Neural Machine Translation between Myanmar (Burmese) and Dawei (Tavoyan)

Thazin Myint Oo
University of Computer Studies, Yangon
thazinmyintoo@ucsy.edu.mm

Ye Kyaw Thu
National Electronics and Computer Technology Center, Thailand
yktnlp@gmail.com

Khin Mar Soe
University of Computer Studies, Yangon
khinmarsoe@ucsy.edu.mm

Thepchai Supnithi
National Electronics and Computer Technology Center, Thailand
thepchai.supnithi@nectec.or.th

Abstract

This work explores the first evaluation of the quality of neural machine translation between Myanmar (Burmese) and Dawei (Tavoyan). We also developed Myanmar-Dawei parallel corpus (around 9K sentences) based on the Myanmar language of ASEAN MT corpus. We implemented two prominent neural machine translation systems: Recurrent Neural Network (RNN) and Transformer with syllable segmentation. We also investigated various hyper-parameters such as batch size, learning rate and cell types (GRU and LSTM). We proved that LSTM cell type with RNN architecture is the best for Dawei-Myanmar and Myanmar-Dawei neural machine translation. Myanmar to Dawei NMT achieved comparable results with PBSMT and HPBSMT. Moreover, Dawei to Myanmar RNN machine translation performance achieved higher BLEU scores than PBSMT (+1.06 BLEU) and HPBSMT (+1.37 BLEU) even with the limited parallel corpus.

Keywords**:** *neural machine translation, Myanmar-Dawei parallel corpus, Recurrent Neural Network and Transformer*

I. INTRODUCTION

The Myanmar language includes a number of mutually intelligible Myanmar dialects, with a largely uniform standard dialect used by most Myanmar standard speakers. Speakers of the standard Myanmar may find the dialects hard to follow. The alternative phonology, morphology, and regional vocabulary cause some problems in communication. Machine translation (MT) has so far neglected the importance of properly handling the spelling, lexical, and grammar divergences among language varieties. In the Republic of the Union of Myanmar, there are many ethnical groups, and dialectal varieties exist within the standard Myanmar language. To address this problem, we are developing a Myanmar (Burmese) and Dawei (Tavoyan) parallel text. We conducted statistical machine translation (SMT) experiments between Myanmar and Dawei and obtained the highest BLEU scores 45.58 BLEU score for Myanmar to Dawei and 63.22 BLEU score for Dawei to Myanmar [25]. Deep learning revolution brings rapid and dramatic change to the field of machine translation. The main reason for moving from SMT to neural machine translation (NMT) is that it achieved the fluency of translation that was a huge step forward compared with the previous models. In a trend that carries over from SMT, the strongest NMT systems benefit from subtle architecture modifications and hyperparameter tuning.

NMT models have advanced the state of the art by building a single neural network that can learn representations better [4]. Other authors conducted experiments with different NMTs for less-resourced and morphologically rich languages, such as Estonian and Russian [5]. They compared the multi-way model performance to one-way model performance, by using different NMT architectures that allow achieving state-of-the-art translation. For the multiway model trained using the transformer network architecture, the reported improvement over the baseline methods was +3.27 BLEU points. Honnet et al., 2017 [6] proposed solutions for the machine translation of a family of dialects, Swiss German, for which parallel corpora are scarcee. The authors presented three strategies for normalizing Swiss German input to address the regional and spelling diversity. The results show that character-based neural machine translation was the most

promising strategy for text normalization and that in combination with phrase-based statistical machine translation it achieved 36% BLEU score.

In their study, NMT outperformed SMT. In our study, we performed the first comparative NMT analysis of Myanmar dialectal language with two prominent architectures: recurrent neural network (RNN) and transformer. We investigated the translation quality of the corresponding hyper-parameters (batch size, learning rate and cell type) in machine translation between the standard Myanmar and Dawei (one of the dialects) languages. We used syllable byte pair encoding (Syllable-BPE) segmentation for all NMT experiments. In addition, we compared the performance of SMT and NMT experiments with the RNN and transformer.

We found that LSTM cell type with RNN architecture is the best for Dawei Myanmar and Myanmar-Dawei neural machine translation. Myanmar to Dawei NMT achieved comparable results with PBSMT and HPBSMT. Moreover, Dawei to Myanmar RNN machine translation performance achieved higher BLEU scores than PBSMT (+1.06 BLEU) and HPBSMT (+1.37 BLEU) even with the limited parallel corpus.

II. Related Work

Karima Meftouh et al. built PADIC (Parallel Arabic Dialect Corpus) corpus from scratch, then conducted experiments on cross dialect Arabic machine translation [10]. PADIC is composed of dialects from both the Maghreb and the Middle-East. Some interesting results were achieved even with the limited corpora of 6,400 parallel sentences.

Using SMT for dialectal varieties usually suffers from data sparsity, but combining word-level and character-level models can yield good results even with small training data by exploiting the relative proximity between the two varieties [11]. Friedrich Neubarth et al. described a specific problem and its solution, arising with the translation between standard Austrian German and Viennese dialect. They used hybrid approach of rule-based preprocessing and PBSMT for getting better performance.

Pierre-Edouard Honnet et al. [6] proposed solutions for the machine translation of a family of dialects, Swiss German, for which parallel corpora are scarce. They presented three strategies for normalizing Swiss German input in order to address the regional and spelling diversity. The results show that character-based neural MT was the most promising one for text normalization and that in combination with PBSMT achieved 36% BLEU score.

III. Dawei Language

The Tavoyan or Dawei dialect of Burmese is spoken in Dawei (Tavoy), in the coastal Tanintharyi Region of southern Myanmar (Burma). The large and quite distinct Dawei or Tavoyan variety (tvn) is spoken in and around Dawei (formerly Tavoy) in Tanintharyi (formerly Tenasserim) by about 400,000 people; its sterotyped characteristic is the mesial ~/I/, found in earlist Bagan inscriptions but by merger there nearly 800 years ago; for further information see Pe Maung Tin (1933) and Okell (1995) [13]. Htawei formerly known as Tavoy is a city of south-eastern Myanmar and is the capital of Tanintharyi Region, formerly known as the Tenasserim is bounded by Mon state to the north, Thailand to the east and south, and the Andaman sea to the west.

Tavoyan retains /-l-/ medial that has since merged into the /-j-/ medial in standard Burmese and can form the following consonant clusters: /gl-/, /kl-/, /kʰl-/, /bl-/, /pl-/, /pʰl-/, /ml-/, /m̥l-/. Examples include မ္လေ (/mlè/ → Standard Burmese /mjè/) for "ground" and ကျောင်း (/kláʊɴ/ → Standard Burmese /tʃáʊɴ/) for "school"[14]. Also, voicing only with unaspirated consonants, whereas in standard Burmese, voicing can occur with both aspirated and unaspirated consonants. Also, there are many loan words from Malay and Thai not found in Standard Burmese. An example is the word for goat, which is hseit (ဆိတ်) in Standard Burmese but be (ဘဲ) in Tavoyan.

In the Tavoyan dialect, terms of endearment, as well as family terms, are considerably different from Standard Burmese. For instance, the terms for "son" and "daughter" are ဖစု (/pʰa̰ òu/) and မိစု (/mḭ òu/) respectively. Moreover, the honorific နောင် (Naung) is used in lieu of မောင် (Maung) for young males.

Another evidence of "Dawei" is "Dhommarazaka" pogoda inscription of Bagan period. It was inscription of Bagan period. It was inscribed in AD 1196 during the region of Bagan King Narapatisithu(AD 1174-1201). In this inscription line 6 to 19, when the demarcation of Bagan is mentioned "Taung-Kar-Htawei" (up to Htawei to the south) and "Taninthaye" (Tanintharyi) are including. Therefore, the name of "Dawei"

appeared particulary since Bagan period, at the time of the first Myanmar Empire (Dawei was established at Myanmar year 1116) is actually meant that the present name Dawei appears as the name of the settlers later and the original name of the city is Tharyarwady, which was established at Myanmar year 1116 according to the saying. As "Dawei" nationality deserves as one nationalist in our country. Actually, Dawei region is a place where local people lived since very ancient Stone Age. After that, Stone Age, Bronze Age and Iron Age culture developed. Moreover, as there has sound evidence of Thargara ancient city, comtemporary to Phu Period, the Dawei people, can be assumed that they are one nationality of high culture in Myanmar.

Dawei(Tavoyan) usage and vocabularies is divided into three main groups. The first one is using Myanmar vocabularies with Dawei speech, the second is the vocabularies same with Myanmar vocabularies and using isolated Dawei words and vocabularies.

In Myanmar word " ထို,ဟို" , " here, there" is used in Dawei "သယ်", "here" and "ဟောက်" , "there". For example "သယ်မျိုး", "ဒီလို" and "ဟောက်မျိုး", "ဟိုလို". The question words in Myanmar "နည်း(သနည်း)", "လဲ(သလဲ)" is used and the Dawei word "လော,လော်" is used instead of "လား(သလား)". For example "ဘာလဲ(what)" and 'ဘာဖြစ်တာလဲ(what happened)" is "ဖြာနူး" and " ဖြာဖြစ်နူး" in Dawei usage. In negative sense of Myanmar word "ဘူး" is not usually used in Dawei word. The negative Dawei word "ဟှ(ရ) " " or "ဟန်း" instead of " No " in Myanmar word. Myanmar adverb word "သိပ်" , "အလွန်", "အလွန်အလွန်" (very , extremely) " is used as Dawei word "ရရာ" , "ရမိရရာ" , "ဗြင်း" . There are many Dawei vocabularies such as "ဝန်းရှင်း " is called in Myanmar " ကိုယ်ဝန်ဆောင်" (pregnant) " , in Dawei word "ကောန်သား" , Myanmar word "ကောင်လေး (boy) " , "Dawei word ဝယ်သား", Myanmar word " ကောင်မလေး (girl) " , Dawei word " ကပ် " , "ပိုက်ဆံ (money) " , Dawei word " ချော့-က်တိုအိုးသီး" , Myanmar word "ကျွဲကောသီး (pomelo)" and Dawei word "သစ်ခတ်ကွား", Myanmar word "ကျားသစ်(leopard)".

The followings are some example parallel sentences of Myanmar (my) and Dawei (dw):

dw : <u>သယ်ဝယ်သား</u> က လှ ဗြင်း ဟှယ် ။
my : <u>ဒီကောင်မလေး</u> က လှ လွန်း တယ် ။
("<u>The girl</u> is so beautiful" in English)

dw : <u>လတ်ဖတ်ရယ်</u> က ရှိ ဗြင်း ဟှယ် ။
my : <u>လက်ဖက်ရည်</u> က ချို လွန်း တယ် ။
("The <u>tea</u> is so sweet" in English)

dw : ကောန်သား <u>ကျောန်း</u> မှန်းမှန် သွား ဟှယ် ။
my : ကောင်လေး <u>ကျောင်း</u> မှန်မှန် တက် တယ် ။
("The boy goes to <u>school</u> regularly" in English)

IV. METHODOLOGY

In this section, we describe the methodology used in the machine translation experiments for this paper.

A. Encoder-Decoder Model

The core idea is to encode a variable-length input sequence of tokens into a sequence of vector representations, and to then decode those representations into a sequence of output tokens. This decoding is conditioned on information from both the latent input vector encodings as well as its own continually updated internal state, motivating the idea that the model should be able to capture meanings and interactions beyond those at the word level. Formally, given source sentence $X= x_1,...,x_n$ and target sentence$Y= y_1,...,y_m$, an NMT system models $p(Y|X)$as a target language sequence model, conditioning the probability of the target word y_t on the target history $Y_{1:t-1}$ and source sentence X . Each x_I and y_t are integer ids given by source and target vocabulary mappings,V_{src} and V_{trg}, built from the training data tokens and represented as one-hot vectors $x_i \in \{0,1\}^{|V_{src}|}$ and $y_t \in \{0,1\}^{|V_{trg}|}$. These are embedded into e-dimensional vector representations, $E_S x_i$ and $E_T y$, using embedding matrices $\mathbb{R}^{e\times|V_{src}|}$ and $E_T \in \mathbb{R}^{e\times|V_{trg}|}$. The target sequence is factorized as $p(Y|X;\theta) = \prod_{m}^{t=1} p(y_t|Y_{1:t-1},X;\theta)$. The model, parameterized by θ, consists of an encoder and a decoder part, which vary depending on the model architecture . $p(y_t|Y_{1:t-1},X;\theta)$ is parameterized via a

softmax output layer over some decoder representation

(1)

where W_o scales to the dimension of the target vocabulary V_{trg}. For parameterizing the encoder and the decoder, Recurrent Neural Networks (RNNs) constitute a natural way of encoding variable-length sequences by updating an internal state and returning a sequence of outputs, where the last output may

$$p(y_t|Y_{1:t-1}, X;) = softmax(W_0 t + b_0)$$

summarize the encoded sequence [15]. However, with longer sequences, a fixed-size vector may not be able to store sufficient information about the sequence, and attention mechanisms help to reduce this burden by dynamically adjusting the representation of the encoded sequence given a state [16].

B. Recurrent Neural Network

The first encoder layer consists of a bi-directional RNN followed by a stack of uni-directional RNNs. Specifically, the first layer produces a forward sequence of hidden states, some non-linear function, such as a Gated Recurrent Unit (GRU) or Long Short Term Memory (LSTM) cell. The reverse RNN processes the source sentence from right to left: and the hidden states from both directions are concatenated. The hidden state h_0 it can incorporate information from both tokens to the left, as well as tokens to the right.

The decoder consists of an RNN to predict one target word at a time through a state vector

$$\mathbf{s}_t = f_{dec}([\mathbf{E}_T \mathbf{y}_{t-1}; \bar{\mathbf{s}}_{t-1}], \mathbf{s}_{t-1}), \tag{2}$$

where f_{dec} is a multi-layer RNN, s_{t-1} the previous state vector, and $\bar{s}_{t-1}$ the source-depend attentional vector. Providing the attentional vector as an input to the first decoder layer is also called input feeding [17]. The initial decoder hidden state is a non-linear transformation of the last encoder hidden state:$s_0 = \tanh(W_{init} h_n + b_{init})$. The attentional vector $\bar{s}_t$ combines the decoder state with a context vector c_t:

$$\bar{\mathbf{s}}_t = \tanh(\mathbf{W}_{\bar{s}}[\mathbf{s}_t; \mathbf{c}_t]), \tag{3}$$

where c_t is a weighted sum of encoder hidden states:$c_t = \sum_{i=1}^{n} \alpha_{ti} h_i$. The attention vector α_t is computed by an attention network $\alpha_{ti} = \text{softmax}(\text{score}(s_t, h_i))$ where score (s,h) could be defined as

$$\alpha_{ti} = \text{softmax}(\text{score}(\mathbf{s}_t, \mathbf{h}_i))$$
$$\text{score}(\mathbf{s}, \mathbf{h}) = \mathbf{v}_a^\top \tanh(\mathbf{W}_u \mathbf{s} + \mathbf{W}_v \mathbf{h}). \tag{4}$$

, the computation of RNN hidden states cannot be parallelized over time.

C. Transformer

The transformer model [17] uses attention to replace recurrent dependencies, making the representation at time step independent from the other time steps. This allows for parallelization of the computation for all time steps in encoder and decoder.

The embedding is followed by several identical encoder blocks consisting of two core sub layers: self-attention and a feed-forward network. The self-attention mechanism is a variation of the dot-product attention [16] but generalized to three inputs: a query matrix $Q \in \mathbb{R}^{n \times d}$, a key matrix $K \in \mathbb{R}^{n \times d}$, and a value matrix $V \in \mathbb{R}^{n \times d}$, where d denotes the number of hidden units. Vaswani et al. [2017] further extend attention to multiple heads, allowing for focusing on different parts of the input. A single head u produces a context matrix

$$\mathbf{C}_u = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_u^Q (\mathbf{K}\mathbf{W}_u^K)^T}{\sqrt{d_u}}\right)\mathbf{V}\mathbf{W}_u^V, \tag{5}$$

where matrices W_u^Q, W_u^K b and W_u^V belong to $R^{d \times du}$. The final context matrix is given by concatenating the heads, followed by a linear transformation : $C = [C_1; ...; C_h]W^O$. The number of hidden units chosen to be a multiple of the number of units per head. Given a sequence of hidden states h_i (or input embeddings), concatenated to $H \in \mathbb{R}^{n \times d}$, the encoder computes self-attention using Q=K=V=H. The second sub network of an encoder block is a feed-forward network with ReLU activation defined as

$$FFN(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{6}$$

which again is easily parallelizable across time steps. Each sub layer, self-attention and feed-forward network, is followed by a post-processing stack of dropout, layer normalization [18], and residual connection. A complete encoder block hence consists of Self-attention → Post-process → Feed-forward → Post-process can be stacked to form a multi-layer encoder. To maintain auto-regressiveness of the model, attention on future time steps is masked out accordingly [17]. In addition to self-attention, a source attention layer which uses the encoder hidden

states as key and value inputs is added. The sequence of operations of a single decoder block is:

Selfattention→Postprocess→Encoderattention→Post -process→Feed-forward→Post-process

Multiple blocks are stacked to form the full decoder network and the representation of the last block is fed into the output layer of Equation.

V. EXPERIMENTS

A.Corpus Statistics

We used 9K Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus [19], which is a parallel corpus in the travel domain. It contains six main categories and they are people (greeting, introduction and communication), survival (transportation, accommodation and finance), food (food, beverage and restaurant), fun (recreation, traveling, shopping and nightlife), resource (number, time and accuracy), special needs (emergency and health). We used 9k Myanmar-Dawei parallel corpus and 6,883 sentences for traning, 1,215 sentences for development and 902 sentences for testing.

B. Syllable Segmentation

We defined Generally, Myanmar words are composed of multiple syllables, and most of the syllables are composed of more than one character. Syllables are composed of Myanmar words. If we only focus on consonant-based syllables, the structure of the syllable can be described with Backus normal form (BNF) as follows:

Syllable := CMW[CK][D]

Here, C stands for consonants, M for medials, V for vowel, K for vowel killer character, and D for diacritic characters. Myanmar syllable segmentation can be done with a rule-based approach, finite state automation (FSA) or regular expressions (RE)

(https://github.com/yekyawthu/sylbreak).

In our experiments, we used RE based Myanmar syllable segmentation tool named. The following is an example of syllable segmentation for a Dawei sentence in our corpus and the meaning is
Unsegmented Dawei sentence:
dw: ဘယ်နားက လာဟို့ ဘယ်နားသွားဟို့ ဘဲ့သေးနူး ။
Syllable Segmented Dawei sentence:
dw: ဘယ် နား က လာ ဟို့ ဘယ် နား သွား ဟို့ ဘဲ့ သေး နူး ။

C. Byte-Pair-Encoding

Sennrich et al., 2016 [21] proposed a method to enable open-vocabulary translation of rare and unknown words as a sequence of subword units as a sequence of subword units representing BPE algorithm [22]. The input is a monolingual corpus for a language (one side of the parallel training data, in our case) and starts with an initial vocabulary, the characters in the text corpus. The vocabulary is updated using an iterative greedy algorithm. In every iteration, the most frequent bigram (based on the currebt vocabulary) in the corpus is added to the vocabulary (the merge operation). The corpus is again encoded the updated vocabulary, and this process is repeated for a predetermined number of merge operations. The number of merge operations is the only hyper-parameter of the system that needs to be tuned. A new word can be segmented by looking up the learnt vocabulary.

D. Moses SMT System

We used the Moses toolkit (Koehn et al., 2007) for training the operation sequence model (OSM) statistical machine translation systems. We did not consider phrase-based statistical machine translation (PBSMT) and hierarchical phrase-based statistical machine translation (HPBSMT), because the OSM approach achieved the highest BLEU [7] and RIBES [8] scores among three approaches [24] for both Myanmar-Dawei to Dawei-Myanmar statistical machine translations. The word-segmented (i.e., Syllable-BPE) source language was aligned with the word-segmented target language using GIZA++. The alignment was symmetrized by grow-diag-final and heuristic. The lexicalized reordering model was trained with the msd-bidirectional-fe option. We used KenLM [9] for training the 5-gram language model with modified Kneser- Ney discounting. Minimum error rate training (MERT) was used to tune the decoder parameters, and the decoding was done using the Moses decoder (version 2.1.1) [23]. We used the default settings of Moses for all experiments.

E. Framework of NMT System

An open-source sequence-to-sequence toolkit for NMT written in Python [20] and built on Apache MXNET, the toolkit offers scalable training and

inference for the three most prominent encoder - decoder architectures: attentional recurrent neural network, self-attentional transformers.

F. Training Details

We used the Sockeye toolkit, which is based on MXNet, to train NMT models. The initial learning rate is set to 0.0001. All experiments are runned with maximum epoch. If the performance on the validation set has improves for 8 checkpoints, the learning rate is multiplied by 8 checkpoints. All the neural networks have eight layers. For RNN Seq2Seq, the encoder has one bi-directional LSTM and six stacked unidirectional LSTMs, and the encoder is a stack of eight unidirectional LSTMs. The size of hidden states is 512. We apply layer-normalization and label smoothing (0.1) in all models. We tie the source and target embedding. The dropout rate of the embedding and transformer blocks is set to (0.1). The dropout rate of RNNs is (0.2). The attention mechanism in the transformer has eight heads.

We applied three different batch sizes (128, 256 ,512 and 1024) for RNN and Transformer architectures. The learning rates varies from 0.0001 to 0.0005. Two memory cell types GRU and LSTM were used for the RNN and transformer. The comparison was done for both SMT (i.e., PBSMT, HPBSMT) and NMT (RNN, Transformer) techniques. All experiments are run on NVIDIA Tesla K80 24GB GDDR5, We trained all models for the maximum number of epochs using the AdaGrad and adaptive moment estimation (Adam) optimizer. The BPE segmentation models were trained with a vocabulary of 8,000.

VI. Results and Discussion

In discussion of Table I, different batch sizes gained comparable results for both RNN and Transformer architecture for both cell types. The results of Table II, Batch size 128 is the best score in RNN architecture. Therefore, we are runned batch size 128 with two different cell types LSTM and GRU with different learning rates. In evidence of the results of Table III, the learning rate 0.0005 is the best score of bi-directional Myanmar-Dawei machine translation. In comparing the results of two cell type of GRU and LSTM, LSTM is more suitable than GRU in RNN architecture. For the results of Table III and Table IV, we can be clearly say that RNN is better results than transformer model architecture.

Discussion about batch size 256 shown in table V and table VI. LSTM cell type of RNN architecture is better results than GRU in bi-directional Myanmar to Dawei machine translation for the result of table V. But in transformer architecture of batch size 256 GRU cell type is get better results than LSTM shown in the results of table VI.

In Table V, the learning rate 0.0005 is gained the highest score the Myanmar-Dawei machine translation that score is 44.68 on batch size 256 on LSTM cell type in RNN architecture and Dawei-Myanmar machine translation highest score is 61.84 is on batch size 128 on LSTM type on RNN. In summarization of above facts, LSTM cell type of RNN architecture is best score for Myanmar to Dawei bi-directional machine translation.

In summarization of the results of batch size 512 on Table VII and Table VIII LSTM cell type is best score for Myanmar to Dawei machine translation and GRU cell type is best score for Dawei to Myanmar machine translation. Comparing the discussion results is shown in Table IX. We found that LSTM cell type with RNN architecture is the best for Dawei-Myanmar bi-directional machine translation. Myanmar to Dawei NMT achieved comparable results with PBSMT and HPBSMT. Moreover, Dawei to Myanmar RNN machine translation performance achieved higher BLEU scores than PBSMT (+1.06 BLEU) and HPBSMT (+1.37 BLEU) even with the limited parallel corpus.

TABLE I. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS BATCH SIZES WITH TWO DIFFERENT CELL TYPE FOR TRANSFORMER MODEL

Batch size	Transformer			
	LSTM		*GRU*	
	my-dw	*dw-my*	*my-dw*	*dw-my*
128	42.32	57.5	42.31	57.50
256	42.16	57.57	**42.37**	**57.84**
512	**42.21**	**57.6**	42.21	57.60
1024	41.85	57.12	41.85	56.38

TABLE II. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS BATCH SIZES WITH TWO DIFFERENT CELL TYPE FOR RNN MODEL

Batch size	RNN			
	LSTM		*GRU*	
	my-dw	*dw-my*	*my-dw*	*dw-my*
128	**40.96**	**57.51**	**41.29**	**57.43**
256	39.5	54.13	40.67	56.63
512	38.86	55.95	38.95	52.48
1024	35.82	51.08	37.43	53.63

TABLE III. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS LEARNING RATES ON DIFFERENT CELL TYPE FOR RNN MODEL ON BATCH 128

Learning	RNN

rate	LSTM		GRU	
	my-dw	dw-my	my-dw	dw-my
0.0002	43.47	58.04	41.6	59.27
0.0003	43.39	60.03	43.22	59.39
0.0004	42.94	60.6	42.94	60.00
0.0005	**44.24**	**61.84**	**43.23**	**60.38**

TABLE IV. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS LEARNING RATES ON DIFFERENT CELL TYPE FOR TRANSFORMER MODEL ON BATCH 128

Learning rate	transformer			
	LSTM		GRU	
	my-dw	dw-my	my-dw	dw-my
0.0002	41.97	58.72	**42.69**	58.49
0.0003	42.62	58.84	42.23	60.17
0.0004	42.72	**59.09**	42.50	59.94
0.0005	**43.03**	58.85	41.84	**60.44**

TABLE V. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS LEARNING RATES ON DIFFERENT CELL TYPE FOR RNN MODEL ON BATCH 256

Learning rate	RNN			
	LSTM		GRU	
	my-dw	dw-my	my-dw	dw-my
0.0002	41.93	59.55	43.34	59.46
0.0003	42.68	59.95	43.79	60.61
0.0004	43.28	59.55	**43.84**	60.77
0.0005	**44.68**	**61.31**	43.14	**61.24**

TABLE VI. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS LEARNING RATES ON DIFFERENT CELL TYPE FOR TRANSFORMER MODEL ON BATCH 256

Learning rate	transformer			
	LSTM		GRU	
	my-dw	dw-my	my-dw	dw-my
0.0002	42.84	58.25	42.32	57.53
0.0003	42.57	58.35	43.79	**58.35**
0.0004	42.48	58.61	**43.84**	59.17
0.0005	**42.70**	**59.17**	43.14	59.17

TABLE VII. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS LEARNING RATES ON DIFFERENT CELL TYPE FOR RNN MODEL ON BATCH 512

Learning rate	RNN			
	LSTM		GRU	
	my-dw	dw-my	my-dw	dw-my
0.0002	42.29	57.27	41.50	57.99
0.0003	43.46	59.64	42.71	57.19
0.0004	**45.58**	59.23	43.37	**60.04**
0.0005	43.72	**59.86**	**44.23**	60.20

TABLE VIII. BLEU SCORES OF SYLLABLE-BPE SEGMENTATION FOR VARIOUS LEARNING RATES ON DIFFERENT CELL TYPE FOR TRANSFORMER MODEL ON BATCH 512

Learning rate	transformer			
	LSTM		GRU	
	my-dw	dw-my	my-dw	dw-my
0.0002	41.83	57.84	**42.50**	57.88
0.0003	**42.97**	58.41	42.04	58.41
0.0004	42.12	58.82	42.12	57.93
0.0005	41.22	**59.71**	41.94	**60.44**

TABLE IX. BLEU SCORES OF COMPARISON OF SMT AND NMT

	PBSMT	HPBSMT	RNN	Trasnsformer
my-dw	44.80	45.44	44.68	43.03
dw-my	60.78	60.47	**61.84**	60.44

VII. CONCLUSION

In this section, This work is the first evaluation of the quality of neural machine translation between Myanmar (Burmese) and Dawei (Tavoyan). We are developing a Myanmar-Dawei parallel corpus (around 9K sentences) based on the Myanmar language of ASEAN MT corpus and implemented on two prominent neural machine translation system : Recurrent Neural Network (RNN) and Transformer with syllable segmentation. We also investigated various hyper-parameters such as batch size, learning rate and cell types (GRU and LSTM). We proved that LSTM cell type with RNN architecture is the best for Dawei-Myanmar bi-directional machine translation. We are compared results between SMT (PBSMT and HPSMT) and NMT (RNN) in Myanmar to Dawei machine translation. We found that LSTM cell type with RNN architecture is the best for Dawei-Myanmar bi-directional machine translation. Myanmar to Dawei NMT achieved comparable results with PBSMT and HPBSMT. Moreover, Dawei to Myanmar RNN machine translation performance achieved higher BLEU scores than PBSMT (+1.06 BLEU) and HPBSMT (+1.37 BLEU) even with the limited parallel corpus. In the near future, we plan to conduct a further study with a focus on NMT models with one more subword segmentation scheme SentencePiece for Myanmar-Dawei NMT. Moreover, we intend to investigate SMT and NMT approaches for Myeik dialectal languages.

ACKNOWLEDGEMENT

We would like to thank U Aung Myo (Leading Charge, Dawei Ethnic Organizing Committee, DEOC) for his advice especially on writing system of Dawei language with Myanmar characters. We are very greatful to Daw Thiri Hlaing (Lecturer, University of Computer Studies Dawei) for her leading the Myanmar-Dawei Translation Team. We would like to thank all students of Myanmar-Dawei translation team namely, Aung Myat Shein, Aung

Paing, Aye Thiri Htun, Aye Thiri Mon, Htet Soe San, Ming Maung Hein, Nay Lin Htet, Thuzar Win Htet, Win Theingi Kyaw, Zin Bo Hein and Zin Wai for translation between Myanmar and Dawei sentences. Last but not least, we would like to thank Daw Khin Aye Than (Prorector, University of Computer Studies Dawei) for all the help and support during our stay at University of Computer Studies Dawei. We are very greatful to Daw Thiri Naing (Lecturer, University of Computer Studies , Dawei) for her leading the Myanmar-Dawei Translation Team.

REFERENCES

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation." in Proc. of HTL-NAACL, 2003, pp. 48–54.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation." in Proc. of ACL, 2007, pp. 177–180.

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation." in Proc. of MT summit, 2005, pp. 79–86.

[4] lya Sutskever, Oriol Vinyals, and Quoc V. Le.2014a. "Sequence to sequence learning with neuralNetworks" In Proceedings of the 27th InternationalConference on Neural InformationProcessing Systems - Volume 2, NIPS'14, pages3104–3112, Cambridge, MA, USA. MIT Press.

[5] Matīss Rikters, Mārcis Pinnis, and Rihard Krišlauks. 2018." Training and Adapting Multilingual NMT for Less-resourced and Morphologically Rich Languages" In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

[6] Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2017. "Machine translation of low-resource spoken dialects: Strategies for normalizing swiss german." CoRR, abs/1710.11035.

[7] 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 Associationfor Computational Linguistics, ACL '02,pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[8] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, KatsuhitoSudoh, and Hajime Tsukada. 2010. "Automatic evaluation of translation quality for distant language pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing", pages 944–952, Cambridge, MA. Association for Computational Linguistics.

[9] Kenneth Heafield. 2011. "Kenlm: Faster and smaller language model queries", In Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT'11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

[10] Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas and Kamel Smaili, "Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus", oin Proc. of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015, pp. 26-34.

[11] Neubarth Friedrich, Haddow Barry, Huerta Adolfo Hernandez and Trost Harald, "A Hybrid Approach to Statistical Machine Translation Between Standard and Dialectal Varieties", Human Language Technology, Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznan, Poland, December 7-9, 2013, Revised Selected Papers, pp .341–353.

[12] Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat and Michael Baeriswyl, "Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German", CoRR journal, volume (abs/1710.11035), 2017.

[13] John Okell , "Three Burmese Dialects", 1981, London Oxford University press, Univeristy of London.

[14] https://en.wikipedia.org/wiki/Tavoyan_dialects

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le.2014a, "Sequence to sequence learning with neural Networks" In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

[16] Minh-Thang Luong, Hieu Pham, Christopher D. Manning, " Effective Approaches to Attention-

based Neural Machine Translation", Computation and Linguistics in Sep 2015

[17] 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 I. Guyon, U. V.Luxburg, S. Bengio, H. Wallach, R. Fergus,S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

[18] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016." Layer normalization." CoRR, abs/1607.06450.

[19] Prachya, Boonkwan and Thepchai, Supnithi, "Technical Report for The Network-based ASEAN Language Translation Public Service Project", Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013

[20] Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: "A toolkit for neuralmachine translation". CoRR, abs/1712.05690.

[21] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. "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. Association for Computational Linguistics.

[22] Philip Gage. 1994. "A new algorithm for data compression"C Users J., 12(2):23–38.

[23] Philipp Koehn, Hieu Hoang, Alexandra Birch,Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics

[24] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe and Thepchai Supnithi " Statistical Machine Translation Between Myanmar (Burmese) and Dawei (Tavoyan)", The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019), 11-12, September 2019, University of Trento, Italy