MyanBERTa: A Pre-trained Language Model For Myanmar

Aye Mya Hlaing
Natural Language Processing Lab.
University of Computer Studies, Yangon
Yangon, Myanmar
ayemyahlaing@ucsy.edu.mm

Win Pa Pa
Natural Language Processing Lab.
University of Computer Studies, Yangon
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract*— In recent years, BERT-based pre-trained language models have dramatically changed the NLP field due to their ability of allowing the same pre-trained model to successfully tackle a wide range of NLP tasks. Many studies highlight that language specific monolingual pre-trained models outperform multilingual models in many downstream NLP tasks. In this paper, we developed monolingual Myanmar pre-trained model called MyanBERTa on word segmented Myanmar dataset and evaluated it on three downstream NLP tasks including name entity recognition (NER), part-of-speech (POS) tagging and word segmentation (WS). The performance of MyanBERTa was compared with the previous pre-trained model MyanmarBERT and M-BERT for NER and POS tasks, and the Conditional Random Field (CRF) approach for WS task. Experimental results show that MyanBERTa consistently outperforms MyanmarBERT, M-BERT and CRF. We release our MyanBERTa model to promote NLP researches for Myanmar language and become a resource for NLP researchers. MyanBERTa model is available at: http://www.nlpresearch-ucsy.edu.mm/myanberta.html

Keywords*— MyanBERTa, Pre-trained Language Model, Name Entity Recognition, Part-of-Speech Tagging, Word Segmentation, Myanmar

I. INTRODUCTION

In recent years, pre-trained language models, especially BERT (Bidirectional Encoder Representations from Transformers) [1], have recently become extremely popular because of the effectiveness of the same pre-trained model to successfully tackle a broad set of natural language processing (NLP) tasks. While pre-trained Multilingual BERT (M-BERT) models[1] have a remarkable ability to support low-resource languages, a monolingual BERT models can get better performance than the Multilingual models.

In terms of pre-trained monolingual language models on Myanmar language, to the best of our knowledge, there are only two published works on monolingual Myanmar language modeling based on BERT, [3] and [4]. The monolingual MyanmarBERT [3] well learns of POS tagging task than M-BERT, however, the effectiveness of that model is less obvious in NER task compared to M-BERT. In [4], sentence-piece segmentation is directly applied on raw Myanmar corpus. Intuitively, pre-trained model based on word-level segmented data might perform well than raw dataset that has not been applied any kind of tokenization.

Due to the above concerns, we developed a BERT-based language model for Myanmar language which we name MyanBERTa. We trained MyanBERTa with byte-level Byte-Pair Encoding (BPE) vocabulary containing 30K subword units, which is learned after word segmentation on Myanmar dataset. We evaluated this model on three downstream NLP tasks: name entity recognition (NER), part-of-speech (POS) tagging and word segmentation (WS).

II. BUILDING MYANBERTA

In this section, we descibe the creation of Myanmar corpus, model architecture and training configuration for pretraining MyanBERTa.

A. Myanmar Corpus

Pretraining BERT language model relies on large size of plain text corpora. MyanBERTa is trained on a combination of MyCorpus [3] and Myanmar data collected from blogs and news websites. For Myanmar language, some preprocessing steps needs to be done on raw data to provide the effective usage of corpus. Therefore, standardizing encoding and word segmentation are done on Myanmar corpus to form the word segmented corpus that might be promote the effective usage of pre-training model on NLP downstream tasks.

Finally, we got the word segmented Myanmar corpus of size 2.1G. It consists of 5.9M sentences with 135M words. The detail statistics for Myanmar corpus used for pretraining language model is shown in Table I.

B. Model Architecture

The architecture of MyanBERTa is based on RoBERTa[2] which makes some modifications on BERT pretraining for more robust performance. Due to limited resources and to facilitate the comparison with MyanmarBERT [3], we opt to train a single RoBERTa based model using the original architectures of $BERT_{BASE}$ (L=12, H=768, A=12) with total parameters of 110M. The number of layers (i.e., Transformer blocks) is denoted as L, the hidden size as H, and the number of self-attention heads as A. Though, the original RoBERTa [2] uses a larger byte-level BPE vocabulary on dataset without any additional preprocessing or tokenization of the input, we apply 30,522 byte-level BPE vocabulary on word tokenized Myanmar corpus.

C. Pretraining

The model was trained for 528K steps using Tesla K80 GPUs. Due to memory constraints, we use the batch size of 8 with sequence length 512. It takes approximately 6 days for pretraining MyanBERTa. We use Huggingface's transformers [5] for both pre-training and fine-tuning.

TABLE I. STATISTICS OF MYANMAR CORPUS

Source	Number of Sentences	Number of Words	Size
MyCorpus [3]	3,535,794	83 M	1.3 G
News and Blog Websites	2,456,505	53 M	0.8 G
Total	5,992,299	136 M	2.1 G

[1] https://github.com/google-research/bert/blob/master/multilingual.md

III. Experimental Setup and Results

The performance of MyanBERT is investigated on three downstream Myanmar NLP tasks: NER, POS and WS by using F1 score.

A. Datasets for NER, POS and WS

In fine-tuning our pre-trained MyanBERTa model, Myanmar NE tagged corpus (60,500 sentences) manually annotated with the predefined NE tags proposed in [6] was employed for NER task and the publicly available POS dataset [7] which consists of 11K sentences for POS tagging. The same experimental data setting as [3] was applied for NER and POS tasks. For WS dataset, the word segmented Myanmar corpus (48,950 sentences) was prepared by applying BIES (Begin, Inside, End, and Single) scheme in labelling each syllable in the corpus.

B. Fine-tuning MyanBERTa

For NER, POS and WS downstream tasks, we fine-tune MyanBERTa for each task and dataset independently as shown in Fig. 1. The task-specific models such as NER, POS and WS models are formed by incorporating MyanBERTa with one additional output layer. In the figure, *E* represents the input embedding which is the sum of token embeddings, T_i represents the contextual representation of token *i*, [CLS] is the special symbol for classification output. Tag 1 to Tag N represent the output tags for each Tok 1 to Tok N, respectively.

During the fine-tune process, the batch size of 32 with the maximum sequence length 256 and the Adam optimizer with the learning rate of $5e^{-5}$ was used. We fine-tune in 30 epochs on each training data and evaluate the task performance on the validation set. And then select the best model checkpoint to report the final result on the test set.

C. Experimental Results

In Table II, we can see that fine-tuning MyanBERTa outperforms MyanmarBERT and M-BERT on both NER and POS tasks. Particularly, MyanBERTa achieves better F1 scores of 93.5% on NER task and 95.1% on POS task which are 0.2% and 3.7% absolute improvement over MyanmarBERT.

The comparison of pre-trained MyanBERTa model and CRF model which is the same setting as [8] on WS task is shown in Table III. As we can see, MyanBERTa get 0.7% higher F1 score than the CRF.

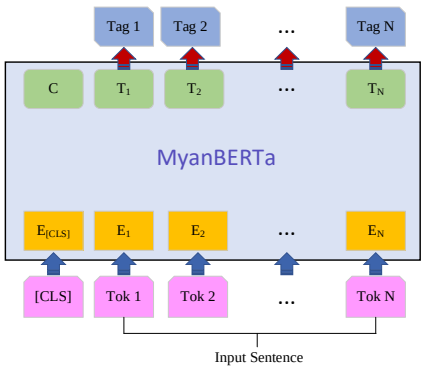

Fig. 1. An illustration of fine-tuning MyanBERTa on downstream NLP tasks: NER, POS and WS

TABLE II. Performance Scores on NER and POS

Pre-trained Models	NER (F1)	POS (F1)
M-BERT	93.4%	86.6%
MyanmarBERT [3]	93.3%	91.4%
MyanBERTa	**93.5%**	**95.1%**

TABLE III. Performance Scores on WS

Models	F1
CRF	96.6%
MyanBERTa	**97.3%**

In summary, we proved that our pre-trained MyanBERTa is effective on all three downstream tasks.

IV. Conclustion

A Myanmar monolingual pre-trained model, MyanBERTa, fine-tuned on three tasks, NER, POS and Word segmentation is described in this paper with the purpose of publicly available resource releasing for various Myanmar NLP tasks. The detail settings of experiments on pre-training and fine-tuning MyanBERTa are discussed in this paper. The performance of MyanBERTa is proved on fine-tuning three downstream tasks, and the absolute improvements over previous MyanmarBERT and M-BERT are achieved.

In the future, the pre-trained MyanBERTa can be used to improve other important Myanmar language processing tasks such as Sentiment Analysis, Natural Language Inference, Machine Translation, Automatic speech recognition and also to benchmark pre-trained Myanmar BERT models with larger dataset.

References

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[3] Saw Win, Win Pa Pa, "MyanmarBERT:Myanmar Pre-trained Language Model using BERT", The 19th IEEE Conference on Computer Applications (ICCA), 2020.

[4] Jiang, Shengyi, Xiuwen Huang, Xiaonan Cai, and Nankai Lin. "Pre-trained Models and Evaluation Data for the Myanmar Language." In International Conference on Neural Information Processing, pp. 449-458. Springer, Cham, 2021.

[5] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).

[6] Mo, Hsu Myat, and Khin Mar Soe. "Myanmar named entity corpus and its use in syllable-based neural named entity recognition." International Journal of Electrical & Computer Engineering (2088-8708) 10, no. 2 (2020).

[7] Htike, Khin War War, Ye Kyaw Thu, Win Pa Pa, Zuping Zhang, Yoshinori Sagisaka, and Naoto Iwahashi. "Comparison of six POS tagging methods on 10K sentences Myanmar language (Burmese) POS tagged corpus." In Proceedings of the CICLING. 2017.

[8] Pa, Win Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. "Word boundary identification for Myanmar text using conditional random fields." In International Conference on Genetic and Evolutionary Computing, pp. 447-456. Springer, Cham, 2015