Syllable-based Myanmar-English Neural Machine Translation

Yi Mon Shwe Sin, Khin Mar Soe
Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
yimonshwesin@ucsy.edu.mm , khinmarsoe@ucsy.edu.mm

Abstract

The paper presents the first large scale evaluation of the quality of Syllable-based Neural Machine Translation (Syllable-NMT) system for Myanmar-English pair. Neural Machine Translation (NMT) system has reached state-of-the-arts results on some languages. However, one of the main challenges that NMT still faces is dealing with very large vocabularies and morphologically rich languages. Like other low-resources languages, Myanmar Language has a lots of morphology information. This issue lead is to increase the ambiguity and to decrease the quality of translation results. Moreover, rule-based and phrase-based techniques were used in the existing research on Myanmar translation with the small amount of parallel corpus. Therefore, a large amount of parallel corpus is prepared and introduces a NMT model that maps a source syllable sequence to a target word sequences to address the morphological problems. In addition, this paper shows some experiments results and compare them. Our results show that syllable-NMT system is able to surpass than the character-based and word-based NMT systems by 5 BLEU.

1. Introduction

Neural machine translation (NMT) is a new methodology for machine translation that has led to remarkable improvements, particularly in terms of human evaluation, compared to rule-based and statistical machine translation (SMT) systems [8]. For the last 20 years, statistical phrase-based machine translation is one of the most popular approaches. Just recently, the neural MT approach has appeared and obtained state-of-the-art results [9]. Neural Machine Translation is an end-to-end architecture to translate by substitution of words or characters or any other in one language to another language. Neural Machine Translation improved the data sparsity problem of traditional SMT models. This is the issue to use the neural networks. Currently, Neural Machine Translation (NMT) is a very popular research title.

Nearly all previous work in machine translation systems use character level or word level as their unit of input and output. Although the usage of word level in some languages reaches into successful state, word-level neural machine translation cannot handle rare words in translating languages with rich morphology such as Myanmar, Czech, Finnish and Turkish. To address this morphological problem, some papers present a character-level NMT model. Character-level models are better suited for multilingual translation than their word-level counterparts which require a separate word vocabulary for each language. Character level models can handle rare morphological variants of a word, and do not require word segmentations.

In Myanmar-English Machine Translation, the previous research mostly learnt by using rule-based as well as statistically based approach. However, there are a little research papers using statistic methods. There is no research paper using neural information. Nevertheless, the previous researches on machine translation only used the small amount of parallel corpus. Actually, the quality of machine translation depends on the size of training data. And Myanmar language is an under-resourced language (known as low-resource languages). So, there are some difficulties in preparation of parallel corpus. Hence, this paper concentrates on building a huge amount of Myanmar-English parallel corpus to provide NLP tasks, especially Machine Translation.

Like other low-resources languages, Myanmar Language is one of the morphology rich languages. This issue lead is to increase the ambiguity and pose a challenge for Myanmar to English machine translation. Unlike the other languages, one Myanmar sentences has one or more Myanmar Words. One word has one or more syllables. And one Syllable has one or more characters. So, this paper introduces a Syllable-based Neural Machine Translation system for Myanmar to English pairs. Syllables-NMT maps a source syllable sequence to a target word sequences to address the morphological problems. And, this paper shows many attempts to develop the machine translation with higher performance using word-based, character-based and syllable-based neural method and compared them. Our results show that syllable-NMT system is able to surpass than the character-based and word-based NMT systems by 5 BLEU.

In this paper, we mainly focus on the Myanmar to English Syllable-based Neural Machine Translation system. This paper is organized as follow. Section 2 describes the related works of the areas of machine translation on both local and international papers. Section 3 presents about the Myanmar language and its formation. Section 4 describes the theory background of Neural Machine Translation. Section 5 presents the detailed experiments settings and section 6 concludes the paper.

2. Related Work

Many researchers have been worked for Neural Machine Translation system in other languages. However, there is no one for implementing neural machine translation system based on syllable for Myanmar language. This section presents some of the related works in the area of Machine Translation systems on Myanmar language and other languages.

T.T. Zin et al. [2] presented about Myanmar phrases translation model with morphology analysis based on statistical approach. Myanmar language is inflected language and there are very few creations and researches of corpora in Myanmar, comparing to other language such as English, French, and Czech etc. Therefore, Myanmar phrases translation model is based on syntactic structure and morphology of Myanmar language. In this paper, Bayes rule is also used to reformulate the translation probability of phrase pairs. Experiment results showed that proposed system can improve translation quality by applying morphological analysis on Myanmar language.

Y.Y. Win and T.H. Nwe [5] presents English-Myanmar bidirectional text to text machine translation based on rule-based machine translation (RBMT) approach. The system applies tree to tree transformation for rule-based translation using Synchronous Context Free Grammar (SCFG) rules to change source sentence structure to target sentence structure. W.P. Pa et al. [4] studied on phrase-based, hierarchical phrase-based, the operational sequence model, sting-to-tree, tree-to-string statistical machine translation methods between English (en) and the under resourced languages Lao, Myanmar, Thai in both directions. The ASEAN-MT parallel corpus [19] without name entity was used and contained 20000 sentences for training, 500 sentences for development and 300 sentences for evaluation. This paper shows that the phrase-based SMT gave the highest BLEU scores than the other methods.

Y. Belinkov and J. Glass [12] presents Machine translation between Arabic and Hebrew. This paper compares standard phrase-based and neural systems on Arabic-Hebrew translation. And external tools are used for tokenization and experiments on subword modeling by character-level neural models. This paper shows that lead to improved translation performance, with a small advantage to the neural models.

J. Lee et al. [13] employs a character-to-character model outperforms a recently proposed baseline with a subword-level encoder on WMT's15 DE-EN and CS-En. This shows the advantage of character-level models that are better suited for multilingual translation than their word-level counterparts which require a separate word vocabulary for each language.

In our work, syllable-based neural machine translation system is proposed to addresses the morphological problems and compares the performance of word-based NMT, char-based NMT and syllable-NMT for Myanmar to English pairs.

3. Overview of Myanmar Language

Myanmar language is the official language and native language of Union of Myanmar. Myanmar is recognized as one of the Tibeto-Burman group. About 34 million people speak as their first language and almost all the educated people speak as second language if their mother tongue is another ethnic language [18]. Myanmar language has (33) Consonants, Independent vowels, Dependent consonant signs (also known as Medials), Dependent vowels signs, Dependent various signs (also known as Parli Word), punctuation and digits. They can be seen in figure 1.

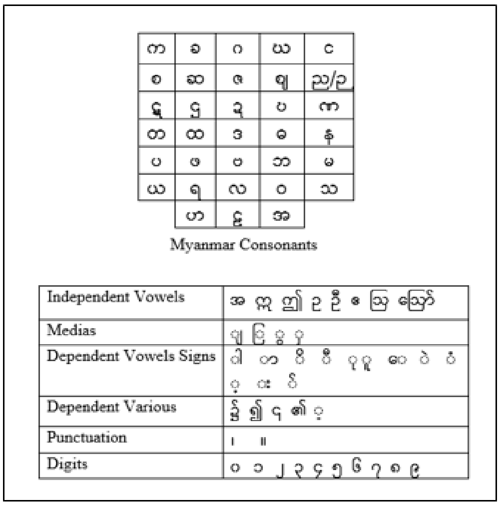

Figure 1: Myanmar Character Pattern

Myanmar language is written from left to right. There are two language styles: Spoken style and Written style. And the structure of the sentences is subject-object-verb (SOV). A word consists of one or more syllables. And a syllable is composed of one or more character such as an initial consonant followed by zero or more medials, zero or more vowels. The formation of Myanmar Word is shown in figure 2. Myanmar language likes other Southeast Asia languages that do not place spaces between words. It is usually written continuously without using space. Sometimes, it is written spaces between phrases. However, there is no rule how to write definitely in Myanmar language.

One Word	ကျန်းမာပါစေ။
Five Syllables	ကျန်း မာ ပါ စေ ။
Consonants	က န မ ပ စ
Medias	ျ
Dependent Vowels Signs	ိ း ာ ါ ေ
Punctuations	။

Figure 2: The formation of Myanmar Word. It is one word and five syllables. This word is formed with consonants, medias, dependent vowel signs and punctuations.

3. Neural Machine Translation System

Neural machine translation refers to a recently proposed approach to machine translation [9]. This approach aims at building an end-to-end neural network that takes as input a source sentence $X = (x_1, \ldots, x_{Tx})$ and outputs its translation $Y = (y_1, \ldots, y_{Ty})$, where x_t and $y_{t'}$ are respectively source and target symbols. This neural network is constructed as a composite of an encoder network and a decoder network.

The encoder network encodes the input sentence X into its continuous representation. This paper closely follows the neural translation model proposed in [9] and use two recurrent neural networks. The forward network reads the input sentence in a forward direction:

$$\overrightarrow{z_t} = \vec{\emptyset}\,(e_x(x_t), \overrightarrow{z_{t-1}})$$

where $e_x(x_t)$ is a continuous embedding of the t-th input symbol and Ø is a recurrent activation function. Similarly, the reverse network reads the sentence in a reverse direction (right to left):

$$\overleftarrow{z_t} = \overleftarrow{\emptyset}(e_x(x_t), \overleftarrow{z_{t+1}})$$

At each location in the input sentence, we concatenate the hidden states from the forward and reverse RNNs to form a context set $C = \{z_1, \ldots, z_{Tx}\}$, where

$$z_t = [\overrightarrow{z_t}\,;\, \overleftarrow{z_t}]$$

Then the decoder computes the conditional distribution over all possible translations based on this context set. This is done by first rewriting the conditional probability of a translation:

$$\log p(Y|X) = \sum_{t'=1}^{T_y} \log p(y_{t'}|y_{<t'}, X)$$

For each conditional term in the summation, the decoder RNN updates its hidden state by

$$h_{t'} = \emptyset(e_y(y_{t'-1}), h_{t'-1}, c_{t'}), \qquad (1)$$

where e_y is the continuous embedding of a target symbol. $c_{t'}$ is a context vector computed by a soft alignment mechanism:

$$c_{t'} = f_{align}(e_y(y_{t'-1}), h_{t'-1}, C), \qquad (2)$$

The soft-alignment mechanism f_{align} weights each vector in the context set C according to its relevance given what has been translated. The weight of each vector z_t is computed by

$$\alpha_{t,t'} = \frac{1}{Z} e^{f_{score}(e_y(y_{t'-1}), h_{t'-1}, z_t)} \qquad (3)$$

where f_{score} is a parametric function returning an unnormalized score for z_t given $h_{t'-1}$ and $y_{t'-1}$. We use a feed-forward network with a single hidden layer in this paper. Z is a normalization constant:

$$Z = \sum_{k=1}^{T_x} e^{f_{score}\left(e_y(y_{t-1}), h_{t-1}, z_k\right)}$$

This procedure can be understood as computing the alignment probability between the *t'-th* target symbol and *t-th* source symbol.

The hidden state $h_{t'}$, together with the previous target symbol $y_{t'-1}$ and the context vector $c_{t'}$, is fed into a feed-forward neural network to result in the conditional distribution:

$$p(y_t|y_{<t}, X)\alpha\ e^{f_{out}^{y_t}(e_y(y_{t-1}), h_t, c_t)} \qquad (4)$$

The whole model, consisting of the encoder, decoder and soft-alignment mechanism, is then tuned end-to-end to minimize the negative log-likelihood using stochastic gradient descent.

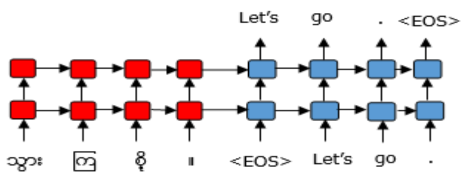

Figure 3: Encoder-Decoder Architecture of Myanmar to English Syllable-NMT

Syllable-based Neural Machine Translation system takes a sequence of syllable of Myanmar source sentences as input and a sequence of word of English target sentences as output. The system is implemented using the Pytorch OpenNMT. Pytorch is a popular OpenNMT tool that is designed to be research friendly in translation, summary, image-to-text, morphology, and many other domains. OpenNMT-py is run as a collaborative open-source project. It is currently maintained by Sasha Rush (Cambridge, MA), Ben Peters (Saarbrücken), and Jianyu Zhan (Shenzhen). The original code was written by Adam Lerer (NYC). The architecture of Syllable-NMT model is shown in figure 3.

4. Experimental Settings

This section describes the dataset and preprocessing, the training and evaluation results and all experimental results.

4.1. Dataset and Preprocessing

Myanmar language is one of the low resource Asian Languages. So, there are some difficulties to collect the parallel sentences. There are a few sources that include bilingual sentences. Therefore, the bilingual sentences from local news website, Wikipedia and text book were collected. Some bilingual sentences are copied from the website and some are copied from the ebooks manually. These bilingual sentences contain about 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) and news. There are nearly 230K parallel sentences.

Table 1. Total Parallel Sentences

Domain	Parallel Sentences
News	100000
Text Books	125600
Total	225600

In preprocessing step, we segment the English sentences into word level using a tokenization script included in Moses. To segment the Myanmar sentences as word level, we use UCSY word segmenter (http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html). For syllable level, we use the syllable segmenter (https://github.com/ye-kyaw-thu/sylbreak). The total parallel sentences are shown in Table 1.

4.2. Training and Evaluation Results

We tested three models: Word-Based Neural Machine Translation, Character-Based Neural Machine Translation and Syllable-based Neural Machine Translation. These models are implemented using a pytorch OpenNMT [22]. In the experiments, we use similar parameters for all experiments: deep LSTM models, 2 layers, 500 hidden units and 500-dimensional embeddings on both the encoder/ decoder. The models are trained on a single GPU using SGD with 13-epochs and learning rate of 1.0. Batches are size of 64 and dropout is 0.3. Decoding is done with beam search using a width of 5. The vocabulary size of each model is shown in Table 2.

Table 2. Vocabulary size of the experiments

Model	Source vocabulary size	Target vocabulary size
Word-NMT	30000	56000
Character-NMT	300	56000
Syllable-NMT	6600	56000

Translation quality was measured BLEU (BiLingual Evaluation Understudy) which is a script from Moses. BLEU is inexpensive and easy to calculate, correlates well with human evaluation, and is language independent. It is among the most widely-used automated methods of determining machine translation quality.

Table 3. BLEU of the experiments

Experiments	BLEU
Word-NMT	21.88
Character-NMT	20.71
Syllable-NMT	26.50

Table 4. Examples of sentences translation

Source	လယ် သ မား များ သည် အ စိုး ရ ဆီ က ချေး ငွေ များ လက် ခံ ရ ရှိ သည် ။
Reference	Farmers receive the loans from the government.
CharacterNMT	Farmers have to accept the government .
WordNMT	Farmers receive loans from government .
SyllableNMT	Farmers received loans from the government .
Source	ဆ ရာ မ သည် သင် ပုန်း ကို ညွှန် ပြ နေ သည် ။
Reference	The teacher is pointing to the blackboard.
CharacterNMT	Teacher was pointing at the blackboard.
WordNMT	The teacher pointed to the dhamma.
SyllableNMT	The teacher is pointing to the blackboard.
Source	ဇန် န ဝါ ရီ လ (၄) ရက် နေ့ သည် မြန် မာ နိုင် ငံ ၏ လွတ် လပ် ရေး နေ့ ဖြစ် သည် ။
Reference	4th of January is Myanmar independent day.
CharacterNMT	January is independence day.
WordNMT	On 4 January is the independence day of Myanmar.
SyllableNMT	January 4 is the independence day of Myanmar.
Source	အ ရပ် ဘက် အ ဖွဲ့ အ စည်း တွေ အ နေ နဲ့ သူ တို့ ရဲ့ ကိုယ် ပိုင် အ ချက် အ လက် တွေ ကို ကာ ကွယ် ဖို့ နည်း လမ်း တွေ ရှာ ဖွေ သင့် ပါ တယ် ။
Reference	Civil society should now be exploring ways to protect their own data.
CharacterNMT	They should find rid of their own organization.
WordNMT	Civil society organisations should find ways to protect their own facts.
SyllableNMT	The civil society organisations should consider ways to protect their own data.

BLEU score of the experiments is shown in Table 3. It can be seen that only Syllable-NMT is better than the Word-NMT and Character-NMT. Mostly, Character-NMT and Word-NMT can translate only short sentences. There are many errors in long sentences and cannot translate it well. And they don't know the usage of verb and sentence reordering well. However, Syllable-NMT can translate not only short sentences but also long sentences. Besides, it knows the usage of verb and sentence reordering and can translate well. Table 4 shows some examples of the sentences translation that are translated by each system. As shown in example, the performance of Syllable-NMT improves up to 5 BLEU over other level NMT for Myanmar-English pair.

5. Conclusion

This paper proposes a syllable-based NMT model that accepts a sequence of syllables in the source sentences and outputs a sequence of words in the target language. And we tested the three experiments with a number of large-scale Myanmar to English parallel sentences and compared them. Our NMT systems are implemented using the Pytorch OpenNMT. Among them, Syllable-based NMT is better than the other level such as words or characters for the low resource languages like Myanmar language. Besides, we investigate the Syllable-based Myanmar to English Neural Machine Translation addresses the morphological problems for low resource languages like Myanmar.

Acknowledgments

This work is partially supported by Institute of Infocomm research (I^2R), Singapore. We are thankful to Aw Ai Ti and Wu Kui, Department of Human Language Technology, I^2R, who provided expertise that greatly assisted for my research. We also thank the reviewers for their valuable comments and suggestions.

References

[1] M.T. Win, M.M. Win and M.M. Than, "Burmese Phrase Segmentation", Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

[2] T.T. Zin, K.M. Soe and N.L. Thein, "Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System", 25th Pacific Asia Conference on Language, Information and Computation, pages 130–139, 2011.

[3] Y.K. Thu, A. Finch, Y. Sagisaka, and E. Sumita, "A study of Myanmar word segmentation

schemes for statistical machine translation", Proceeding of the 11th International Conference on Computer Applications, 2013, pages 167–179.

[4] W.P. Pa, Y.K. Thu, A. Finch and E. Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages", 5th Workshop on Spoken Language Technology for Under-resources Languages, SLTU 2016 Yogyakarta, Indonesia, 9-12 May 2016, pages 250-257.

[5] Y.Y. Win and T.H. Nwe, "Myanmar-English Bidirectional Machine Translation system by using Transfer Based Approach", 13th International Conference on Computer Applications, ICCA (2015).

[6] K.T. Nwet and K.M. Soe, "Myanmar - English Machine Translation Model", International Conference on Genetic and Evolutionary Computing (ICGEC): Genetic and Evolutionary Computing, 2016, pages 195-203.

[7] M. Luong, I. Sutskever, Q.V. Le, O. Vinyals and W. Zaremba, "Addressing the Rare Word Problem in Neural Machine Translation", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 26-31, 2015, pages 11–19.

[8] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation", Technical report, Google. https://arxiv.org/abs/1609.08144, 2016.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", Published as a conference paper at ICLR 2015.

[10] M. Luong and C. D. Manning "Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1054–1063, 2016.

[11] M. R. Costa-Jussà and J. A. R. Fonollosa, "Character-based Neural Machine Translation", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 357–361, 2016.

[12] Y. Belinkov and J. Glass, "Large -Scale Machine Translation between Arabic and Hebrew: Available Corpora and Initial Results", Proceedings of the Workshop on Semitic Machine Translation, pages7-12, 2016.

[13 J. Lee, K. Cho and T. Hofmann, "Fully Character-Level Neural Machine Translation without Explicit Segmentation", Transactions of the Association for Computational Linguistics, vol. 5, pp. 365–378, 2017.

[14] G. Klein, Y. Kim, Y. Deng, J. Senellart and A. M. Rush, 2017, "OpenNMT: Open-Source Toolkit for Neural Machine Translation",

[15] Department of the Myanmar Language Commission, Ministry of Education Yangon, 1998. Myanmar-English Dictionary.

[16] Department of the Myanamar Langue Commission, Ministry of Education, Union of Myanmar. 2005. Myanmar Grammar.

[17]https://www.britannica.com/topic/Burmese-language

[18] https://en.wikipedia.org/wiki/Burmese_language

[19] http://www.aseanmt.org/index.php

[20]http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html

[21] https://github.com/ye-kyaw-thu/sylbreak

[22] https://github.com/OpenNMT/OpenNMT-py