) words.

a shared vocabulary was learned using fastBPE⁸ from the monolingual data using 60,000 operations. The statistics of both the monolingual corpus and the parallel corpus after the preprocessing is given in Table 2 respectively.

Corpus		Sentences	Tokens	
Mono	en	378,693	10,172,299	
	mni	132,071	3,509,945	
Dev	en	1,000	29,801	
	mni	1,000	31,109	
Test	en	500	13,131	
	mni	500	13,575	

Table 2: Statistics of the Monolingual and the Develop-
ment Corpora for the English (en) and Manipuri (mni).

4.1.2 Manipuri Side Multiple Reference

In addition to the original parallel data for the Manipuri side test reference, we also include another test reference to handle linguistic diversity. Any manual translation is bound to have variations with the translation of other translator at lexical, semantic and syntactic level. And, considering the fact that most machine translation systems are evaluated via one of the string matching methods which penalizes the paraphrasing. Hence to maximize the string overlapping, we include an additional Manipuri side test reference.

Source:	I am crazy about Thai, Mughlai
	and Bengali food.
Reference-1:	ঐনা থাই মুঘলাই অমসুং বেঙ্গলি চিঞ্জাক য়ামা পামজৈ।
Transliteration	ai-na thai mughlai ama-sung:
	bengali chinjak yaamnaa pam-
	jei.
English-	I extremely love Thai, Mughlai
Translation:	and Bengali food.
Reference-2:	ঐ থাই মুঘলাই অমদি বেঙ্গলি চিঞ্জাকশিংদা য়ামা ঙাওজৈ।
Transliteration	ai thai mughlai amadi bengali:
	chinjak-shing-da yaamnaa ngou-
	jei.
English-	I am very crazy about Thai,
Translation:	Mughlai and Bengali food.

⁴https://www.thesangaiexpress.com/ ⁵http://tdil.meity.gov.in/ ⁶https://github.com/moses-smt/ mosesdecoder ⁷https://github.com/anoopkunchukuttan/ indic_nlp_library ⁸https://github.com/glample/fastBPE

4.2 Baseline Systems

We compare our proposed model with the XLM (Lample and Conneau, 2019) and Artetxe et al. (2018b). XLM uses a transformer based shared encoder and a shared decoder with a cross-lingual language model pretraining as the initialization step. While, Artetxe et al. (2018b) uses a GRU based shared encoder and language specific decoder with cross-lingual word embedding mapping as the initial step.

4.2.1 Baseline-1

We use the Lample and Conneau (2019) system as our first baseline (UNMT_{baseline-1}). For both the cross-lingual language model pretraining and the translation finetuning, the XLM⁹ toolkit is used. A Transformer based architecture is used (Vaswani et al., 2017) for the encoder and decoder with 6 layers, 8 multi-head attention units, 1024 hidden units, GELU activation with a dropout (Srivastava et al., 2014) rate of 0.1 accompanied with a positional encoding. In addition to this, Adam optimizer (Kingma and Ba, 2014) for the optimisation, a linear warmup (Vaswani et al., 2017) and varying learning rates from 10^{-4} to 5.10^{-4} is used. A stream of 256 tokens is used with a batch size of 32 instead of 64 in (Lample and Conneau, 2019) for the Masked Language Model (MLM) objective. Averaged perplexity over the two languages is used as the stopping criterion for the MLM pretraining. For the unsupervised MT task, all the hyper-parameters are same as that of the MLM with the addition of noise (word shuffle, word dropout, word blank) and 2000 tokens per batch. The stopping criterion is the average of the tokenized BLEU score¹⁰ considering the unsupervised criterion for the two directions.

4.2.2 Baseline-2

The second baseline (UNMT_{baseline-2}) is based on the work done by Artetxe et al. (2018b). For the sake of comparison, we use the same settings as described in the paper of Artetxe et al. (2018b) and the same cross-lingual embedding mapped vectors from our proposed model (UNMT*proposed*) in Section 4.3 for the intialization step. Furthermore, the system is a GRU based bi-directional encoderdecoder network with global attention (Luong et al., 2015) with 600 hidden units for each GRU cells and the embeddings of 300 dimensions. For optimization, Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of 0.0002. Additionally, this setting incorporates a single shared encoder and two language specific decoders and we do not use any additional parallel data for parameter tuning purpose. The model is trained for 290,000 iterations with a batch size of 50 and implemented using the codebase¹¹ of Artetxe et al. (2018b).

4.3 Proposed Model Setup

Our proposed model (UNMT_{proposed}) operates in twofold. First, it learns a cross-lingual word embedding mapping and then followed by an iterative backtranslation.

Cross-lingual mapping: First, we use the monolingual corpora to independently train the embeddings for *en* and *mni* separately using fast-Text (Bojanowski et al., 2017) with the skip-gram model having 10 negative samples, a context window of 10 words, 300 dimensional embedding vector, a sub-sampling of 10^{-5} and 10 training iterations. After getting monolingual embedding for each languages, we map the fastText embeddings into a common embedding space using vecmap¹² (Artetxe et al., 2018a) without using any parallel data.

MT Task: After getting the cross-lingual embeddings mapping, we perform the denoising and iterative backtranslation task. For this, we use a transformer based shared encoder and language specific decoders. The transformer(Vaswani et al., 2017) model has 5 encoder-decoder layers, 6 attention heads, 300 hidden units. We use GELU activation with a dropout (Srivastava et al., 2014) rate of 0.1 and optimized using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001. The training is conducted for 290,000 iterations. We implement our model using PyTorch¹³ by extending the implementation of Artetxe et al. (2018b).

4.4 Evaluation Metrics Used

We measured the performance of our systems we used both sentence level and character level similarity between the reference and the hypothesis.

⁹https://github.com/facebookresearch/ XLM

¹⁰https://github.com/moses-smt/ mosesdecoder/blob/master/scripts/ generic/multi-bleu.perl

¹¹https://github.com/artetxem/undreamt

¹²https://github.com/artetxem/vecmap

¹³https://pytorch.org/



Figure 2: BLEU vs Iterations graph: The left side graph shows the BLEU vs Iterations for the $en \rightarrow mni$ direction considering the reference "ref1" only. Whereas, the right side is for the $mni \rightarrow en$ direction. The blue line represents the UNMT_{baseline-2} model while the orange line is for the UNMT_{proposed} model.

Systems	Directions	BLEU			ChrF	
		ref1	ref2	ref1+ref2	ref1	ref2
UNMT _{baseline-1}	en-mni	pprox 0	pprox 0	pprox 0	13.5531	13.0772
	mni-en	pprox 0	pprox 0	pprox 0	14.0372	-
UNMT _{baseline-2}	en-mni	1.8	1.4	2.7	19.51	19.30
	mni-en	2.4	-	-	22.27	-
UNMT _{proposed}	en-mni	2.0	1.7	3.1	21.21	20.73
	mni-en	2.7	-	-	24.89	-

Table 3: BLEU score and the character n-gram F-score (ChrF) of the systems for en-mni and mni-en translation tasks using the multiple references (ref1 and ref2) for the Manipuri side.

For the sentence level we used **BLEU** (Papineni et al., 2002) and **ChrF** (Popović, 2015) for the character n-gram F-score.

5 Results and Analysis

In this section, we discuss the results and the performance of our proposed setting (UNMT_{proposed}) in comparison with the baselines (UNMT_{baseline-1}) and (UNMT_{baseline-2}). The reported BLEU score is calculated upon the de-tokenized text using sacrebleu¹⁴ (Post, 2018) while ChrF is calculated using ChrF toolkit¹⁵ for the $en \leftrightarrow mni$ directions. The scores of the systems is given in Table 3 with the inclusion of additional Manipuri side multiple test references (ref1 and ref2). We find that the two UNMT systems with cross-lingual embeddings (UNMT_{baseline-2} and UNMT_{proposed}) as the initialization step yields better BLEU and ChrF scores than the one with the cross-lingual language model pretraining (UNMT_{baseline-1}) as the precursor to UNMT for this language pair. Furthermore, a shared decoder setting of the UNMT_{baseline-1} fails to mitigate a proper parameter sharing between this distant language pairs. On the contrary, the other two UNMT settings with distinct language decoders performs better.

Additionally, our transformer based premise (UNMT_{proposed}) significantly improves over the GRU based stronger baseline (UNMT_{baseline-2}) both in terms of BLEU and ChrF scores for the $en \rightarrow mni$ and $mni \rightarrow en$ directions respectively. The UNMT_{baseline-1} produced a BLEU score of \approx 0 for almost every checkpoints so we present in Figure 2 the comparative score for the other two unsupervised settings over 290,000 iterations.

5.1 Effect of the Multiple Reference

The machine translation system seldom generates a perfect paraphrase due to the possible valid linguistic variations. However, the n-gram based evaluation metrics like BLEU requires exact word overlaps thus penalizing a valid diverse word form. To cope up with this issue we use an additional Manipuri side multiple test references (ref1 and ref2). In Table 3, we find that both UNMT_{baseline-2} and

¹⁴https://github.com/mjpost/sacrebleu

¹⁵https://github.com/m-popovic/chrF

Input English-1	Frequent malarial infection can cause anaemia	
Reference	মেলেরিয়াগী ইনফেক্সন তোইবনা এনেমিয়া থোকহল্লি	
UNMT _{baseline-2}	তোইনা মেলেরিয়াগী লায়না থীংবা	
Transliteration	Toina malariagi laina thingba	
en Translation	Frequently prevent the malarial disease	
UNMT _{proposed}	মেলেরিয়া ইনফেক্সন তোইনা থোকহনবা য়াই	
Transliteration	Malaria infection toina thokhanba yai	
en Translation	Malaria infection can be caused frequently	
Input	Containers for carrying and storing water need to be kept clean inside	
English-2	and outside and covered to keep the water clean	
Reference	ঈশীং শোক্লবা অমসুং থম্নবা পাত্রশীং অসিগী মনুং-মপান অনিমক শেংহনগদবনি অমসুং ঈশীং শেংনা থম্নবা মখুম	
	কুপকদবনি	
UNMT _{baseline-2}	পর্ট্রেজ অমসুং হার পুথোক পুশিন তৌনবা টোইলেট অমসুং তরু - তরুবা ঈশিং থম্নবা হোত্বগদবনি।	
Transliteration	Pertraize amashung haar puthok pushin tounaba toilet amashung taru-	
	taruba eeshing thamnaba hotnagadabni.	
en Translation	Toilet and storage of clean water should be practiced for pertraize and	
	fertilizers transportation.	
UNMT _{proposed}	ঙা ফানবা অমসুং থক্লবা ঈশীং তরু - তনানবা অমসুং মখুম শেংনা ঈশীং থমগদবনি।	
Transliteration	Nga fanaba amashung thaknaba eeshing taru-tananba amashung	
	makhum shengna eeshing thamgadabni.	
en Translation	Water should be stored with a clean cover for fishing and drinking.	

Table 4: Sample Input and Output for the two test sentences for the reference "ref1".

UNMT_{proposed} improves significantly in terms of BLEU score after using the multiple references for en-mni direction. The UNMT_{baseline-2} improves by +0.9 and +1.3 cumulative BLEU score (ref1+ref2) over the separate scores for ref1 and ref2 respectively. While, the cumulative BLEU score of UNMT_{proposed} improves by +1.1 and +1.4 over the ref1 and ref2 respectively. And finally our approach, UNMT_{proposed} significantly improves by a BLEU score of +0.4 over UNMT_{baseline-2} with the inclusion of multiple reference.

5.2 Sample Input-Output and Qualitative Analysis

We present a critical case study of the sample input and output of two randomly selected *en* test sentences w.r.t the reference "ref1" and their respective *mni* translations by our UNMT_{baseline-2} and UNMT_{proposed} systems. We do not report the UNMT_{baseline-1} since it struggled to obtain even a proper word form. Although, we perform a bidirectional translation task, the analysis is focused on the *en* \rightarrow *mni* part. The *en* test sentences are selected randomly, one being short and the other a longer one in order to see the effect of the length on translation quality. Furthermore, we select the translated outputs of UNMT_{baseline-2} and UNMT_{proposed} with their respective best BLEU scoring checkpoint. Additionally, in Table 4 we provide the en transliteration of the mni sentences and their en translations. We assess the quality of our translations under the criterion such as subjectobject-verb¹⁶ agreement, adequacy and fluency. Considering the first test sentence (Input English-1), it is observed that both the systems captured the similar word form of the subject মেলিরয়াগী ইনেফক্সন (malarial infection) of the reference while dropping the object এনেমিয়া (anaemia). Although, both the systems do not converge to the actual meaning of the reference sentence, they follow a perfect SOV word agreements thus making fluent translations, albeit with a relatively poor adequacy. Meanwhile, the systems performed comparatively good for a shorter length test sentence, but the systems struggles when the length increases as it is evident in the translations of the Input English-2 sentence. The two systems generated translations with a proper sentential form. Although, both the

¹⁶Manipuri language follows subject-object-verb (SOV) word order.

systems covered the key words such as ঈশীং (eeshing : water), থমগদবনি (thamgadabni : should be stored), তরু - তনানবা (taru-taananba : clean), তরু - তরুবা (taru-taruba : clean) and the conjunction অমসুং (amashung : and), but failed to generate a corresponding word for পাত্রশীং (paatrashing : containers). Rather the systems translated extraneous yet related words like ঙা ফানবা (ngaa faanaba : fishing) and ঈশীং থরুবা (eeshing thaknaba : drinking water). The UNMT_{baseline-2} generates an absurd word পর্দ্রেজ (Pertraize) which is highlighted with an underline. This word is highly likely to have been generated by the BPE subwords.

While we demonstrate here the translation of two samples, but for the test set as a whole we observe that the performance of these two systems are synonymous for a shorter length test sentence. Additionally, we observe that when the sentence length increases our proposed system with transformer produces a more fluent translation and a relatively adequate one (by generating the related words) than the GRU based baseline (UNMT_{baseline-2}) as per human evaluation.

6 Conclusion

In this work, we report an unsupervised neural machine translation system on low resource setting for English (en) - Manipuri (mni) language pair using a transformer based shared encoder, language specific decoders and monolingual data only. We observe an improvement in BLEU score (+0.4 for $en \rightarrow mni$ and +0.3 for $mni \rightarrow en$) over the stronger baseline (UNMT_{baseline-2}) by incorporating a transformer based shared encoder and independent decoders along with a multiple reference scenario. Similarly, the ChrF scores of UNMT_{proposed} surpassed both the baselines by a large margin. Moreover, it is found that the cross lingual embedding mappings is more effective than a cross lingual language model pretraining for this language pair. One of the reason being that the XLM model objective pretrains the encoder and the decoders separately and the decoders are shared between the two languages. This shared decoder may be useful for a similar language pair but for the distant pairs it fails to capture the cross-lingual representations. Besides, the automatic scoring mechanism BLEU fails to mitigate and capture the linguistic inflections of the morphologically richer Manipuri language which is tackled by using multiple references. Meanwhile, the translation quality is reasonably fluent and adequate considering the facts that the language pairs are relatively unrelated and the use of a single test reference.

Finally, our unsupervised premise made a descent performance considering the relatively smaller monolingual data used hence this work highlights the potential of the unsupervised MT for this distant language pairs. In our future work, we plan to extend this preliminary unsupervised setting by incorporating linguistic features and devise an improved initialization step along with a better use of the synthetic data suitable for this language pair.

References

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798, Melbourne, Australia. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semisupervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. pages 1724– 1734.
- Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. arXiv preprint arXiv:1710.04087.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 866–875, San Diego, California. Association for Computational Linguistics.
- Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.
- Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 127–133.
- Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS).
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only.

In International Conference on Learning Representations (ICLR).

- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the* 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.
- Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.
- 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186– 191, Belgium, Brussels. Association for Computational Linguistics.
- Nima Pourdamghani and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the* 2017 Conference on Empirical Methods in Natural Language Processing, pages 2513–2518, Copenhagen, Denmark. Association for Computational Linguistics.

- Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 12– 21, Portland, Oregon, USA. Association for Computational Linguistics.
- Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 241–248.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715– 1725, Berlin, Germany. Association for Computational Linguistics.
- 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, Atlanta, Georgia. Association for Computational Linguistics.
- 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, Beijing, China. Coling 2010 Organizing Committee.
- 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, Beijing, China. Coling 2010 Organizing Committee.
- Thoudam Doren Singh and Sivaji Bandyopadhyay. 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, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *In*-

ternational Conference on Machine Learning, pages 5926–5936.

- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27, pages 3104–3112. Curran Associates, Inc.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the* 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.