Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2980–2983
Marseille, 11–16 May 2020

A Myanmar (Burmese)-English Named Entity Transliteration Dictionary

Aye Myat Mon[†‡], Chenchen Ding[†*], Hour Kaing[†], Khin Mar Soe[‡], Masao Utiyama[†], Eiichiro Sumita[†]
[†] Advanced Translation Technology Laboratory, ASTREC, NICT
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
[‡] Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar
[†] {ayemyatmon, chenchen.ding, hour_kaing, mutiyama, eiichiro.sumita}@nict.go.jp
[‡] {ayemyatmon.ptn, khinmarsoe}@ucsy.edu.mm

Abstract

Transliteration is generally a phonetically based transcription across different writing systems. It is a crucial task for various downstream natural language processing applications. For the Myanmar (Burmese) language, robust automatic transliteration for borrowed English words is a challenging task because of the complex Myanmar writing system and the lack of data. In this study, we constructed a Myanmar-English named entity dictionary containing more than eighty thousand transliteration instances. The data have been released under a CC BY-NC-SA license. We evaluated the automatic transliteration performance using statistical and neural network-based approaches based on the prepared data. The neural network model outperformed the statistical model significantly in terms of the BLEU score on the character level. Different units used in the Myanmar script for processing were also compared and discussed.

Keywords: Myanmar (Burmese), named entity, transliteration, machine translation, neural network

1. Introduction

Transliteration is the task of transcribing words from a source script to a target script. Generally, the transcription is phonetically based. Transliteration processing is important for many downstream natural language processing (NLP) tasks, such as machine translation and information retrieval. In recent years, general sequence-to-sequence processing techniques for NLP tasks have been developed significantly. However, a lack of resources is still a problematic issue for many understudied languages.

In this study, we focus on transliteration between Myanmar (Burmese) and English. To facilitate the application of data-driven approaches, we manually collected a dictionary containing more than eighty thousand Myanmar-English transliteration instances. The data have been released under a CC BY-NC-SA license for research purposes.[1] Based on the dictionary, we conducted experiments on automatic transliteration between Myanmar and English. Specifically, we conducted experiments using two neural network (NN)-based approaches: the Transformer model using the OpenNMT system[2] (Vaswani et al., 2017; Klein et al., 2017) and a joint agreement bidirectional long short-term memory (LSTM)-based recurrent NN (RNN) using the JANUS[3] tool (Liu et al., 2016). A traditional phrase-based statistical machine translation (PBSMT) system using the Moses[4] toolkit (Koehn et al., 2007) was set as a baseline. The experimental results were evaluated using the BLEU score (Papineni et al. 2002) on the character level. The experimental approaches performed well on transliteration tasks. The NN-based approaches outperformed the traditional PBSMT by large gains. The effect of using units at different granularities in the Myanmar script was also investigated. To the best of our knowledge, this study is the first systematic work on the topic of Myanmar-English transliteration driven by a relatively large-scale dataset.

The remainder of this paper is organized as follows: In Section 2, related work is described. In Section 3, issues for the transliteration of Myanmar are addressed and the collected dictionary is described. In Section 4, the experimental results are reported, and in Section 5, a discussion is provided. In Section 6, the paper is concluded and future work is presented.

2. Related Work

Many Asian languages apply special writing systems, and efforts have been made on transliteration processing for major languages such as Chinese, Japanese, and Korean (Merhav and Ash, 2018). However, studies are required on understudied languages with limited resources.

Generally, the transliteration task can be modeled as a simplified translation task on the character/grapheme level rather than the word/phrase level, with no (or few) reordering operations. The technical background has been well established in the field of NLP. A PBSMT system (Koehn et al., 2003) can be used as an off-the-shelf tool once adequate data are provided. In recent years, NN-based frameworks such as the LSTM-RNN (Cho et al., 2014) have been widely applied to many NLP tasks. The Transformer model (Vaswani et al., 2017), which introduces a self-attention mechanism, is a state-of-the-art NN architecture in the NLP field.

Regarding specific studies on transliteration, Finch et al. (2016) proposed an agreement model of bidirectional RNN, which outperformed PBSMT on various language pairs. Wu and Yarowsky (2018) compared several machine translation methods for the transliteration of 591 languages into English. Their conclusion was that a PBSMT system outperformed other systems including NN-based approaches. Regarding the case of Myanmar processing, there are few previous works. Ding et al. (2017) first attempted a Myanmar name Romanization task, where NN-based approaches did not outperform the traditional approaches of the conditional random field and support vector machine. It can be considered that the transliteration task may be sensitive to the quality and quantity of training data, in addition to the features of the specific languages involved, so proper approaches should be investigated case by case.

* Corresponding author

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/my-en-transliteration.zip

[2] http://opennmt.net/OpenNMT-py/

[3] https://github.com/lemaoliu/Agtarbidir

[4] https://github.com/moses-smt/mosesdecoder

3. Transliteration of Myanmar

3.1 Overview

The Myanmar language belongs to the Sino-Tibetan language family. It is the official language of the Republic Union of Myanmar and the native language of the Bamar people. The Myanmar script is an abugida system. It consists of 33 characters of standalone consonants, four dependent consonants, and tens of diacritic marks that represent vowels and tones. The orthography of Myanmar is generally syllable-based, although syllables may be merged in special writing forms. One word can be composed of multiple syllables and one syllable can be composed of multiple characters. Between the characters and syllables, sub-syllable units may be identified for specific tasks. More details are available in Ding et al. (2017).

The contemporary Myanmar language contains a considerable number of borrowed words from English. Temporarily, there is no consistent standard in Myanmar to transcribe the growing number of borrowed words from or through English.

3.2 Difficulty

The difficulty in modeling transliteration between Myanmar and English is two-fold. The first issue is the difference in the phonological inventory of the two languages. Chang (2013) provided a study on the phonology of English loanword adaptation in Myanmar. This study was based on a list of 280 collected borrowed English words. As typical features of the Myanmar consonant system, stops and affricates make a three-way contrast with voiced, voiceless, and voiceless aspirated; a two-way voicing contrast is also present with nasals and approximants. These contrasts do not completely exist in English. Myanmar has a typical seven-vowel system, and the syllable codas are restricted to nasal and glottal. Although the vowel system of Myanmar is somehow simpler than that of English, three tones in Myanmar are superfluous when transliterating into English. By contrast, consonant clusters are limited in Myanmar, but abundant in English.

The second issue is non-phonetic orthographies. Both English and Myanmar apply an orthography that is more etymologically based than phonetically based. The Myanmar script has relative redundancy regarding its phonology inventory because many phonemes can be realized in more than one way. Moreover, some special spellings may be used intentionally to spell borrowed words for an exotic appearance. English spellings may also cause irregular transliteration where the transcription follows the spelling rather than the real pronunciation.

As a result of the above-mentioned issues, transliteration between Myanmar and English becomes correspondingly complex. In the following are some examples. <eu> in "Eurasia" and <u> in "United Kingdom," although spelt differently, are transliterated identically as <ယူ> (/jù/) into Myanmar. This is also the case for some pronunciations that cannot be distinguished in Myanmar, for example, <aun> in "auntie" and <un> in "uncle" are both transcribed into <အန်> (/ʔàɴ/). By contrast, <j> in "Japan" is transcribed by pronunciation as <ဂျ> (/d͡ʒ.../); however, <j> is more traditionally and consistently transcribed as <ဇ> (/z.../) in the case of "January," "June," and "July."

3.3 Dictionary

To collect the transliteration instances, we began with two English-Myanmar parallel corpora, and then moved to resources on the Internet to enlarge the scale of the data. Specifically, we used the ALT corpus (Riza et al., 2016; Ding et al., 2018; Ding et al., 2019; Ding et al., 2020)[5] and UCSY corpus (Sin et al., 2018).[6] The ALT corpus consists of twenty thousand parallel sentences from news articles and the UCSY corpus contains two hundred thousand parallel sentences collected from different domains, including local news articles and textbooks. We used the GIZA++ toolkit[7] (Och and Ney, 2003) to obtain the raw alignments between the source and target language, based on which the transliteration instances were filtered. From the Internet, we further collected instances of places, organizations, and person names.

The dictionary was encoded in Unicode. For experimental investigation, we divided it into three parts by taking the first one thousand lines as the test data and the final one thousand lines as the development data. Statistics for the data are listed in Table 1, where #Char. denotes Unicode characters, #Syl. denotes syllables, and #Sub-Syl. denotes the sub-syllable units proposed in Ding et al. (2017). Table 2 shows examples of the different units of the Myanmar transcription "ဆွစ်ဇာလန်" for "Switzerland."

	Myanmar			English
	#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

Table 1: Statistics for the data.

Char.	<ဆ><ွ><စ><်><ဇ><ာ><လ><န><်>
Sub-Syl.	<ဆ><ွစ်><ဇ><ာ><လ><နဲ>
Syl.	<ဆွစ်><ဇာ><လန်>

Table 2: Different units in Myanmar processing.

4. Experiment

The Myanmar data were segmented into different units, as shown in Table 2, and all English data were lowercased. Moses was used to train the baseline PBSMT system. Character alignment was generated using GIZA++ and symmetrized using the grow-diag-final-and heuristics. A 5-gram language model on character was trained using the SRILM toolkit (Stolcke, 2002).[8] Minimum error rate training (Och, 2003) was used to tune the weight of the features. All other parameters were adapted using the default setting of Moses. Regarding the JANUS toolkit, which implements a bidirectional agreement LSTM-RNN model, the hyperparameters followed the original paper (Liu et

[5] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

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

[7] http://www.statmt.org/moses/giza/GIZA++.html

[8] http://www.speech.sri.com/projects/srilm/

al., 2016). Specifically, the settings were 500 for embedding, 500 for the hidden unit dimensions, and 16 for the batch size. AdaDelta was used as the optimizer, with a decay rate of 0.95 and ε of 10^{-6}. The model was trained with 100 epochs. Regarding the Transformer model using the OpenNMT toolkit. The hyperparameters used in the experiments are listed in Table 3.

Parameter	Setting
rnn_size	512
word_vec_size	512
transformer_ff	2048
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

Table 3: Hyperparameter settings for the Transformer model.

	Char.	Sub-Syl.	Syl.
PBSMT	0.82	0.81	0.84
LSTM-RNN	0.87	0.90	0.92
Transformer (L6, H8)	0.92	0.92	0.92
Transformer (L2, H2)	0.89	0.92	0.92
Transformer (L4, H4)	0.91	0.92	0.91
Transformer (L2, H8)	0.92	**0.93**	**0.93**
Transformer (L4, H8)	0.91	0.92	0.92

Table 4: Myanmar-to-English results.

	Char.	Sub-Syl.	Syl.
PBSMT	0.73	0.64	0.77
LSTM-RNN	0.84	0.84	0.76
Transformer (L6, H8)	0.75	0.74	0.76
Transformer (L2, H2)	0.85	0.86	0.78
Transformer (L4, H4)	0.86	0.84	0.76
Transformer (L2, H8)	0.86	0.80	0.85
Transformer (L4, H8)	0.**87**	0.76	0.86

Table 5: English-to-Myanmar results.

The BLEU score (Papineni et al. 2002) on the character level was used in the evaluation. We used bleukit[9] for the calculation. The experimental results for English-to-Myanmar (En→My) transliteration in addition to the reversed Myanmar-to-English (My→En) transcription are provided in Tables 4 and 5, respectively. For the Transformer model, different combinations of layers (L) and heads (H) were compared in the experiments.

5. Discussion

In this section, we discuss case studies on transliteration instances that were difficult to process. In Tables 6 and 7, the outputs of different systems are compared for identical transliteration instances in both directions.

	Unit	En→My	My→En
PBSMT	Char.	ကာဒီဖ်	cardift
	Sub-Syl.	ကာဒီအက်ဖ်	cardif
	Syl.	ကာဒီအက်ဖ်အက်ဖ်	carဒစ်ဖ်
LSTM-RNN	Char.	ကာဒီဖ်	kadif
	Sub-Syl.	**ကားဒစ်ဖ်**	kadif
	Syl.	ကာဒစ်	kadif
Transformer (L2, H8)	Char.	**ကာဒစ်ဖ်**	**cardiff**
	Sub-Syl.	ကာဒစ်	cardif
	Syl.	ကာဒစ်	cars
Transformer (L4, H8)	Char.	**ကာဒစ်ဖ်**	cardif
	Sub-Syl.	ကာဒစ်	**cardiff**
	Syl.	ကာဒစ်	carbock

Table 6: Results for “cardiff” ↔ “ကားဒစ်ဖ်”.

Some non-native Myanmar spellings may appear in the transcription of borrowed English words. Table 6 presents a typical example of the pair “cardiff” and “ကားဒစ်ဖ်.” The spelling of <စ်ဖ်> used to transcribe <iff> is not permitted in Myanmar native words. It can be observed that syllable-based processing cannot handle such exceptional structures, regardless of which system is used.

	Unit	En→My	My→En
PBSMT	Char.	ဒဂျိုးကိုဗစ်	jokovic
	Sub-Syl.	ဒီဂျိုကိုဗစ်	jokovic
	Syl.	ဂျော်ကိုဗစ်	jokovic
LSTM-RNN	Char.	ဒီဂျိုကိုဗစ်	jokovic
	Sub-Syl.	ဒီဂျိုကိုဗစ်	jokovic
	Syl.	ဒီဂျိုးကိုဗစ်	jokovic
Transformer (All)	Char.	ဒီဂျိုကိုဗစ်	jokovic
	Sub-Syl.	ဒီဂျိုကိုဗစ်	jokovic
	Syl.	ဒီဂျိုကိုဗစ်	jokovic

Table 7: Results for “djokovic” ↔ “ဂျိုကိုဗစ်.”

[9] http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/

Table 7 presents a difficult example of the pair “djokovic” and “ဂျိုကိုဗစ်,” for which no system provided the correct results for either the En→My or My→En direction. The word “djokovic” itself is not a native English word, so all systems provided the more common spelling of “jokovic” in My→En processing. <djo> also caused difficulty in En→My processing, and in most cases, <d> was transcribed separately as <ဒ…>. The most approximate result was provided by PBSMT using Myanmar syllables, where <dj> was correctly transcribed as <ဂျ…>; however, the vowel was not transcribed exactly.

As the best performances of different Myanmar units do not differ so obviously in Tables 4 and 5, we consider that using character or sub-syllable units in Myanmar is a better option than syllables to avoid errors caused by irregular spellings, in most cases.

6. Conclusion and Future Work

In this study, we built and released a dictionary of Myanmar-English transliteration instances. Experiments were conducted on a baseline PBSMT system and two NN-based approaches to provide benchmark performance based on the prepared data.

Temporarily, we are collecting more instances so that the scale of the dictionary exceeds one hundred thousand instances. We will investigate the effect of using the data to improve the performance of Myanmar-English translation in the near future.

7. Bibliographical References

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078.

Chang, C. (2003). “High-interest loans”: The phonology of English loanword adaptation in Burmese.

Ding, C., Pa, W. P., Utiyama, M., and Sumita, E. (2017). Burmese (Myanmar) name romanization: A sub-syllabic segmentation scheme for statistical solutions. In Proc. of PACLIC, pp. 191—202.

Ding, C., Utiyama, M., and Sumita, Eiichiro. (2018). NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. ACM TALLIP, Vol. 18, Issue 2, Article No. 17.

Ding, C., Aye, H. T. Z, Pa, W. P., Nwet, K. T., Soe, K. M., Utiyama, M., and Sumita, E. (2019). Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-Speech Tagging. ACM TALLIP, Vol. 19, Issue 1, Article No. 5.

Ding, C., Yee, S. S. S, Pa, W. P., Soe, K. M., Utiyama, M., and Sumita, E. (2020). A Burmese (Myanmar) Treebank: Guideline and Analysis. ACM TALLIP, Vol. 19, Issue 3, Article No. 40.

Finch, A., Liu, L., Wang, X., and Sumita, E. (2016). Target-bidirectional neural models for machine transliteration. In Proc. of NEWS, pp. 78—82.

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

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In Proc. of ACL, poster and demo., pp. 177—180.

Liu, L., Finch, A., Utiyama, M., and Sumita, E. (2016). Agreement on target-bidirectional LSTMs for sequence-to-sequence learning. In Proc. of AAAI, pp. 2630—2637.

Merhav, Y., and Ash, S. (2018). Design Challenges in Named Entity Transliteration. arXiv:1808.02563.

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

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In Proc. of ACL Vol. 1, pp. 160—167.

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.

Riza, H., Purwoadi, M., Gunarso, Uliniansyah, T., Aw Ai Ti, Aljunied, S. M., Luong C. M., Vu T. T., Nguyen P. T., Chea, V., Sun, R., Sam, S., Seng, S., Soe, K. M., Nwet, K. T., Utiyama, M., and Ding, C. (2016) Introduction of the Asian Language Treebank. In Proc. of O-COCOSDA, pp. 1—6.

Sin, Y. M. S., Soe, K. M., and Htwe, K. Y. (2018). Large scale Myanmar to English neural machine translation system. In Proc. of GCCE, pp. 464—465.

Stolcke, A. (2002). SRILM: An extensible language modeling toolkit. In Proc. of ICSLP, pp. 901—904.

Wu, W., and Yarowsky, D. (2018). A comparative study of extremely low-resource transliteration of the world’s languages. In Proc. of LREC, pp. 938—943.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser. L, and Polosukhin, I. (2017). Attention is all you need. In Proc. of NIPS, pp. 5998—6008.