ISBN: 978-981-18-1791-5
2021 the 11th International Workshop on Computer Science and Engineering (WCSE 2021)
doi: 10.18178/wcse.2021.02.012

Neural Named Entity Transliteration for Myanmar to English Language Pair

Aye Myat Mon [1+], Khin Mar Soe [2]

[1, 2] Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar

Abstract. Named entity (NE) transliteration is mainly a phonetically based transcription of names across languages using different writing systems. This is a crucial task for various downstream natural language processing applications, such as information retrieval, machine translation, automatic speech recognition and so on. Robust transliteration of named entities is still a challenging task for Myanmar language because of the complex writing system and the lack of data. In this paper, we proposed our Myanmar-English named entity terminology dictionary and experimented on transformer-based neural network model. Furthermore, we evaluated the performance of neural network-based approach on the transliteration tasks using BLEU score. Different units in the Myanmar script, i.e., character units, sub-syllable units and syllables units are compared in the experiments.

Keywords: myanmar language, transformer, neural network, named entity, transliteration

1. Introduction

According to the linguistic point of view, transliteration is the tasks of representing words from source language script using the approximate phonetic or spelling equivalents of target language script. Meanwhile, the quality of machine translation has improved significantly but there are still many problems emerging to be solved to emend machine transliteration. Precise transliteration of named entities plays a very significant role in improving the quality of machine translation and cross language information retrieval and their attainment depends extremely on accurate transliteration of named entities.

In this study, the tasks of Myanmar named entity Transliteration is indicated from initial raw transliteration instances collection to manual annotation and final experiments. Like Myanmar, one major obstacle of low resource languages is the problem of out-of-vocabulary (OOV) words. Thus, our in-house Myanmar-English bilingual named entity terminology dictionary is contrived to promote Myanmar natural language processing research areas. Our experiments aim to compare Myanmar (My)-English (En) transliteration directions with character units (Char), sub-syllable units (Sub-Syl) and syllable units (Syl) on Myanmar side and standard character units on English side using Openmt open source toolkit for transformer model. This approach performs well on cross lingual transliteration tasks. To the best of our knowledge, we believe that our work will be the first attempt in this direction.

The paper's structure is organized as follows. In session 2, we discuss the related work and in session 3, we present the nature and collation of Myanmar language. The issues of transliteration and construction of My-En terminology dictionary are described in session 4 and 5. We show the experiments in session 6 and conclude the final in session7.

2. Related Work

Although prior research made to improve the transliteration process based on many languages such as English, Chinese, Korean, Japan and Thailand etc., it still need to accomplish for Myanmar language due to lack of efficient resources. According to surveys, there are a few researches for Myanmar Language. Transliteration process is similar to Romanization or Transcription process for rendering Myanmar Latin

[+] Corresponding author. Tel.: +95-9-750989163; fax: +95-1-610633.
E-mail address: ayemyatmon.ptn@ucsy.edu.mm.

alphabet which can cast as a clarified translation process on grapheme level or phrase level without reordering the operations.

In prior work [1], the authors performed Myanmar Name Romanization task by using sub-syllable and syllable units based on small amount of training data. Although the system gets the efficient results in statistical way, it still has some necessities because LSTM network require more training data. In the proposed system, we have prepared enough training data to apply neural network approaches by tuning the different hyper parameters to improve the performance of transliteration task. In reference [2], the authors applied grapheme to phoneme (G2P) conversion on Myanmar Language. This system is mainly emphasized for speech recognition rather than NLP. The transformer model was presented in [3] which rely on only self-attention by avoiding the needs for sequential processing. Unlike the encoder-decoder architecture, there is no information bottleneck in hidden state vectors in transformer model.

3. Nature and Collation of Myanmar Language

The Myanmar alphabet (မြန်မာ အက္ခရာ) is an abugida and their original calligraphy was a square format but currently changed to rounded format. It is the first language of native Myanmar people and also the official language of the Republic of the Union of Myanmar. Basically, the spelling of Myanmar script is syllable-based. One word belongs to multiple syllables and one syllable by multiple characters. In between the characters and symbols, sub-syllable units may be designed for specific task [1]. Figure 1 exposes the syllable structure defined by Myanmar script. The initial consonant (C) on the left is obliged; more characters for consonant clusters, alternated vowels, syllable codas and tones can be added gradually.

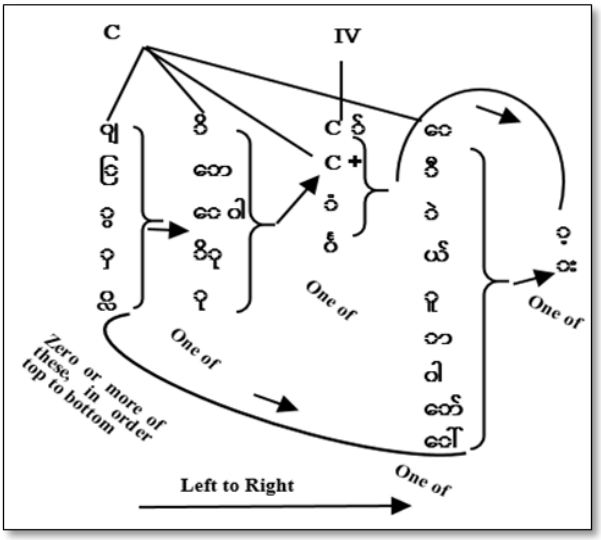

Fig. 1: The Collation of Myanmar Syllable

4. Transliteration Issues

Continuous growth of out of vocabulary names (loan words) to be transliterated, there are no systematic rules in Myanmar language. Myanmar loan word is overwhelmingly in the form of proper nouns (i.e., person names, place names and organization names). As a natural consequence of British rule in Myanmar, English has been another major source of vocabulary, especially about technology, measurements and modern institutions. In major case of Myanmar name transliteration, the adoption of an English word, adapted to the Myanmar phonology, known as direct loan. For example, “Confucius” (Chinese philosopher) is “ကွန်ဖြူးရှပ်” and “Sidney” (City in Australia) is “ဆစ်ဒနီ” in Myanmar terms respectively. There are many issues on transliteration because Myanmar writing and pronunciation have some conflicts between native and foreign words. In some writing script of Myanmar alphabets such as “ဆ”, “စ” and “သ” are pronounced into “Sa”, “S” and “Tha”. The “သ” is not clear today because “သ” is usually pronounced as “Tha”. Even though this pronunciation has been accepted in other nationalities, i.e., other ethnic group like ‘Mon’, “သ” is pronounced as “S”. For instance, other Myanmar vowels like “အော”, “သြ”, “သြော” and “သြော်”, are written as “Aw” which are pronounced as “O”. Anyway, “သီရိ-လင်္ကာ” (Democratic Socialist Republic of Sri- Lanka)

is written as “Sri-Lanka” but pronounced as “Thiri-Linka”. Generally, Transliterating names is an easy way to pronunciation as much as accurate representation for native speakers, but the above confusions can cause a problem of spelling difficulty and effect on the accuracy of transliteration tasks.

5. Myanmar–English Named Entity Dictionary Construction

One of the main reasons is the lack of resources such as annotated-corpus, gazetteers or name-mapping dictionary and name-lists etc. That is to say, Myanmar language is resource-constrained language. As a matter of this fact, My-En bilingual named entity terminology dictionary is proposed to coverage these problems.

We used University of Computer Studies (UCSY) corpus [4] and Asian Language Treebank (ALT) corpus [4] in constructing Myanmar-English bilingual NE dictionary. All of the sentences in these corpora are normalized and tokenized for both Myanmar and English languages. The UCSY corpus comprises 200K Myanmar-English pairs of parallel sentences which are collected from textbooks and local news articles [5] developed by NLP lab, UCSY, Myanmar. ALT corpus is one of the segments of ALT project launched by ASEAN IVO. It is composed of 20K Myanmar-English pairs of parallel sentences from Wikipedia news.

In construction the dictionary, we utilized GIZA++ open source toolkit [6] to get the raw alignment for source and target language. To filter the transliteration sentence pairs, we have manually annotated the transliteration Myanmar term of public figures, places, well-known person and organizations for each English named entity in this aligned coarse sentence pairs.

We performed the experiments based on bilingual dataset which contains 84,057 named entity instance pairs. We first divided these instance pairs into two types: parallel data and monolingual data. We then subdivided these data into three parts for training, development and testing purpose. The 1K of dataset is made as testing, 1K as development whiles the rest of the dataset as training. All of the collected named entities are standardized with Unicode format.

Table 1: Data statistics on Train, Dev and Test

Data	My			En
	# Char	# Sub-Syl	# Syl	# Char
Train	2,120,773	1,824,597	1,433,745	1,884,108
Dev	15,800	14,476	10,314	13,604
Test	16,288	14,221	11,201	14,253

6. Experiments

6.1. Data Preprocessing

Myanmar is a complicated language. So, it needs to be precise data pre-processing. For all experiments, we performed both My→En and En→My directional sub-tasks on character, sub-syllable and syllable units for Myanmar side and typically smallest character units for English side. The statistical data are mentioned in the previous NE dictionary construction section.

Finally, we implemented our home-made character unit segmenter to segment character and sub-syllable unit segmenter developed by [1], to segment sub-syllable units respectively. We also used syllable segmenter [7] using regular expression for syllable segmentation scheme. All names are lowercased for both languages, and characters separated by space. The Moses script [8] clean-n-corpus.perl is only applied on preprocessed My-En monolingual data to remove lines containing more than 80 tokens. We described the sample data format for My-En NE instance pair for person names in Table 2.

Table 2: Sample data format for My↔En (NE) tasks

Unit	My↔EN	
Char	<ဘ ာ><ရ က ်><ဟ ူ><စ ိ န ်><အ ိ ု><ဘ ာ ား><မ ာ ား>	b a r a c k h u s s e i n o b a m a
Sub-Syl	<ဘာ><ာ><ရ><က်><ဟ><ူ><စ><ိန်><အ><ို><ဘာ><ား><မ><ား>	b a r a c k h u s s e i n o b a m a
Syl	<ဘာ><ရက်><ဟူ><စိန်><အို><ဘား><မား>	b a r a c k h u s s e i n o b a m a

6.2. Experimental Setting

To build contemporary NMT systems, we choose to rely on the transformer neural network architecture [3] since it has been substantiated to outperform in quality and efficiency, the two other mainstream architectures for NMT known as deep recurrent neural network (deep RNN) and convolutional neural network (CNN).

As the original paper indicated [3], Transformer has been used the attention-mechanism we saw earlier. Like LSTM, Transformer is the architecture for transforming one sequence into another one with the help of two parts encoder and decoder, but it differs from the conventional existing sequence to sequence models because it does not imply any recurrent networks (GRU, LSTM, etc.). Additionally, because there is no longer a sequential recurrent network, model training can be better parallelized, minimizing model training time. To train for Transformer model, we exert on Opennmt toolkit that are publicly released by [9]. All our Transformer system was consistently trained with the following hyper-parameters on Opennmt except to changing for layers of network and heads. Model hyper-parameter settings are described as the following Table 3.

Table 3: Hyper-parameter setting for Transformer

Parameters	**Setting**
layers	2,4,6
rnn_size	512
word_vec_size	512
transformer_ff	2048
heads	2,4,8
encoder_type	Transformer
decoder_type	Transformer
position_encoding	-
train_steps	50000
max_generator_batches	2
dropout	0.1
batch_size	1024
batch_type	tokens
normalization	tokens
accum_count	2
optim	adam
adam_beta2	0.998
decay_method	noam
warmup_steps	8000
learning_rate	2
max_grad_norm	0
param_init	0
param_init_glorot	-
label_smoothing	0.1
valid_steps	10000
save_checkpoint_steps	10000
world_size	1
gpu_ranks	0

6.3. Experimental Results

The BLEU score [10] on the character level, sub-syllable level and syllable level was used in the evaluation. The experimental results for En→My transliteration in addition to the reversed My-to-En transcription are expressed in Table 4. For the Transformer model, different combinations of layers (L) and heads (H) were compared in the experiments.

Table 4: Experimental Results for My-En (NE) Transliterations

System	**My→En**			**En→My**		
	Char	Sub-Syl	Syl	Char	Sub-Syl	Syl
Open NMT/Transformer (L6,H8)	0.92	0.92	0.92	0.75	0.74	0.76
Open NMT/Transformer (L2,H2)	0.89	0.92	0.92	0.85	0.86	0.78
Open NMT/Transformer (L4,H4)	0.91	0.92	0.91	0.86	0.84	0.76
Open NMT/Transformer (L2,H8)	0.92	**0.93**	**0.93**	0.86	0.80	0.85
Open NMT/Transformer (L4,H8)	0.91	0.92	0.92	0.**87**	0.76	0.86

By analysing the performance of linguistic feature for My→En task, sub-syllable units and syllable units perform well to transcribe names than character units on Transformer (L2, H8). The syllable structures have clearer and play an important role in Myanmar NLP pre-processing tasks. Likewise, sub-syllable units are also flexible and precise units for statistical approaches and depend on an insightful consideration of Myanmar phonology. Transformer (L4, H8) also achieves the impressive results upon character units but dramatically falloff the BLEU points on reverse direction.

7. Conclusion and Future Works

In this paper, we introduce our in-house Myanmar named entity (NE) terminology dictionary and address the case of (NE) transliteration between Myanmar and English with a systematic comparison of character units, sub-syllable units and syllable units using neural based approach; transformer model on our prepared segmented data. Although, our NE corpus is not so big, neural network models produce satisfied results for transliteration tasks, we still believe with more data and more experiments, neural network transliteration models will have a bright future in this field. This work can be further developed in various directions. Anyway, this exploration of using neural networks for Myanmar NE transliteration is the first work on Myanmar language.

8. References

[1] Ding, C., Pa, W. P., Utiyama, M., & Sumita, E. (2017, August). Burmese (Myanmar) name romanization: A sub-syllabic segmentation scheme for statistical solutions. In International Conference of the Pacific Association for Computational Linguistics (pp. 191-202). Springer, Singapore.

[2] Thu, Y. K., Pa, W. P., Sagisaka, Y., & Iwahashi, N. (2016, December). Comparison of grapheme-to-phoneme conversion methods on a myanmar pronunciation dictionary. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016) (pp. 11-22).

[3] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.

[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/

[5] Sin, Y. M. S., Oo, T. M., Mo, H. M., Pa, W. P., Soe, K. M., & Thu, Y. K. (2018). UCSYNLP-Lab Machine Translation Systems for WAT 2018. In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation.

[6] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19-51.

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

[8] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Proc. of NAACL, Vol. 1, pp. 48—54.

[9] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. arXiv:1701.02810.

[10] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proc. of ACL, pp. 311—318.