Neural Machine Translation between Myanmar Sign Language and Myanmar Written Text

Swe Zin Moe[†] **Ye Kyaw Thu**[‡] **Hnin Aye Thant**[†] **Nandar Win Min**[†]
[†]University of Technology (Yatanarpon Cyber City), Myanmar
[‡]Language and Speech Science Research Lab., Waseda University, Japan
{swezinmoe.1011, wasedakuma, hninayethant, nandarwinmin}@gmail.com

Abstract

We explore Neural Machine Translation (NMT) between Myanmar Sign Language (MSL) and Myanmar Written Text (MWT). Our developing MSL-MWT parallel corpus was used and the experiments were carried out using three different NMT approaches: Recurrent Neural Network (RNN), Transformer, and the Convolutional Neural Network (CNN). In addition, four different segmentation schemes for word embedding were studies, these were syllable segmentation, word segmentation (sign unit based word segmentation for MSL), SentencePiece and the Byte-Pair-Encoding (BPE). The results show that the highest quality NMT and Statistical Machine Translation (SMT) performances were attained with syllable segmentation for both MSL and MWT. We found that Transformer outperformed both CNN and RNN for MWT-to-MSL and MSL-to-MWT translation tasks.

Keywords: Neural Machine Translation (NMT), Myanmar Sign Language (MSL), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Byte-Pair-Encoding (BPE)

1 Introduction

Most of the deaf people are having serious problems when expressing themselves in written languages or understanding written texts. Naturally, hearing problems can affect the ability to read or write the written or spoken languages. This fact can cause deaf people to have problems when accessing information, education, job, social relationship, knowledge, etc. In Myanmar, MSL is the primary means of communication for about 673,126 deaf people [1], although there are not enough sign language interpreters and communication systems. Machine Translation (MT) of MSL would be useful in enabling hearing people who do not know MSL to communicate with Deaf individuals.

Our research contribution is to explore NMT between MSL and MWT and presenting appropriate hyper-parameters (batch size and optimizer) for MWT-MSL and MSL-MWT translations NMT experiments. One more contribution is we are developing a parallel corpus of MSL and MWT and we used the current version of the corpus for our experiments. Furthermore, we can make a comparison between SMT and NMT for MWT-MSL and MSL-MWT MT. We did NMT experiments with RNN, Transformer, and the CNN.

The structure of the paper is as follows. In the next section, we present a brief review of machine translation systems for text to SL. Section 3 presents a sketch of MSL and Myanmar language. Section 4 presents preparation of the MSL corpus for machine translation experiments. Section 5 gives the detail information about all four segmentation schemes. Then, in Section 6, we describe the methodologies used in the machine translation experiments. Section 7 presents statistical information of the corpus and the experimental settings. The results together with some discussions are presented in Section 8. Section 9 presents the error analysis of translated sentences. Finally in Section 10, we present our conclusions and indicate promising results for future research.

2 MT for Sign Language

MT systems between spoken and sign languages had a start in the late 90s. Strategies used for developing MT system are also used for developing text to sign language MT system including direct MT, template-

based MT, transfer-based MT, interlingua-based MT, rule-based MT, knowledge-based MT, example-based MT, syntax-based MT and statistical-based MT. Details of each strategy can be found in several books as follows: Hutchins and Somers, 1992 [2]; Hutchins, 2000 [3]; Nirenburg and Raskin, 2004 [4]. A number of text to sign language translation systems have been carried out around the world, e.g. TESSA system (Bangham & Cox, 2000) [5], weather reports generate system (Angus & Smith, 1999) [6], ViSiCAST Translator (Safar & Marshall, 2000) [7], TEAM Project (Zhao & Kipper, 2000) [8], ZARDOZ system (Veale & Collins, 1998) [9], ASL Workbench (Armond & Speers, 2001) [10], South African sign language machine translation system (Zijl & Barker, 2003) [11], TGT system-polish text into sign language (Suszczanska & Szmal, 2002) [12], spatial and planning models of ASL classifier predicates for MT and American sign language generation: Multimodal natural language generation (NLG) with multiple linguistic channels (Huenerfauth, 2004, 2005) [13], [14], [15], [16], [17], [19], experiments in sign language machine translation using examples (Morrissey & Way, 2006) [20] and Morpho-syntax base statistical methods for automatic sign language translation (Stein, Bungeroth, & Ney, 2006) [21]. Most of them are English-to-American Sign Language (ASL).

3 MSL and Myanmar Language

MSL like other known Sign Languages (SLs) depends on three basic factors that are used to represent the Manual Features (MFs): hand shape, hand location and orientation. In addition to the MFs, MSL also has Non-Manual Features (NMFs) that are related to head, face, eyes, eyebrows, shoulders and facial expressions like puffed checks and mouth pattern movements. Postures or movements of the body, head, eyebrows, eyes, cheeks, and mouth are used in various combinations to show several categories of information, including lexical distinction, grammatical structure, adjectival or adverbial content, and discourse functions [22]. Grammatical structure that is shown through non-manual signs includes questions, negation, relative clauses [23], boundaries between sentences [24], and the argument structure of some verbs [25]. Similar to ASL and British Sign Language (BSL), MSL use non-manual marking for yes/no questions. They are shown through raised eyebrows and a forward head tilt [26], [27], [28]. Figure 1 shows an example of MSL sentence "မိသားစု (family)" "ဘယ်လောက် (how many) + NMFs – chin up and raised eyebrows for wh-question". The meaning of the MSL sentence is "မိသားစု မှာ လူ ဘယ် နှစ် ယောက် ရှိ သလဲ ။" in Myanmar language and "How many people are there in your family?" in English respectively.

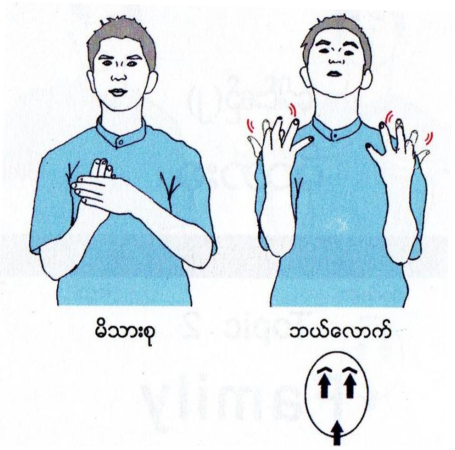

Figure 1. An example of MSL sentence that used non-manual features [28]

Sign language is different in Yangon and Mandalay regions with many dialects. To the best of our knowledge, MSL using in the Mary Chapman School for the Deaf Children, Yangon is mainly different with MSL of Mandalay region. This difference gives the difficulty of communicating and dealing between Deaf or hearing disabilities in different cities. A government project was set up in 2010 to establish a national sign language with the aid of the Japanese Federation of the Deaf.

MSL is a full natural language that includes various linguistic structures (e.g., grammars, vocabularies, word order, etc.) distinct from Myanmar written language. Myanmar language is tonal and syllable-based. Examples of different grammar,

word order and vocabulary used between Myanmar and MSL can be seen in the followings.

English: What time do you wake up?
Myanmar: ဘယ် အချိန် အိပ်ယာ က ထ သလဲ ။
MSL: အိပ်ယာထ (wake up) အချိန် (time) ဘာလဲ (what)

English: I wake up at six o'clock.
Myanmar: မနက် ခြောက် နာရီ မှာ ထ လေ့ ရှိ ပါတယ် ။
MSL: မနက် (morning) နာရီ (o'clock) ခြောက် (six)

English: Daughter-in-law
Myanmar: ခွေးမ ။
MSL: သား (son) လက်ထပ် (marries) မိန်းကလေး (girl)

4 Corpus Preparation

Currently, there is no parallel corpus for MSL. Therefore, as a first step, we are building a multimedia parallel MSL corpus with the purpose of developing a MT-based approach for using technology to assist hearing and speaking disabilities with limited Myanmar language in their daily life basic conversation.

For this purpose data collection with 22 SL trainers and Deaf people: males and females, age range from 11 to 48, from School for the Deaf (Mandalay), Mary Chapman School for Deaf Children (Yangon), School for the Deaf (Tamwe), Myanmar Deaf Society and Literacy and Language Development for the Deaf in Yangon and Mandalay regions has been carried out. We also considered covering different MSL dialects. The MSL corpus contains MSL video, a textual representation of Myanmar sign language and translated Myanmar written text.

5 Segmentation

In this section, we give detail information about four different segmentation schemes for word embedding.

5.1 Word Segmentation

In Myanmar text, spaces are used for separating phrases for easier reading. It is not strictly necessary, and these spaces are rarely used in short sentences. There are no clear rules for using spaces in Myanmar language, and thus spaces may (or may not) be inserted between words, phrases, and even between root words and their affixes. Although we can implement conditional random fields (CRF) approach word segmentation model by using freely available word segmented Myanmar corpus such as myPOS [29], we did manual word segmentation for Myanmar text of our corpus. The reasons are the current myPOS corpus size is only 12K and we assumed that manual word segmentation is more suitable for the domain of our corpus. We applied the word segmentation rules proposed by (Win Pa Pa et al., 2015) [30].

We considered different segmentation schemes for Myanmar language sentence and MSL sentence. For MSL sentence, segmentation is based on meaningful MSL word unit. Sign unit based word segmentation was done manually for the whole parallel corpus. The followings show the different word segmentation between MSL and MWT ("She is engaged with Wunna." in English):

Word segmentation for MWT:
သူ ဝတ္ထာ နဲ့ စေ့စပ်ကြောင်းလမ်း ထား တယ် ။

Sign Unit based segmentation for MSL:
သူ စေ့စပ် ကောင်လေး နာမည် စာလုံးပေါင်း ဝတ္ထာ ။

5.2 Syllable Segmentation

Generally, Myanmar words are composed of multiple syllables and most of the syllables are composed of more than one character. Syllables are also basic units for pronunciation 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 := CMV[CK][D]

Here, C for consonants, M for medials, V for vowels, K for vowel killer character and D for diacritic characters [32]. Myanmar syllable segmentation can be done with rule based [33], [31], finite state automaton

(FSA) [34] or regular expression (RE) [35]. In our experiments, we used RE based Myanmar syllable segmentation tool named "sylbreak" [35]. The following example shows the syllable segmentation for MWT ("She is engaged with Wunna." in English):

Syllable segmentation for MWT:
သူ ဝတ္ထာ နဲ့ စေ့ စပ် ကြောင်း လမ်း ထား တယ် ။

5.3 SentencePiece

SentencePiece, a language-independent subword tokenizer and detokenizer designed for Neural-based text processing, including NMT [36]. It provides open-source C++ and Python implementations for subword units. While existing subword segmentation tools assume that the input is pre-tokenized into word sequences, SentencePiece can train subword models directly from raw sentences, which allows us to make a purely end-to-end and language independent system. The following example shows the SentencePiece segmentation for MWT ("She is engaged with Wunna." in English):

SentencePiece segmentation for MWT:
_သူ ဝ ဏ္ဍ ဏာ နဲ့စေ့စပ် ကြောင်း လမ်း ထားတယ် ။

5.4 Byte-Pair-Encoding

(Sennrich et al., 2016) [37] proposed a method to enable open-vocabulary translation by representing rare and unknown words as a sequence of subword units. This is achieved by adapting a compression algorithm called Byte-Pair-Encoding [38]. The essential idea is to start with a vocabulary of characters and keep extending the vocabulary with the most frequent n-gram pairs in the data set. One can choose to either build separate vocabularies for training and test sets or build one vocabulary jointly. After the vocabulary is built, an NMT system with some seq2seq architecture can be directly trained on these word segments. Notably, this method won top places in Workshop on Machine Translation (WMT) 2016. The following example shows the BPE segmentation for MWT ("She is engaged with Wunna." in English):

BPE segmentation for MWT:
သူ ဝ@@ ဏ္ဍ@@ ဏာ နဲ့ စေ့စ@@ ပ်@@ က@@ ေြာင်း@@ လမ်း ထား တယ် ။

6 Experimental Methodology

In this section, we describe the methodology used in the SMT and NMT experiments for this paper.

6.1 Phrase-based Statistical Machine Translation (PBSMT)

A PBSMT translation model is based on phrasal units [39], [40]. Here, a phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase. A phrase-based translation model typically gives better translation performance than word-based models. We can describe a simple phrase-based translation model consisting of phrase-pair probabilities extracted from the corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table [41].

6.2 Encoder-Decoder Models for NMT

6.2.1 Stacked Recurrent Neural Network (RNN) with Attention

The encoder consists of a bi-directional RNN followed by a stack of uni-directional RNNs. An RNN at layer l produces a sequence of hidden states $\mathbf{h}_1^l \ldots \mathbf{h}_n^l$:

$$\mathbf{h}_i^l = f_{enc}(\mathbf{h}_i^{l-1}, \mathbf{h}_{i-1}^l), \qquad (1)$$

where f_{enc} is some non-linear function, such as a Gated Recurrent Unit (GRU) or Long Short Term Memory (LSTM) cell, and $\mathbf{h}_i^{l-1}$ is the output from the lower layer at step i. The bidirectional RNN on the lowest layer uses the embeddings $\mathbf{E}_S\mathbf{x}_i$ as input and concatenates the hidden states of a forward and a reverse RNN: $\mathbf{h}_i^0 = [\overrightarrow{\mathbf{h}_i^0}; \overleftarrow{\mathbf{h}_i^0}]$. With deeper networks, learning turns increasingly difficult (Hochreiter et al., 2001) [42], (Pascanu et al., 2012) [43] and residual connections of the form $\mathbf{h}_i^l = \mathbf{h}_i^{l-1} + f_{enc}(\mathbf{h}_i^{l-1}, \mathbf{h}_{i-1}^l)$ become essential (He et al., 2016) [44].

The decoder consists of an RNN to predict one target word at a time through a state vector $\mathbf{s}$:

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

where f_{dec} is a multi-layer RNN, $\mathbf{s}_{t-1}$ the previous state vector, and $\bar{\mathbf{s}}_{t-1}$ the source-dependent *attentional vector*. Providing the attentional vector as an input to the first decoder layer is also called *input feeding* (Luong et al., 2015) [45]. The initial decoder hidden state is a non-linear transformation of the last encoder hidden state: $\mathbf{s}_0 = \tanh(\mathbf{W}_{init}\mathbf{h}_n + \mathbf{b}_{init})$. The attentional vector $\bar{\mathbf{s}}_t$ combines the decoder state with a *context vector* $\mathbf{c}_t$:

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

where $\mathbf{c}_t$ is a weighted sum of encoder hidden states: $\mathbf{c}_t = \sum_{i=1}^{n} \alpha_{ti}\mathbf{h}_i$. The attention vector α_t is computed by an attention network (Bahdanau et al., 2014) [46], (Luong et al., 2015) [45] :

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

6.2.2 Self-attentional Transformer

The transformer model (Vaswani et al., 2017) [47] uses attention to replace recurrent dependencies, making the representation at time step i independent from the other time steps. This requires the explicit encoding of positional information in the sequence by adding fixed or learned positional embeddings to the embedding vectors.

The encoder uses several identical blocks consisting of two core sublayers, self-attention and a feed-forward network. The self-attention mechanism is a variation of the dot-product attention (Luong et al., 2015) [45] generalized to three inputs: a query matrix $\mathbf{Q} \in \mathbb{R}^{n\times d}$, a key matrix $\mathbf{K} \in \mathbb{R}^{n\times d}$, and a value $\mathbf{V} \in \mathbb{R}^{n\times d}$, where d denotes the number of hidden units. [47] further extend attention to multiple *heads*, allowing for focusing on different parts of the input. A single *head* u produces a context matrix

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

where matrices $\mathbf{W}_u^Q$, $\mathbf{W}_u^K$ and $\mathbf{W}_u^V$ are in $\mathbb{R}^{d\times d_u}$. The final context matrix is given by concatenating the heads, followed by a linear transformation: $\mathbf{C} = [\mathbf{C}_1; \ldots; \mathbf{C}_h]\mathbf{W}^O$. The form in Equation 5 suggests parallel computation across all time steps in a single large matrix multiplication. Given a sequence of hidden states $\mathbf{h}_i$ (or input embeddings), concatenated to $\mathbf{H} \in \mathbb{R}^{n\times d}$, the encoder computes self-attention using $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{H}$. The second subnetwork 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 is also easily parallelizable across time steps. Each sublayer, self-attention and feedforward network, is followed by a post-processing stack of dropout, layer normalization (Ba et al., 2016) [48], and residual connection.

The decoder uses the same self-attention and feed-forward networks subnetworks. To maintain auto-regressiveness of the model, self-attention on future time steps is masked out accordingly [47]. In addition to self-attention, a source attention layer which uses the encoder hidden states as key and value inputs is added. Given decoder hidden states $\mathbf{S} \in \mathbb{R}^{m\times s}$ and the encoder hidden states of the final encoder layer $\mathbf{H}^l$, source attention is computed as in Equation 5 with $\mathbf{Q} = \mathbf{S}, \mathbf{K} = \mathbf{H}^l, \mathbf{V} = \mathbf{H}^l$. As in the encoder, each sublayer is followed by a post-processing stack of dropout, layer normalization (Ba et al., 2016) [48], and residual connection.

6.2.3 Fully Convolutional Models (ConvSeq2Seq)

The convolutional model (Gehring et al., 2017) [49] uses convolutional operations and also dispenses with recurrency. Hence, input embeddings are again augmented with explicit positional encodings.

The convolutional encoder applies a set of (stacked) convolutions that are defined as:

$$\mathbf{h}_i^l = v(\mathbf{W}^l[\mathbf{h}_{i-\lfloor k/2\rfloor}^{l-1}; \ldots; \mathbf{h}_{i+\lfloor k/2\rfloor}^{l-1}] + \mathbf{b}^l) + \mathbf{h}_i^{l-1}, \tag{7}$$

where v is a non-linearity such as a Gated Linear Unit (GLU) (Gehring et al., 2017)

[49], (Dauphin et al., 2016) [50] and $\mathbf{W}^l \in \mathbb{R}^{d_{cnn} \times kd}$ the convolutional filters. To increase the context window captured by the encoder architecture, multiple layers of convolutions are stacked. To maintain sequence length across multiple stacked convolutions, inputs are padded with zero vectors.

The decoder is similar to the encoder but adds an attention mechanism to every layer. The output of the target side convolution

$$\mathbf{s}_t^{l*} = v(\mathbf{W}^l[\bar{\mathbf{s}}_{t-k+1}^{l-1}; \ldots; \bar{\mathbf{s}}_t^{l-1}] + \mathbf{b}^l) \quad (8)$$

is combined to form $\mathbf{S}^*$ and then fed as an input to the attention mechanism of Equation 5 with a single attention head and $\mathbf{Q} = \mathbf{S}^*, \mathbf{K} = \mathbf{H}^l, \mathbf{V} = \mathbf{H}^l$, resulting in a set of context vectors $\mathbf{c}_t$. The full decoder hidden state is a residual combination with the context such that

$$\bar{\mathbf{s}}_t^l = \mathbf{s}_t^{l*} + \mathbf{c}_t + \bar{\mathbf{s}}_t^{l-1}. \quad (9)$$

To avoid convolving over future time steps at time t, the input is padded to the left.

7 Experiments

7.1 Corpus statistics

We used 2,510 MWT and MSL parallel sentences of MSL corpus, which is a collection of everyday basic conversation expressions. It contains six main categories and they are people (greeting, introduction, family, daily activities, education, occupations, and communication), food (food, beverage and restaurant), fun (shopping, hobbies and sports), resource (number, time, weather and accuracy), travel (bus, train and airport) and emergency (health, accident, police, fire, earthquake, flood and storm). In our MSL data, 6% of sentences are containing Myanmar fingerspelling characters. 2,000 sentences were used for training, 310 sentences for development and 200 sentences for evaluation.

7.2 Moses SMT system

We used the Phrase-based SMT provided by the Moses toolkit [51] for training the PBSMT statistical machine translation system. The word segmented source language was aligned with the word segmented target languages using GIZA++ [61]. The alignment was symmetrized by grow-diag-final-and heuristic [62]. The lexicalized recording model was trained with the msd-bidirectional-fe option [63]. We used KenLM for training the 5-gram language model with interpolated modified Kneser-Ney discounting [64], [65]. Minimum error rate training (MERT) [66] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1) [51]. We used default settings of Moses for all experiments.

7.3 Framework for NMT

Sockeye is an open-source sequence-to-sequence toolkit for NMT (Hieber et al., 2017) [52], written in Python and built on Apache MXNet (Chen et al., 2015) [53]. To the best of our knowledge, Sockeye is the only toolkit that includes implementations of all three major neural translation architectures: attentional recurrent neural networks (Schwenk, 2012) [54], (Kalchbrenner and Blunsom, 2013) [55], (Sutskever et al., 2014) [56], (Bahdanau et al., 2014) [46], (Luong et al., 2015) [45], self-attentional transformers (Vaswani et al., 2017) [47], and fully convolutional networks (Gehring et al., 2017) [49].

7.4 Training Details

We used the Sockeye [52] toolkit, which is based on MXNet [53], to train NMT models. The initial learning rate is set to 0.0002. If the performance on the validation set has not improved for 8 checkpoints, the learning rate is multiplied by 0.7. We set the early stopping patience to 32 checkpoints. All the neural networks have 8 layers. For RNN Seq2Seq, the encoder has 1 bi-directional LSTM and 6 stacked uni-directional LSTMs, and the decoder is a stack of 8 unidirectional LSTMs. The size of embeddings and hidden states is 512. We apply layer-normalization and label smoothing (0.1) in all models. We tie the source and target embeddings. The dropout rate of embeddings and Transformer blocks is set to (0.1). The dropout rate of RNNs is (0.2). The attention mechanism in Transformer has 8 heads.

Table 1. BLEU scores of word segmentation for three NMT approaches: RNN, Transformer and CNN on MWT→MSL and MSL→MWT translations tasks

Source - Target	NMT Approach	Batch size	Optimizer	BLEU
MWT→MSL	RNN	256	Adam	13.10
	CNN	128	Adam	**28.80**
	Transformer	256	Adagrad	27.91
MSL→MWT	RNN	256	Adam	12.30
	CNN	256	Adam	28.78
	Transformer	1024	Adam	**29.38**

Table 2. BLEU scores of syllable segmentation for three NMT approaches: RNN, Transformer and CNN on MWT→MSL and MSL→MWT translations tasks

Source - Target	NMT Approach	Batch size	Optimizer	BLEU
MWT→MSL	RNN	256	Adam	15.14
	CNN	256	Adam	**32.76**
	Transformer	512	Adagrad	29.68
MSL→MWT	RNN	256	Adam	12.17
	CNN	256	Adam	35.02
	Transformer	1024	Adam	**38.21**

Table 3. BLEU scores of SentencePiece segmentation for three NMT approaches: RNN, Transformer and CNN on MWT→MSL and MSL→MWT translations tasks

Source - Target	NMT Approach	Batch size	Optimizer	BLEU
MWT→MSL	RNN	256	Adam	12.02
	Transformer	256	Adam	22.13
	CNN	1024	Adam	**22.69**
MSL→MWT	RNN	256	Adam	8.97
	Transformer	1024	Adam	22.13
	CNN	256	Adam	**23.64**

Table 4. BLEU scores of Byte-Pair-Encoding segmentation for three NMT approaches: RNN, Transformer and CNN on MWT→MSL and MSL→MWT translations tasks

Source - Target	NMT Approach	Batch size	Optimizer	BLEU
MWT→MSL	RNN	1024	Adam	10.56
	Transformer	256	Adagrad	**29.39**
	CNN	256	Adam	26.91
MSL→MWT	RNN	256	Adam	28.05
	Transformer	256	Adagrad	**32.92**
	CNN	512	Adam	27.73

All experiments are run on a single GeForce GTX 1080 8GB ROG STRIX GPU. We trained all models for maximum epoch using the stochastic gradient descent (SGD) (Robbins and Monro, 1951) [57], Adagrad (Duchi et al., 2011) [58] and adaptive mo-

Table 5. SMT performances comparison on four types of segmentation units for MWT→MSL and MSL→MWT translations tasks

Approach	Segmentation	MWT→MSL	MSL→MWT
PBSMT	Word	29.69	32.26
	Syllable	**35.81**	**35.82**
	SentencePiece	28.80	26.30
	BPE	28.20	33.14

Table 6. NMT performances comparison on four types of segmentation units for MWT→MSL traslation

Approach	Segmentation	MWT→MSL
CNN	Word	28.80
CNN	Syllable	**32.76**
CNN	SentencePiece	22.69
Transformer	BPE	29.39

Table 7. NMT performances comparison on four types of segmentation units for MSL→MWT translation

Approach	Segmentation	MSL→MWT
Transformer	Word	29.38
Transformer	Syllable	**38.21**
CNN	SentencePiece	23.64
Transformer	BPE	32.92

ment estimation (Adam) (Kingma and Ba, 2014) [59] optimizers and the size of the training batches were set to 1024, 512, 256, 128 and 64. The BPE models were trained with a vocabulary size of 5,000.

7.5 Evaluation

We used automatic criteria for the evaluation of the machine translation output. The de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [60]. The BLEU score measures the adequacy of the translations language pairs such as Myanmar and English. The higher BLEU scores are better. Before computing BLEU, the translations were decomposed into their constituent syllables in order to ensure the results were cross-comparable.

8 Result and Discussion

The BLEU score results for the three NMT approaches (RNN, Transformer and CNN) for word segmentation, syllable segmentation, SentencePiece and Bye-Pair-Encoding are shown in Table 1, 2, 3 and 4 respectively. Bold numbers indicate the highest BLEU score among different batch size and optimizer combinations.

When we focus on NMT approaches, for RNN models, the combination of batch size 256 and Adam optimizer achieve the highest BLEU scores for all segmentation schemes on both MWT-MSL and MSL-MWT translations excluding the MWT-MSL translation of BPE segmentation (see Table 4). For Transformer models, the combination of batch size 1024 and Adam optimizer achieve the highest BLEU scores on MSL-MWT translation of word, syllable and SentencePiece segmentation schemes. The combination of batch size 256 and Adagrad optimizer achieve the highest BLEU scores for the BPE segmentation of Transformer models on both MWT-MSL and MSL-MWT translations (see Table 4). For CNN models, the combination of batch size 256 and Adam optimizer achieve the highest BLEU scores on MSL-MWT translation of word, syllable and SentencePiece segmentations.

From the overall results in Table 1, 2, 3 and 4, it can be clearly seen that CNN and Transformer approaches are significantly better than RNN. Although CNN results higher than RNN, it is significantly lower than Transformer especially for syllable and BPE segmentations. Transformer outperformed both CNN and RNN for MWT-to-MSL and MSL-to-MWT translation tasks.

Table 5 shows the BLEU score results for the SMT performances for word, sylla-

ble, SentencePiece and BPE segmentation on both MWT-MSL and MSL-MWT translations. Table 6 and 7 present the top NMT performance scores for each segmentation schemes (word, syllable, SentencePiece and BPE).

When we focus on four different segmentation schemes (see Table 5, 6 and 7), the highest BLEU scores 28.80 and 32.76 respectively were achieved by CNN for word and syllable segmentation on MWT-MSL translation. Transformer gave the highest BLEU scores 29.38 and 38.21 respectively for word and syllable segmentation on MSL-MWT translation. For SentencePiece segmentation, CNN gave the highest BLEU score 22.69 on MWT-MSL translation and BLEU score 23.64 on MSL-MWT translation. For BPE segmentation, Transformer gave the highest BLEU scores 29.39 and 32.92 respectively on both MWT-MSL and MSL-MWT translations.

Looking at the results in Table 5, 6 and 7, it is clear that the highest quality NMT and SMT performances were attained with syllable segmentation for both MSL and MWT. For MWT-MSL translation, the highest BLEU score 35.81 was achieved by SMT. For MSL-MWT translation, Transformer gave the higher BLEU score 38.21 than SMT. Surprisingly, if we only focus on BPE segmentation of MWT-MSL translation, Transformer gave the BLEU score 29.39 which is higher than that of SMT on our limited data size.

9 Error Analysis on NMT Approaches

In this paper, we focus on the performances of three NMT approaches (RNN, CNN, and Transformer). We analyzed the translated outputs of NMT models using Word Error Rate (WER). We used SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK version 2.4.10 (http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm) for making dynamic programming based alignments between reference (ref) and hypothesis (hyp) and calculation of WER. The formula for WER can be stated as equation (10):

$$WER = (I + D + S)100/N \qquad (10)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words and N is the number of words in the reference ($N = S + D + C$) [57]. Note that if the number of insertions is very high, the WER can be greater than 100%.

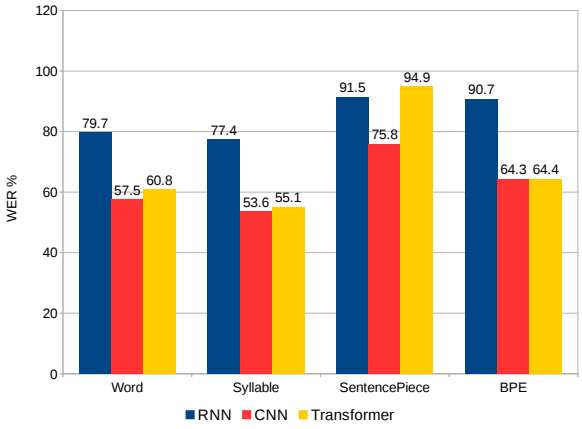

Figure 2. WER of three NMT approaches for MWT to MSL translation for four segmentation schemes

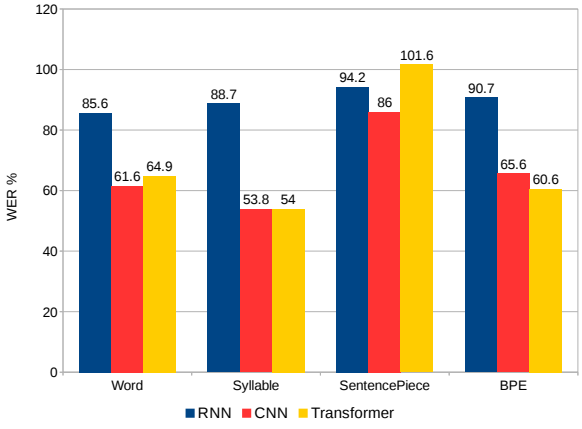

Figure 3. WER of three NMT approaches for MSL to MWT translation for four segmentation schemes

Figure 3 and 4 present the WER percentages of translation between MWT and MSL. The results show that syllable segmentation gave the lowest WER values for Transformer and CNN models and the difference

is higher for the SentencePiece segmentation for both MWT-MSL and MSL-MWT translation tasks (see Figure 3 and 4).

We also made manual error analysis on translated outputs of NMT models, and we found that dominant errors are different in sentence level. Several missing words errors are found in BPE segmentation and confusion words errors are found in word, SentencePiece and BPE segmentation of Transformer model on the MSL-MWT translation. The followings are some examples of missing words error (see underline word) that we found on BPE segmentation ("Can I borrow a book?" in English):

Word Segmentation

Scores: (#C #S #D #I) 6 0 0 0
REF: စာအုပ် ငှား လို့ ရ မလား ။
HYP: စာအုပ် ငှား လို့ ရ မလား ။
Eval:

Syllable Segmentation

Scores: (#C #S #D #I) 8 0 0 0
REF: စာ အုပ် ငှား လို့ ရ မ လား ။
HYP: စာ အုပ် ငှား လို့ ရ မ လား ။
Eval:

SentencePiece Segmentation

Scores: (#C #S #D #I) 5 0 0 0
REF: _ စာအုပ် ငှား လို့ရမလ ား။
HYP: _ စာအုပ် ငှား လို့ရမလ ား။
Eval:

BPE Segmentation

Scores: (#C #S #D #I) 5 0 1 0
REF: စာအုပ် ငှား လို့ ရ မလား ။
HYP: **** ငှား လို့ ရ မလား ။
Eval: D

The followings are some examples of confusion word errors (see underline words) that we found on word, SentencePiece and BPE segmentation ("I am getting stomach pain." in English):

Word Segmentation

Scores: (#C #S #D #I) 3 2 0 1
REF: ကျွန်တော် **** ဗိုက် နာ တယ် ။
HYP: ကျွန်တော် ပန်းနာရင်ကျပ် ဖြစ် နေ တယ် ။
Eval: I S S

Syllable Segmentation

Scores: (#C #S #D #I) 6 0 0 0
REF: ကျွန် တော် ဗိုက် နာ တယ် ။
HYP: ကျွန် တော် ဗိုက် နာ တယ် ။
Eval:

SentencePiece Segmentation

Scores: (#C #S #D #I) 3 2 0 1
REF: _ကျွန်တော် ********* ဗိုက် နာ တယ် ။
HYP: _ကျွန်တော် လည် ချောင်း နာ နေတယ် ။
Eval: I S S

BPE Segmentation

Scores: (#C #S #D #I) 4 1 0 0
REF: ကျွန်တော် ဗိုက် နာ တယ် ။
HYP: ကျွန်တော် မျက်စိ နာ တယ် ။
Eval: S

10 Conclusion

This paper has presented the first study of the neural machine translation between Myanmar sign language and Myanmar written text. We implemented three NMT systems (RNN, Transformer and CNN) with our developing MSL-MWT written text corpus. We also investigated the effectiveness of four word segmentation schemes (word segmentation, syllable segmentation, SentencePiece and Byte-Pair-Encoding) for NMT. Our results clearly show that the highest quality NMT and SMT performances were attained with syllable segmentation for both MSL and MWT. We found that Transformer outperformed both CNN and RNN for MWT-to-MSL and MSL-to-MWT translation tasks.

We plan to extend our study on NMT approaches with our MSL corpus data to explore the appropriate hyper-parameters such as the number of hidden layers and initial learning rate, etc. Furthermore, we also planning our research work for MT between MSL video and MWT in the near future.

Acknowledgment

We would like to thank principals, teachers, SL trainers and students of School for the Deaf (Mandalay), Mary Chapman Chapman School for the Deaf Children (Yangon), School for the Deaf, Tamwe (Yangon), Myanmar Deaf Society and Literacy and Language Development for the Deaf for their kind contribution to our research.

References

[1] The population and housing census of Myanmar, 2014

[2] Hutchins, W. J., "Early years in machine translation", John Benjamins Publishing, 2000, USA doi: 10.1075/sihols.97

[3] Hutchins, W. J., & Somers, H. L., "An introduction to machine translation", Academic Press, 1992, London, ISBN-13: 978-0123628305

[4] Nirenburg, S., & Raskin, V., "Ontological semantics", The MIT Press, 2004 ISBN: 9780262140867

[5] Bangham, J. A., & Cox, S. J., "Signing for the deaf using virtual humans", In Proceeding of the Speech and Language Processing for Disabled and Elderly People (Ref. No. 2000/025), IEE Seminar, 2000, London, UK

[6] Angus B., & Smith, G., "English to American sign language machine translation of weather reports", In Proceeding of the 2nd high desert student conference in linguistics (HDSL2), 1999, Albuquerque, New Mexico, pp. 23–30

[7] Safar, E., & Marshall, I., "The architecture of an English-text-to-sign-language translation system", In Angelova, G. (Ed.), Recent advances in natural language processing (RANLP), 2000, Tzigov Chark, Bulgaria, pp. 223–228

[8] Zhao, L., & Kipper, K., "A machine translation system from English to American sign language", In Proceeding of the 4th conference of the association for machine translation in the americas on envisioning machine translation in the information future, 2000, Springer-Verlag, pp. 54– 67

[9] Veale, T., & Collins, B., "The Challenges of Cross-modal Translation: English to sign language translation in the ZARDOZ system", Machine Translation, 13, 1998, pp. 81–106.

[10] Armond, D., & Speers, L., "Representation of American sign language for machine translation", Ph.D. Thesis, 2001, Department of linguistics, Georgetown University.

[11] Zijl, L. V., & Barker, D., "South African sign language machine translation system", In Proceeding of the 2nd international conference on computer graphics, virtual reality, visualisation and interaction in Africa (ACM SIGGRAPH), 2003, Cape Town, South Africa, pp. 49–52

[12] Suszczanska, N., & Szmal, P., "Translating Polish text into sign language in the TGT system", In Proceeding of the 20th IASTED international multi-conference applied informatics AI, 2002, Innsbruck, Austria, pp. 282–287

[13] Huenerfauth, M. (2004a), "A multi-path architecture for machine translation of English text into American sign language animation", In Proceeding of the student workshop at the human language technology conference/North American chapter of the association for computational linguistics annual meeting (HLT-NAACL), May 02 - 07, 2004, Boston, MA, USA, pp. 25-30

[14] Huenerfauth, M. (2004b), "Spatial and planning models of ASL classifier predicates for machine translation", In Proceeding of the 10th international conference on theoretical and methodological issues in machine translation (TMI 2004), Baltimore, MD, USA.

[15] Huenerfauth, M. (2004c), "Spatial representation of classifier predicates for machine translation into American Sign Language", In Proceeding of the workshop on the representation and processing of signed languages, 4th international conference on language resources and evaluation (LREC 2004), Lisbon, Portugal.

[16] Huenerfauth, M. (2005a), "American Sign Language generation: Multimodal NLG with multiple linguistic channels", In Proceeding of the student research workshop, the 43rd annual meeting of the association for computational linguistics, Ann Arbor, MI, USA.

[17] Huenerfauth, M. (2005b), "American Sign Language, natural language generation and machine translation", ACM SIGACCESS Accessibility and Computing (Vol. 81). New York: ACM Press.

[18] Huenerfauth, M. (2005c), "American sign language spatial representations for an accessible user-interface", In Proceeding of the 3rd international conference on universal access in human-computer interaction, Las Vegas, NV, USA.

[19] Huenerfauth, M. (2005d), "Representing coordination and non-coordination in an American sign language animation", In Proceeding of the 7th international ACM SIGACCESS conference on computers and accessibility (ASSETS 2005), Baltimore, MD, USA

[20] Morrissey, S. & Way, A., "Experiments in sign language machine translation using examples", In Proceeding of the IBM CASCON 2006 Dublin symposium, Dublin, Ireland.

[21] Stein, D., Bungeroth, J., & Ney, H., "Morpho-syntax based statistical methods for sign language translation", In Proceeding of the 11th annual conference of the European association for machine translation, June, 2006, Oslo, Norway.

[22] Wikipedia of Fingerspelling: https://en.wikipedia.org/wiki/Fingerspelling

[23] Boudreault, Patrick; Mayberry, Rachel I., "Grammatical processing in American Sign Language:Age of first-language acquisition effects in relation to syntactic structure". Language and Cognitive Processes, Volume 21, 2006 – Issue 5, pages 608-635, https://doi.org/10.1080/01690960500139363

[24] Fenlon, Jordan; Denmark, Tanya; Campbell, Ruth; Woll, Bencie, "Seeing sentence boundaries", Sign Language & Linguistics, 2008, 10 (2), pp. 177 – 200. http://dx.doi.org/10.1075/sll.10.2.06fen

[25] Thompson, RobYin; Emmorey, Karen; Kluender, Robert, "The Relationship between Eye Gaze and Verb Agreement in American Sign Language: An Eye-tracking Study". Natural Language & Linguistic Theory. 24 (2), 2006, pp. 571–604 doi:10.1007/s11049-005-1829-y.

[26] Baker, Charlotte, and Dennis Cokely, "American Sign Language: A teacher's resource text on grammar and culture", 1980, Silver Spring, MD: T.J. Publishers.

[27] Sutton-Spence, Rachel, and Bencie Woll, "The linguistics of British Sign Language", Cambridge: Cambridge University Press, 1998

[28] "Myanmar Sign Language Basic Conversation Book", Ministry of Social Welfare, Relief and Resettlement, Department of Social Welfare, Japan International Cooperation Agency, 1st Edition, August 2009, Daw Yu Yu Swe, Department of Social Welfare.

[29] myPOS (Myanmar Part-of-Speech Corpus): https://github.com/ye-kyaw-thu/myPOS

[30] Win Pa Pa, Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita, "Word Boundary Identification for Myanmar Text Using Conditional Random Fields", In Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing, August 26-28, 2015, Yangon, Myanmar, pp. 447-456.

[31] Ye Kyaw Thu, Andrew Finch, Yoshinori Sagisaka and Eiichiro Sumita, "A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation", In Proceedings of the 11th International Conference on Computer Applications (ICCA 2013), February 26 27, 2013, Yangon, Myanmar, pp. 167-179.

[32] Myanmar Unicode Table, Range:1000–109F, http://www.unicode.org/charts/PDF/U1000.pdf

[33] Zin Maung Maung and Yoshiki Makami. "A rule-based syllable segmentation of Myanmar Text", In Proceedings of the IJCNLP-08 workshop of NLP for Less Privileged Language, January, 2008, Hyderabad, India, pp. 51-58.

[34] Tin Htay Hlaing, "Manually constructed context-free grammar for Myanmar syllable structure", In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12), Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 32-37.

[35] Syllable Segmentation Tool for Myanmar Language: https://github.com/ye-kyaw-thu/sylbreak

[36] Taku Kudo and John Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing", EMNLP2018.

[37] Rico Sennrich, Barry Haddow and Alexandra Birch, "Neural Machine Translation of Rare Words with Subword Units", In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany.

[38] Philip Gage, "A New Algorithm for Data Compression", C Users J., 12(2):23–38, February 1994.

[39] Kohen, P., Och, F .J., Marcu, D., "Statistsical phrase-based translation", In HLT-NAACL, 2003, url: http://acl.ldc.upenn.edu/N/N03/N03-1017.pdf

[40] Och, F .j., Marcu, D., "Statistical phrase-based translation", 2003, p.127-133.

[41] Specia, L.. Tutorial, fundamental and new approaches to statistical machine translation, In: International Conference Recent Advances in Natural Language Processing, 2011

[42] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies", In A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press, 2001.

[43] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "Understanding the exploding gradient problem". CoRR, abs/1211.5063, 2012.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition". In CVPR, 2016.

[45] Thang Luong, Hieu Pham, and Christopher D. Manning, "Effective approaches to attention-based neural machine translation". In EMNLP, 2015.

[46] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate". CoRR, abs/1409.0473, 2014.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need". CoRR, abs/1706.03762, 2017.

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

[49] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin, "Convolutional sequence to sequence learning". CoRR, abs/1705.03122, 2017.

[50] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks". CoRR, abs/1612.08083, 2016.

[51] Kohen, P., Haddow, B.. "Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses", In Proceedings of the 4th Workshop on Statistical Machine Translation. 2009, pp. 160-164.

[52] Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. "Sockeye: A Toolkit for Neural Machine Translation". ArXiv e-prints, December 2017.

[53] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems". CoRR, abs/1512.01274, 2015.

[54] Holger Schwenk, "Continuous space translation models for phrase-based statistical machine translation", In COLING, 2012.

[55] Nal Kalchbrenner and Phil Blunsom, “Recurrent continuous translation models”, In EMNLP, 2013.

[56] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, “Sequence to sequence learning with neural networks”, In NIPS, 2014.

[57] Robbins, Herbert and Sutton Monro, “A stochastic approximation method”, The annals of mathematical statistics, pages 400–407, 1951.

[58] Duchi, John C., Elad Hazan, and Yoram Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization”, Journal of Machine Learning Research, 12:2121–2159, 2011.

[59] Kingma, Diederik P. and Jimmy Ba, “Adam: A Method for Stochastic Optimization”, CoRR, abs/1412.6980, 2014.

[60] Papineni, K., Roukos, S., Ward, T., Zhu, W., “Bleu: a Method for Automatic Evaluation of Machine Translation”. IBM Research Report rc22176 (w0109022), 2001, Thomas J. Watson Research Center, In ACL ’02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, July 07 - 12, 2002, Philadelphia, Pennsylvania, pp. 311-318

[61] Och, F.J., Ney, H.. “Improved statistical alignment model”, In ACL00. Hong Kong, China, 2000, pp. 440-447

[62] Koehn, P., Och, F.J.., Marcu, D.., “Statistical phrase-based translation”, In Proceedings of the Human Language Technology Conference, 2003, Edmonton, Canada, pp. 48-54

[63] Tillmann, C.., “A unigram orientation model for statistical machine translation”, In Proceedings of HLT-NAACL 2004: Short Papers; HLT-NAACL-Short’04. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 1-932432-24-8; 2004, pp.101-104, http://d1.acm.org/citation.cfm?id-1613984.1614010.

[64] Heafield, Kenneth, “KenLM: Faster and Smaller Language Model Queries”, Proceedings of the 6^{th} Workshop on Statistical Machine Translation; WMT ’11, 2011, Association for Computational Linguistics, Edinburgh, Scotland, pp. 187-197 ISBN- 978-1-937284-12-1

[65] Chen, S.F., Goodman, J., “An empirical study of smoothing techniques for language modeling”, In Proceedings of the 34^{th} annual meeting on Association for Computational Linguistics. Association for Computational Linguistics; 1996, pp. 310-318.

[66] Och, F.J., “Minimum error rate training for statistical machine translation”, In Proceedings of the 41^{st} Meeting of the Association for Computational Linguistics (ACL 2003). Sapporo, Japan