Named Entity Recognition for Myanmar Language

Hsu Myat Mo
Natural Language Processing Lab
University of Computer Studies, Yangon (UCSY)
Yangon, Myanmar
hsumyatmo@ucsy.edu.mm

Khin Mar Soe
Natural Language Processing Lab
University of Computer Studies, Yangon (UCSY)
Yangon, Myanmar
khinmarsoe@ucsy.edu.mm

***Abstract*—Named Entity Recognition (NER) is extracting useful information called entities from sentences. NER for Myanmar language is a vital step for Myanmar Natural Language Processing (NLP) research. Addressing the NER problem for Myanmar is not a straightforward task. For Myanmar NER, experiments have been conducted in various ways form statistical to deep learning approach, and even with the way of fine-tuning BERT based Myanmar pre-trained language model. In this paper, researches that have been conducted for Myanmar NER are presented and the comparative results from each model are also presented.**

Keywords—Myanmar language, Named Entity Recognition, Named Entity corpus, syllable based, Bidirectional LSTM CRF, fine-tuning pre-trained model

I. Introduction

Named Entity Recognition (NER) is the process of automatically extracting useful name entities in text and classifying the extracted name entities into predefined set of Named Entities (NE) categories. NER for Myanmar Language is a crucial step for Myanmar NLP. For Myanmar Language, to identify names in text automatically, it is more complicated compared to other languages for some reasons.

Myanmar language is the official language of the Republic of the Union of Myanmar. Myanmar texts are written in sequence form left to right with no regular space inserted between words but spaces are sometimes inserted between phrases. This would be one of the issues in identifying NE in Myanmar scripts. Moreover, Myanmar language has complex morphological structures and there is no indicator of names such as capital letter like in English language. NER task for Myanmar language is a complicated task because of the complex nature of language. Moreover, the writing structure has no definite order, thus it can also make the NER a difficult process. Another fact is that Myanmar names also take all morphological inflections which can lead to ambiguity. It can be said that how to address the issue of NER for Myanmar language is not an easy task.

There were attempts with rule-based approach to solve the problem of NER for Myanmar language [2], [3]. In this paper, various experiments conducted on the UCSY NE corpus, ranging from statistical way to deep learning approach and even by fine-tuning the pretrained BERT model, are presented. All the experiments are run on Tesla K80 GPU. In the next section, Myanmar NE corpus, statistical NER for Myanmar language, deep neural NER and fine-tuning the pretrained language model for Myanmar NER will be described.

II. Methdology

A. Myanmar Named Entity Corpus

One of the reasons that NER for Myanmar language has been a challenging task was the lack of linguistic resources. The very first step to carry out experiments which employs machine learning approaches for NER research is having the annotated NE corpus. As part of the NER research for Myanmar language, a manually annotated Myanmar NE tagged corpus was developed. There are over 60K sentences and containing the total number of 174,133 named entities of six different name categories (person name, location, organization, race, time and number) in this NE corpus. For more information and statistics about the manually annotated NE tagged corpus and detail construction of this Myanmar NE tagged corpus, check in the paper [6].

In order to convert the NER problem into a sequence labeling problem, a label is assigned to each token to indicate the NE in sentences. As tagging scheme, IOBES (Inside, Outside, Begin, End and Single) scheme was used for all the experiments. Moreover, syllables were considered as basic input token.

B. Statistical Named Entity Recognition for Myanmar Language

As for the statistical experiments and baseline training, Conditional Random Fields (CRF) was applied. The F-measure of 91.47 was gained from experiments that were carried out by tuning various parameters with different features [4],[5]. From those experiments, it showed that CRF performs the best when feature engineering is well prepared.

C. Neural Named Entity Recognition for Myanmar Language

For neural architecture, there are three main parts: character sequence representation layer, syllable sequence representation layer and inference layer. For each input syllable sequence, syllables are represented with syllable embeddings. The character sequence layer can be used to automatically extract syllable level features by encoding the character sequence within the syllable. As the input of the character sequence layer, character embeddings represent characters. CNN is first used to encode character-level information of a syllable into its character-level representation. With CNN, it takes a sliding window to capture local features, and then uses a max-pooling for aggregated encoding of the character sequence. Moreover, for learning character embedding from training data, bidirectional LSTM network as well as GRU was applied in comparison. However, according to the conducted experiments, when CNN is applied in character sequence representation layer, the performance is slightly better. Syllable representations are the concatenation of syllable embeddings and character sequence

encoding hidden vector. Then the syllable sequence layer takes the syllable representations as input; feeds them into bidirectional LSTM and extracts the sentence level features from left to right and also from right to left, which are fed into inference layer to assign a label to each syllable. Although GRU was also applied, the performance is not been satisfied as expected. Among all experiments, bidirectional LSTM can give the best performance. A sequential CRF is used to jointly decode labels for the whole sentence as CRF can take into account neighbouring tags. Experiments were performed with different hyperparameters settings and among all the experiments, CNN_BiLSTM_CRF model with Adam optimizer outperforms. For more information, check in the paper [5].

D. Fine-tuing Pretrained MyanBERTa Model for Myanmar NER

Fine-tuning is a way of utilizing transfer learning. Fine-tuning pre-trained language models has recently become a common practice in building NLP models for various tasks. Moreover, MyanBERTa1[1], a BERT based Myanmar pre-trained language model has recently been released publicity. MyanBERTa fine-tuning results on NER has been explored. For fine-tuning, a batch size of 16 was used, learning rate was set to 2e-5, weight decay was set as 0.01 and tuned with the number of training epoch from 3 to 5 and saved the model with the best performance.

III. EXPERIMENTAL RESULT

In this paper, various experiments conducted on Myanmar NER have been described. The comparative results of each model are shown in table I. Among all the models, deep neural architecture, CNN_BiLSTM_CRF with Adam optimizer gives the best performance. Moreover, MyanBERTa is completely effective and the result of fine-tuning the MyanBERTa [7] on Myanmar NER is also promising.

TABLE I. F-SCORE FROM DIFFERENT NER MODELS

Models	**Precision**	**Recall**	**F-score**
Baseline CRF	91.39	89.72	90.45
CNN_BiLSTM_CRF (SGD)	93.15	93.68	93.41
BiLSTM_BiLSTM_CRF (Adam)	94.79	94.57	94.68
CNN_BiLSTM_CRF(Adam)	**95.04**	**94.89**	**94.97**
With Pretrained MyanBERTa model	92.57	92.28	92.42

IV. CONCLUSION

NER is a major task in NLP field and it can help us quickly extract important information from text. Research on NER for Myanmar language is the major step in order to develop the Myanmar NLP research. In this paper, the F-score value from different models which were trained on Myanmar NE corpus was described. In the future, we will explore the NER performance by fine-turning the multilingual RoBERTa models also. With more experiments, better results will be reported in the future.

REFERENCES

[1] Ashish Vaswani, et al., ”Attention Is All You Need”, 6 DEC 2017.

[2] Thi Thi Swe, Hla Hla Htay, “A Hybrid Methods for Myanmar Named Entity Identification and Transliteration into English”, http://www.lrec-conf.org/proceedings/lrec2010/workshops/W16.pdf., 2010.

[3] Thida Myint, Aye Thida, “Named Entity Recognition and Transliteration in Myanmar Text”, PhD Research, University of Computer Studies, Mandalay, 2014.

[4] Mo H.M., Nwet K.T., Soe K.M, “CRF-Based Named Entity Recognition for Myanmar Language”. In: Pan JS., Lin JW., Wang CH., Jiang X. (eds) Genetic and Evolutionary Computing. ICGEC 2016. Advances in Intelligent Systems and Computing, vol 536. Springer, Cham, 2017.

[5] Mo, Hsu Myat, and Khin Mar Soe. "Syllable-Based Neural Named Entity Recognition for Myanmar Language." International Journal on Natural Language Computing (IJNLC) Vol.8, No.1, February 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) 10.2, 2020, pp. 1544-1551.

[7] https://huggingface.co/UCSYNLP/MyanBERTa/tree/main

[8] https://github.com/huggingface/notebooks/blob/main/examples/token_classification.ipynb

[1] https://huggingface.co/UCSYNLP/MyanBERTa/tree/main