View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institute of Advanced Engineering and Science

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 2, April 2020, pp. 1544~1551
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i2.pp1544-1551

Myanmar named entity corpus and its use in syllable-based neural named entity recognition

Hsu Myat Mo, Khin Mar Soe
Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar

Article Info

Article history:

Received Apr 14, 2019
Revised Oct 15 2019
Accepted Oct 25, 2019

Keywords:

Bidirectional LSTM
Deep learning
Myanmar language
Named entity corpus
Neural NER
Syllabel-based

ABSTRACT

Myanmar language is a low-resource language and this is one of the main reasons why Myanmar Natural Language Processing lagged behind compared to other languages. Currently, there is no publicly available named entity corpus for Myanmar language. As part of this work, a very first manually annotated Named Entity tagged corpus for Myanmar language was developed and proposed to support the evaluation of named entity extraction. At present, our named entity corpus contains approximately 170,000 name entities and 60,000 sentences. This work also contributes the first evaluation of various deep neural network architectures on Myanmar Named Entity Recognition. Experimental results of the 10-fold cross validation revealed that syllable-based neural sequence models without additional feature engineering can give better results compared to baseline CRF model. This work also aims to discover the effectiveness of neural network approaches to textual processing for Myanmar language as well as to promote future research works on this understudied language.

Copyright © 2020 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Hsu Myat Mo,
Natural Language Processing Lab,
University of Computer Studies, Yangon,
No.4 Main Road, Shwe Pyi Thar Township, Yangon, Myanmar.
Email: hsumyatmo@ucsy.edu.mm

1. INTRODUCTION

Named Entity Recognition (NER) is the process of automatically tagging, identifying or labeling different named entities (NE) in text in accordance with the predefined sets of NE categories, for instance, person, location and organization and so on. The task of NER for Myanmar language is absolutely necessary to Myanmar natural language processing. The task of identifying names automatically in Myanmar text is more complicated compared to other languages for many reasons and it has been a challenging issue.

At the present time, Myanmar NLP is at an initial stage and well-prepared lexical resources required for Myanmar NLP research have not been available sufficiently until now. The lack of resources such as annotated NE corpus, name lists, gazetteers or name dictionary is the main issue in resolving NER for Myanmar language. The resource corpus is vital while conducting experiments to address this NER problem. For this reason, in this work, a very first Myanmar NE tagged corpus was manually annotated with the defined NE tags, and was constructed.

On the other hand, Myanmar language has both complex and rich morphology and ambiguity as well. Besides, Myanmar language has distinct characteristics and having no capitalizing feature which is the main indicator of proper names for some other languages like English. Further, its writing structure is of the free order, makes the NER a complex process. Some proper names of foreign person and location are lowanwords or transliterated words so that there are wide variations in some Myanmar terms. Names in Myanmar texts also take all morphological inflections wich can lead to ambiguity. This ambiguity of NE may lead to problem in classifying named entities into predefined types. Moreover, word segmentation is necessary to detect words boundary in written Myanmar text and this segmentation result might affect the NER performance. It can be said how to perform the task of recognizing names in Myanmar scripts automatically is still challenging until now.

One of the reasons why we try to solve this NER problem is to provide NER model to integrate to other NLP research and applications for Myanmar language. Moreover, the current Myanmar-English machine translation system could not recognize names written in Myanmar scripts properly. Futhermore, the main motivation behind this work is that, at present time, there is no publicly available NER tool that can extract named entities in written Myanmar texts.

As far as being aware and up to our knowledge, there is no publicly available NE tagged corpus, and also no work has been published for using deep neural networks on NER for Myanmar language. However, there are many benchmark data resources available for other languages. CoNLL 2003 [1] dataset which includes 1,393 English and 909 German news articles, MUC-6 and MUC-7 [2] provided through their Shared Task. Development of Bengali NE tagged corpus was described in [3]. The authors in [4] presented a workflow of building an English-Vietnamese NE corpus from an aligned bilingual corpus. In [5], a method to automatically build a NE corpus based on the DBpedia ontology was proposed. Likewise, construction of Portuguese NE corpus was proposed by using DBpedia as well [6].

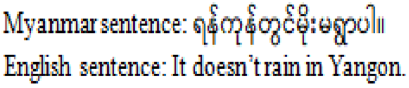
Although statistical approaches such as CRFs have been extensively applied to NER tasks, those approaches heavily rely on feature engineering. A general neural architecture for sequence labeling tasks was developed by [7]. Following this work, various deep neural networks have been got popularity for NER and have been widely applied on different languages, e.g., neural architectures for Japanese NER [8], Italian NER [9], Mongolian NER [10], and Russian NER [11] and so on. Likewise, deep neural architectures also have been deployed on NER for different domains, e.g., Biomedical Neural NER [12], Neural NER for Medical Entities in Twitter [13], NER for Twitter Massages [14], and NER for disease names [15]. All these works revealed that neural networks have the great capability for NER tasks and significantly outperform statistical algorithms.

Previous attempts on Myanmar NER had been done by applying rule-based and statistical approaches. A method for Myanmar Named Entity Identification was proposed by means of hybrid approach in [16]. Their approach is a combination of rule-based and statistical N-grams based method; plus small size of name database was also jointly used. A Myanmar Named Identification algorithm was proposed in [17]. It defines the names by using some of the POS information, NE identification rules and clue words around the contexts of NEs that carry information for NE identification. As limitation, input sentence must be established with POS tags. As weakness, it is incapable of sematic implication of proper names. Moreover, these approaches totally rely on linguistic knowledge and feature engineering. CRF-based NER for Myanmar language can be seen in [18]. The authors of [19] also try to explore the various combinations of features for CRF-based Myanmar NER.

In this work, a very fast manually annotated NE tagged for Myanmar language was developed and proposed to provide NE tagged corpus for future NER research. Moreover, to reduce the need of expensive feature engineering and to provide a good quality NER model for Myanmar language, experiments were conducted by modelling defferent deep neural network architectures. Myanmar NER was solved by means of deep neural networks modeling and it was considered as sequence labeling problem. In this neural modeling, a sentence is taken to be a sequence of syllables and syllables are considered as basic input units to the networks rather than characters or words to avoid the reliance on word segnetation. Among all experiments, the proposed neural model which is implemented by representing syllables as a combination of syllable embedding with Convolutional Neural Network (CNN) over the characters of the syllables, following this with bidirectional Long Short-Term Memory (LSTM) layer over the syllable representaions of a sentence, and finally adding CRF decoding layer above gives the best performance. The performance of neural NER model was also compared with baseline Conditional Random Field (CRF) model. Result from 10-fold cross validation revealed that deep neural networks can give the better results without using any external features.

2. MYANMAR LANGUAGE

Myanmar language also called Burmese is the official language of the Republic of the Union of Myanmar and has more than one thousand years' history. According to the documents, it belongs to the Sino-Tibetan language family and Myanmar scripts were descended from the Brahmi script of ancient South India. Myanmar scripts are written in sequence from left to right without inserting regular white space between words or syllables but white space may sometimes be inserted between phrases see Figure 1 for the example. The Myanmar scripts usually have 75 characters in total and those characters can be classified into 12 groups. For more details, you can check the paper [20].



Myanmar sentence: ရန်ကုန်တွင်မိုးမရွာပါ။
English sentence: It doesn't rain in Yangon.

Figure 1. Example of Myanmar writing

Myanmar language is syllabic language. Words in Myanmar language are composition of one or more syllables and a syllable may also contain one or more characters. A word 'နိုင်ငံ' can be separated into two syllables 'နိုင်' and 'ငံ'. A character can stand as a syllable itself or a syllable in Myanmar language may be made up of one or several characters. For example, five Unicode characters, i.e., 'န, ိ, ု, င' and '်' constitute to form the syllable 'နိုင်' see Figure 2.

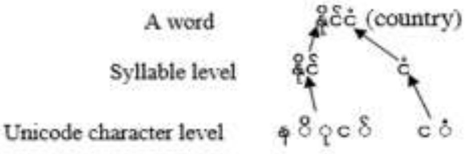


Figure 2. An Example of a Myanmar word formation

For Myanmar language, there are some complex encoding problems. Most Myanmar people are more familiar with encodings that do not follow standard Unicode point. To make Myanmar syllable structure represented in a definite way, in this paper, Unicode encoding is used. A Myanmar syllable structure formation is quite definite and straightforward. Although the constituents can appear in different sequences, a Myanmar syllable usually consists of one base initial consonant and may also have (or not) medials, vowels and optional dependent various signs. Independent vowels, independent various signs and digits can stand as stand-alone syllables and some consonants can act as a syllable as well. According to the Unicode standard, regardless of the appearance of the characters on screen, vowels are stored after the consonant. Detailed explanation of syllable structure segmentation for Myanmar language was described in [20, 21].

Myanmar language is very complex compared to English and other languages. On the other hand, language resources for Myanmar NLP researches have not been well prepared until now. As one of agglutinative languages, Myanmar has complex morphological structure so that for models in which words are treated as basic unit to construct distributed representation, there may be problems for those rich morphological words. On the other hand, it is necessary to deal with out-of-vocabulary words and word segmentation problem is also one of the important problems for Myanmar language. As described above, words in Myanmar language are not always separated by white spaces, so word segmentation is necessary and segmentation errors will affect the NER performance. In Myanmar language, syllable is the smallest linguistic unit that can carry information about word and syllable structure can be easily detected when with Unicode encoding. For these reasons, syllable is treated as the basic unit for label tagging in all our NER experiments.

3. DEVELOPMENT OF MYANMAR NE TAGGED CORPUS

Annotated corpora are vital resources for NLP and information extraction approaches which employ machine learning techniques. Building annotated NE tagged corpus is also the first step in training NER model and implementing NER system especially for low-resourced languages. As mentioned before, annotated corpra for Myanmar language are limited and scattered. This is one of the main reasons why Myanmar NLP lagged behind when compared to others. For our experiments and further research on Myanmar NER, we developed a manually annotated Myanmar NE tagged corpus. It took nearly 7 months to annotate the NE corpus manually. Currently we have over 60K sentences in total and manually annotated according to the defined six NEs tags. To be exact, there are totally 60,500 sentences and containing total number of 174,133 named entities. There is no other available Myanmar NE corpus that has as much data as our NE tagged corpus.

3.1. Data collection and preparation

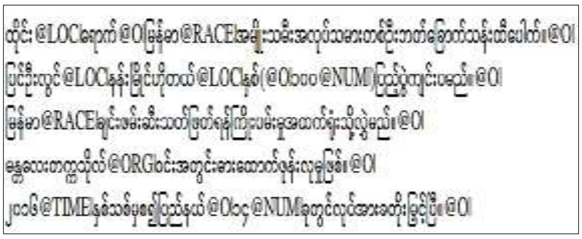
Nowadays, huge amount of web data are available and become the main source of data for computational research processing. In developing our Myanmar NE tagged corpus, news sentences written in Myanmar scripts within a range of year from 2017 to 2019, from online official news websites are collected and used. Different types of news gender including business, crime, health, tourism, education, environment, technology, sport, religion, and also politic are organized. In addition, translated sentences from wiki news as well as sentences supported from ALT-Parallel-Corpus, which is one part of the Asian Language Treebank (ALT) project under ASEAN IVO, are also utilized in construction of our NE tagged corpus. Sentences from ALT corpus are translated from International news so that a lot of transliterated names are appeared in this corpus.

As data preparation, data cleaning is firstly carried out. All kinds of mistyped errors are corrected manually. Moreover, different encodings need to be in uniform encoding. For encoding consistency, all the collected data are converted into Unicode encoding. In sentences, some typing errors are found. Especially, the digit "၀" (zero) and the consonant "ဝ" ("Wa") are usually mistyped. Thus, it is necessary to correct such kind of wrongly typed errors because the quality of data strongly affects the performance.

3.2. Defined NE types

Totally six types of NE tags are defined for manual annotation: PNAME, LOC, ORG, RACE, TIME and NUM. PNAME tag is used to indicate person names including nickname or alias, while LOC tag is defined for location entities. In this case, location entities include politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountain, etc.). Location also includes man-made structures like airports, highways, streets, factories and monuments. Names of organizations (government and non-government organizations, corporations, institutions, agencies, companies and other groups of people defined by an established organizational structure) are annotated with ORG tag.

In our Myanmar language, some location names and names of national races have same spelling in writing scripts. For example, the location name "ကချင်"(Kachin State) and one of the national races "ကချင်" (Kachin race). For this reason, the NE tag RACE is defined to indicate names of national races. TIME is used for dates, months and years. NUM tag is used to indicate number format in sentences. Example annotated sentences are described in Figure 3. Another tag "O" is used to indicate words which are not part of any defined NE types in sentences. The symbol "|" is used to separate the boundary of each tag. Further, the description of defined NE tags and their usage is shown in Table 1. Table 2 lists the entiites distribution in our manually annotated NE tagged corpus. The occurrence of each defined NE type in the annotated tagged corpus is also shown in Figure 4. It is observed that the location entity is the most appeared entity in the corpus which is about 36% out of all NEs. The race type is the least occurred entity which is less than 5%.



ထိုင်း@LOC|နောက်@O|မြန်မာ@RACE|အမျိုးသမီးအလုပ်သမားတစ်ဦးဘတ်စ်ကားတိုက်ခံရသေဆုံးပေါင်း။@O|
ပြင်ဦးလွင်@LOC|နှင့်မြို့ဟိုတယ်@LOC|နှစ်(@O|၁၀၀@NUM|)ပြည့်ပွဲကျင်းပမည်။@O|
မြန်မာ@RACE|ချင်းဖမ်းဆီးသတ်ဖြတ်ရန်ကြိုးပမ်းနေသောထင်ရှားသူလွှဲမည်။@O|
မန္တလေးတက္ကသိုလ်@ORG|ဝင်းအတွင်းဓားထောက်လုယက်မှုဖြစ်။@O|
၂၀၁၆@TIME|နှစ်သစ်မှစ၍ပြည်နယ်@O|၁၄@NUM|ခုတွင်လုပ်အားခတိုးမြှင့်ပြီ။@O|

Figure 3. Example sentences from Myanmar NE corpus

Table 1. Defined NE types and their usage

Defined NE Types	Example Usage
PNAME	ကြယ်စင်၊ အေးအေးမွန်၊ အိုဘားမား
LOC	မြန်မာ၊ မားစ်၊ ရန်ကုန်၊ မန္တလေး၊ မုံရွာ
ORG	ရန်ကုန်ကွန်ပျူတာတက္ကသိုလ်၊ ရိုးမဘဏ်
RACE	ဗမာ၊ ကချင်၊ ချင်း၊ ကိုရီးယား
TIME	နိုဝင်ဘာ၊ တဆောင်မုန်း၊ သောကြာ၊ ၁၅.၆.၂၀၁၉
NUM	၅၆၇၊ ၁၀၀၀၊ ၃.၁၄

Table 2. Corpus data statistics

Data	Total Number
Sentences	60,500
Number of NE	174,133
PNAME	35,405
LOC	63,492
ORG	19,740
RACE	5,720
TIME	29,472
NUM	2,0304

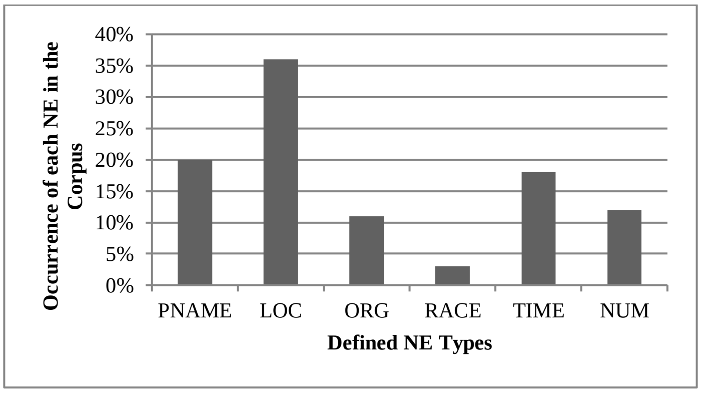


Figure 4. Occurrence of defined NE types in Myanmar NE corpus

4. EXPERIMENTAL SETUP

Experiments are performed by comparing different neural models on syllable level text rather than word level and the performance results are compared with baseline statistical CRF model. In our neural training, we try to investigate various neural network architectures that automatically detect syllable and character level features using bidirectional Long Short-Term Memory (LSTM) [22], Convolutional Neural Network (CNN) and also Gated Recurrent Unit (GRU) [23] architecture to eliminate the need for most feature engineering. Both CNN and bidirectional LSTM networks have been investigated for modeling character level information. As inference layer, CRF is used rather than softmax layer. We also conducted experiments with sofimax inference layer before applying CRF as inference layer. However, the performance results were not as good as using CRF as inference layer.

4.1. Tagging scheme

In order to convert the NER problem into a sequence labeling problem, a label is assigned for each token (syllable) to indicate the NE in sentences. We used the Myanmar syllable segmentation algorithm "sylbreak" of [24] on sentences for the syllable data representation and syllable-based labeling. As tagging scheme, IOBES (Inside, Outside, Begin, End and Single) scheme is used for all the experiments.

4.2. Neural training

For the implementation of the neural network model training, we utilized the PyTorch framework [25] which provides flexible choices of feature inputs and output structure. Experiments were run on Tesla K80 GPU. Based on different parameter settings, the training time for each experiment is different. Given a Myanmar sentence, syllable is treated as the basic training unit for label tagging. In this 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 is used to automatically extract syllable level features by encoding the character sequence within the syllable. We firstly use CNN to encode character-level information of a syllable into its character-level representation. 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; feed them into bidirectional LSTM and extracts the sentence level features, which are fed into inference layer to assign a label to each syllable. Instead of using the softmax output from this layer, we use a sequential CRF to jointly decode labels for the whole sentence as CRF can take into account neighbouring tags see Figure 5. The proposed neural architecture gives the best performance among all other architectures that have been carried out during experiments. It will be referred as CNN_BiLSTM_CRF for short in the following sections. Likewise, other networks architectures will be referred in short form in accordance with the applied network for each part.

As to input embedding setting, we also tried experiments on using pre-trained 100-dim of embedding on syllable level and character level data, respectively. The data used for training the syllable and character embedding includes 200K sentences in total. However, pre-trained embedding did not provide better results. This may be due to the noisy nature of data. Therefore, data cleaning is necessary before training. As to optimization, both the stochastic gradient descent algorithm (SGD) and Adam algorithm were tried. For the SGD, it was performed with initial learning rate of 0.015 and momentum 0.1. The learning decay rate was set as 0.05. For the Adam algorithm, the initial learning rate was set as 0.0015.

Both optimizations had batch sizes set as 10. Early Stopping was used based on the performance on validation sets. A dropout of 0.5 was set for both embedding and output layers to mitigate overfitting in the training process. The hidden dimension was set to 200 in the whole experiment. The best accurancy happens at 45 epoches.

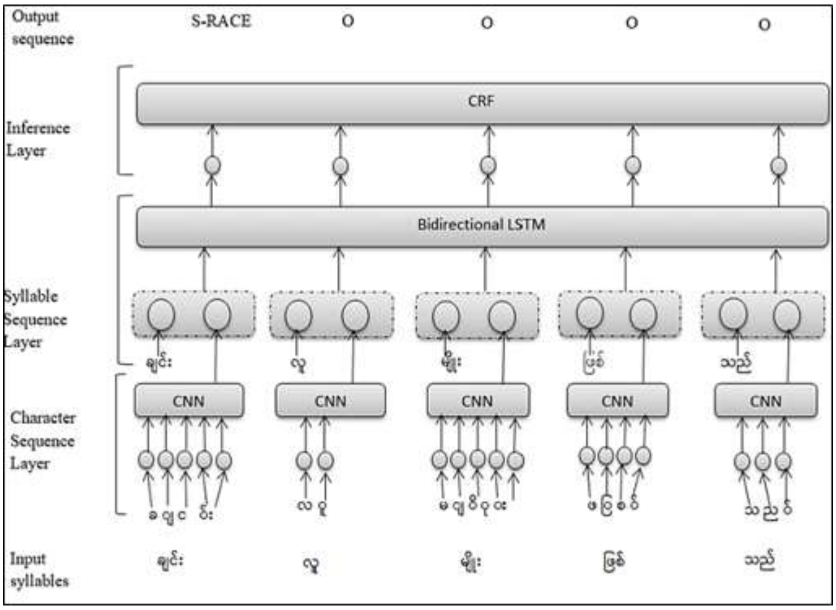


Figure 5. The architecture of neural network for Myanmar NER

5. EXPERIMENTAL RESULTS AND ANALYSIS

Before syllable-based experiments, character-based experiments were also conducted. However, character-based models do not perform as well as syllable-based models. For the reason that syllables as inputs hold more information than individual characters as inputs, it is reasonable that syllable-based models perform better than the character-based models. We separate our NE tagged corpus into 10-fold for all experiments. In Table 3, we listed 10-fold cross validation results from the best neural architecture (CNN-BiLSTM-CRF) among different neural model structures. The best outcome appears when we relied on a CNN model to learn character features first and then concatenated it with the syllables' embeddings as the input of to the bidirectional LSTM network rather than utilizing bidirectional LSTM model to learn character festures. The character features learned from LSTM is not as good as CNN in our experiments. When the GRU is applied on syllable, it makes the F-score value drops continuously as the epoch increases, which is out of our expectation. The performance of using SGD optimization algorithm is slightly worse than using Adam. From the experiments, it can be seen that CNN performs better than others for charcter sequence representation layer, bidirectional LSTM is suitable for syllable representation layer and CRF works better than softmax in inference layer. Therefore, the resut from Table 3 is resulted from CNN-BiLSTM-CRF model which is the best model among all experiments.

Table 3. The performance result of 10-fold cross validation (CNN-BiLSTM-CRF)

N-fold	Dev (Preciion/Recall/F-score)	Test (Preciion/Recall/F-score)
1-fold	90.35/89.65/90.00	94.34/94.19/94.27
2-fold	86.43/86.29/86.36	92.57/92.28/92.42
3-fold	88.98/88.90/88.94	94.74/94.22/94.48
4-fold	90.80/88.77/89.77	89.92/87.85/88.87
5-fold	90.73/90.56/90.65	90.44/88.88/89.65
6-fold	91.39/89.72/90.55	90.09/88.23/89.15
7-fold	90.68/88.65/89.65	91.47/90.97/91.22
8-fold	89.44/88.93/89.19	94.72/94.79/94.71
9-fold	90.28/88.28/89.27	89.69/87.96/88.81
10-fold	91.38/90.67/91.02	90.19/88.59/89.38
Average	90.05/89.04/89.54	91.82/90.80/91.30

For baseline CRF training, an open source toolkit for linera chain CRF [26] was utilized. When only used the tokens and their neighbouring contents as features and the window size was set as 5, F-score of 86.96% was obtained. When a small-sized named dictionary and clue words list was added as additional features, it made the F-score have around 2.4% increase (89.36%). It shows that CRF works the best when feature engineering is well prepared.

If compared the results in Table 4, we can see that neural models perform better than statistical CRF models. By comparison, syllable-based neural models without additional feature engineering perform better than CRF models. Syllable-based CRF model with additional feature is approaching the results of neural models. Although these experiments give the promising results, the size of data used for neural network model training is not so big compared to other NE corpra of other languages. Normally more data can help neural network learn better. Moreover, due to time and data limit, the hyper-parameters used in the experiment may be not the best. This needs modelling experiences and also large amount of trial and error experiments to decide.

Table 4. The performance comparison between neural sequence model and baseline CRF

Model	F-score (10-fold)
CNN-BiLSTM-CRF	**91.30**
BiLSTM-BiLSTM-CRF	90.23
CRF (with additional features)	89.36
CRF	86.96

6. CONCLUSION

In this work we have developed a manually annotated NE tagged corpus for Myanmar language with the intension of developing resources for Myanmar NLP and providing resource for further research. Moreover, we had explored the effectiveness of neural network on Myanmar NER and conducted a systematic comparison between neural approaches and traditional CRF approaches on our manually annotated NE tagged corpus. Experiment results revealed that the performance of neural networks on Myanmar NER is quite promising, because neural models did not use any handcrafted features or additional resources. Moreover, from the experiments, it can be seen that neural networks work well on syllable level data. Although our NE corpus is not so big, neural network models produce better performance than CRF models for Myanmar NER, we still believe with more data and more experiments, neural networks can learn better so as to produce better results. Anyway, this exploration of using neural networks for Myanmar NER is the first work to apply neural networks on Myanmar language. It showed us that bidirectional LSTM network with CRF decoding layer above on syllable level data jointly with CNN to extract character feature can facilitate Myanmar NER. With more data and more experiments, better results will be reported in the future and we will keep exploring neural networks on other Myanmar NLP works.

REFERENCES

[1] X T. Kim Sang, *et al.*, "Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, vol. 4, pp. 142-147, 2003.

[2] R. Grishman, *et al.*, "Message Understanding Conference-6: A Brief History," in *Proceeding of the 16th Conference on Computational Linguistics,* vol. 1, pp. 466-471, August 1996.

[3] A. Ekbal, *et al.*, "Development of Bengli Named Entity Tagged Corpus and its Use in NER Systems," in *The 6th Workshop on Asian Language Resources, 2008*, Jan 2008.

[4] H. Quoc Ngo, *et al*., "Building English-Vietnamese Named Entity Corpus with Aligned Bilingual News Articles", in *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing*, Jan 2014.

[5] H. Younggyun, *et al*., "Named Entity Recognition Corpus Construction using Wikipedia and DBpedia Ontology," in *Proceeding of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 2565-2569, May 2014.

[6] C. Weber and R. Vieira, "Building a Corpus for Named Entity Recognition using Portuguese Wikipedia and DBpedia," 2014.

[7] R. Collobert, *et al.*, "Natural Language Processing (Almost) from Scrtch," *Journal of Machine Learning Research,* vol. 12, pp. 2493-2537, 2011.

[8] S. Misawa, *et al.*, "Character-based Bidirectional LSTM-CRF with words and characters for Japenese Named Entity Recognition," in *Porceeding of the First Workshop on Subword and Character Level Models in NLP, Association for Computational Linguistics*, pp 97-102, 2017.

[9] D. Bonadiman, *et al*., "Deep Neural Networks for Named Entity in Italian," in *Proceeding of Second Italian Conference on Computational Linguistics CLiCit*, Dec 2015.

[10] W. Wang, F. Bao and G. Gao, "Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks," *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, CA, 2016, pp. 495-500.

[11] L. T. Anh, *et al*., "Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition," in *Artificial Intelligence and Natural Language Conference (AINL 2017)*, 2017.

[12] Lishuang Li, Liuke Jin, Zhenchao Jiang, Dingxin Song and Degen Huang, "Biomedical named entity recognition based on extended Recurrent Neural Networks," *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Washington, DC, 2015, pp. 649-652.

[13] A. Jimeno Yepes and A. Mackinlay, "NER for Medical Entities in Twitter using Sequence to Sequence Neural Networks," in *Proceedings of the Australasian Language Technology Association Workshop (ALTA 2016)*, pp. 138-142, 2016.

[14] N. Limsopatham and N. Collier, "Bidirectional LSTM for Named Entity Recognition in Twitter Messages," in *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pp. 145-152, 2016.

[15] Z. Zhehuan, *et al*., "ML-CNN: a novel deep learning based disease named entity recognition architecture," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015.

[16] T. Thi Swe and H. Hla Htay, "A Hybrid Method for Myanmar Named Entity Identification and Transliteration into English," in *Seventh International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[17] T. Myint and A. Thida, "Named Entity Recognition and Transliteration in Myanmar Text,", 2014.

[18] H. Myat Mo, *et al*., "CRF-Based Neamed Entity Recognition for Myanamr Language," in *Genetic and Evolutionary Computing, ICGEC 2016, Advances in Intelligent Systems and Computing*, vol. 536, pp. 204-211, 2016.

[19] H. Myat Mo, *et al*., "Exploring Features for Myanmar Named Entity Recognition," in *Proceeding of 15th International Conference on Computer Application (ICCA2017)*, pp. 429-433, 2017.

[20] Z. M. Maung and Y. Mikami, "A rule-based Syllable Segmentation of Myanmar Text," in *Proceedings of the IJCNLP-08 on NLP for Less Privileged Languages*, pp 51-58, 2008.

[21] T. H. Hlaing and Y. Mikami, "Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer," *International Journal on Advances in ICT for Emerging Regions (ICTer)*, vol.6(2), July 2014.

[22] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in *Neural Computation*, vol. 9(8), pp. 1735-1780, Nov 1997.

[23] C. Junyoung *et al*., "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *NIPS 2014 Deep Learning and Representation Learning Workshop*, 2014.

[24] Y. K. Thu, "Syllable Segmentation tool for Myanmar Language (Burmese)," 2017. https://github.com/ye-kyaw-thu/sylbreak.

[25] J. Yang, *et al*., 2017. https://github.com/pytorch/pytorch.

[26] Y. Kudo, "CRF++: Yet another CRF Toolkit," 2005. https://crfpp.sourceforge.net.

BIOGRAPHIES OF AUTHORS

Hsu Myat Mo got her B.C.Sc (Hons) in 2010, followed by M.C.Sc (Credit:) in 2012, respectively. Currently she is doing her Ph.D research focusing on Myanmar NER in Natural Language Processing Lab, at the University of Computer Studies, Yangon.

Dr. Khin Mar Soe got Ph.D(IT) in 2005. Currently she is working as a Professor and also the Head of Natural Language Processing Lab, at the University of Computer Studies, Yangon. She has been supervising Master thesis and Ph.D researches on Natural Language Processing. Moreover, she participated in the project of ASEAN MT, the machine translation project for South East Asian languages.