ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 95-100
doi:10.18178/wcse.2019.03.017

Joint Word Segmentation and Stemming with Neural Sequence Labeling for Myanmar Language

Yadanar Oo [1+] and Khin Mar Soe [2+]

[1] yadanaroo@ucsy.edu.mm (University of Computer Studies, Yangon, Myanmar)
[2] khinmarsoe@ucsy.edu.mm (University of Computer Studies, Yangon, Myanmar)

Abstract. Word segmentation is widely-studies sequence labeling problem using machine learning method like conditional random fields. In word segmentation, deep learning approaches have achieved state-of-the-art performance. Normally, segmentation is considered as a separate process from stemming. Our approach proposes a joint model that has stronger capabilities for Myanmar word segmentation and stemming. As far as we know, this is the first work on joint Myanmar word segmentation and stemming. In this paper, we evaluate the performance of neural network architecture that relies on two sources of information about syllable- and character-level representation, by using LSTM, CNN, GRU and CRF. For the comparison and analysis process, we examine the importance of different network designs and different factors such as the last layer of the network and different optimizers.

Keywords: Myanmar word segmentation, Stemming, joint model, neural networks.

1. Introduction

The large amount of online data are available, to retrieve accurate data for user query is very essential. In information retrieval system, stemming acts as an important tool to increase the retrieval accuracy. In Myanmar language, stemming is performed by stripping suffix and affix from the given sentence. Unlike English and other western languages, many Asian languages such as Chinese, Japanese, Thai and Myanmar, do not delimit words by whitespace. Myanmar language faces a similar problem due to the fact that a word contains more than one separate syllable and therefore whitespace is not always the word separator. Word segmentation is essential step for later NLP process. Without word segmentation, other processing steps cannot be done.

In this approach, we consider stemming as a typical sequence tagging problem over segmented words, while segmentation also can be modelled as a syllable-level tagging problem that identify the word boundaries via predicting the labels. Recently, neural network models have achieved state-of-the-art performance in the NLP community. Our approach proposes an effective joint neural sequence labelling model which predicts the combinatory labels of segmentation boundaries and stemming tag at the syllable level. The contributions of this paper are: (i) to compare the effectiveness of neural sequence labelling architectures in joint word segmentation and stemming and (ii) in analysis of different optimizers.

The NCRF++ toolkit [1] was used to build neural sequence labelling architecture for joint word segmentation and stemming of Myanmar word. An overview of the contents of the paper is a follow. Section 2 presents a brief overview of related work. Section 3 describes the proposed model. Experimental setup and results comparison and analysis of our system will give in Section 4. Finally, Section 5 concludes.

2. Related Work

[+] Yadanar Oo. Tel.: +96-95149552
E-mail address: yadanaroo@ucsy.edu.mm1, khinmarsoe@ucsy.edu.mm2

Word segmentation and stemming are important and essential pre-processing steps for Myanmar language processing tasks. In Myanmar language, [2] Word Segmentation system consists of four components, sentence splitting, tokenization, initial segmentation by Maximum Matching Algorithm and statistical combined model (bigram model and modified word juncture model) for final segmentation.

In recent year, most methods have treated the task as a sequence labelling problem. In [3], Conditional random field is used to identify Myanmar word boundaries within a supervised framework. CRF approach is compared against a baseline based on maximum matching using dictionary from Myanmar Language Commission Dictionary (word only) and manually segmented subset of the BTEC1 corpus.

In the recent research literature, neural models can be challenging. In [4], it explored three neural model designs: character sequence representation, word sequence representation and inference layer. Experiments show that character information improves model performance. In our approach, such a joint work is performed as a syllable-based neural sequence labelling architecture.

3. Proposed System Architecture

Our main contribution lies in combining neural network models for joint word segmentation and stemming task. We present a hybrid model of RNNs and CNNs that learn both character and syllable-level features, presenting the first evaluating of such architecture on Myanmar language evaluation datasets.

3.1. Syllable segmentation

Syllable is a basic sound unit. A word can be consisted of one or more syllables. Every syllable boundary can also be word boundary. Some word can include other words; it is called a compound word. Syllable breaking is a necessary step for Myanmar word segmentation. For syllable segmentation, this system uses the algorithm from [5]. Example of syllable segmentation are shown in Figure 1.

ကျန်းမာခြင်းသည် လာဘ် တစ်ပါး ဖြစ်သည် ။

ကျန်း မာ ခြင်း သည် လာဘ် တစ် ပါး ဖြစ် သည် ။

Fig. 1: Example of syllable segmentation in Myanmar Language.

3.2. Word formation in Myanmar language

The basic order of Myanmar sentence is subject-object-verb. There are nine Part-of-Speech classes for all Myanmar words. They are Noun, Pronoun, Verb, Adjective, Adverb, Conjunction, Postpositional Marker, particles and Interjection. In this paper, syllable is classified with four classes of word types.

Root Words

Root words can be Noun, Verb, Adjective and Adverb but it is a common form of word without any suffix or prefix. For example, in the word “ ကျောင်းသားများ ” (students) the root word is “ **ကျောင်းသား** ” (student).

Simple words

Simple words are Particle, Conjunction, Postpositional Marker. Like a stop words, these words appear so frequently that their usefulness is limited. In Information Retrieval ignores stop words at the time of searching a user query.

Prefix

Verbs are negated by the particle “ မ “[-ma], which is prefix to the verbs to form the negative verb and which also unchanged the root of verb. “ အ ”[-a] by affixing as ‘လုပ် ” can change verb form to noun ‘‘အလုပ် ” without changing the meaning.

Suffix

Adjectives are used to modify the noun. Myanmar adjectives can be form by combining verb and particles. For example, ‘ ပေါ်ထွက်လာခဲ့သော ’[appeared] is the adjective that combine the verb “ ပေါ်ထွက် ”[appear] and adjective suffix “ လာခဲ့သော”. Moreover, a word that modify verb is adverb. Myanmar adverbs are always before verb and there can be more than one adverb for one verb. Adverb also has suffix “ စွာ “[-swar]. Their stem form remains unchanged when suffix removal.

3.3. Tagging schemes

The task of joint word segmentation and stemming is to assign word type label to every syllable in a sentence. A single word could span several syllables within a sentence. In order to indicate the word boundaries, BIO format is represented where every syllable is labelled as B-label if the syllable is beginning of a word, I-label if it is inside a word but not the first token within the word, or O otherwise. The sentence is first segmented into syllable. Then, from the output, syllable boundary tagging is used to classify the word type and detect the boundary of words. For stemming, each syllable is tagged with one of four word types: Root word (R), Single word(S), Prefix (Pre) and Suffix (Suf). Figure 2 shows the example of syllable tagging in joint word segmentation and stemming.

ကျန်း/B-Rs မာ/I-Rs ခြင်း/B-Suf သည်/B-S လာဘ်/B-Rs
တစ်/B-S ပါး/B-Suf ဖြစ်/B-Rs သည်/B-Suf ။/O

Fig. 2: Syllable-tagging for joint segmentation and stemming.

In Figure 3, the root word is placed within the boundary marker [], and suffix words are marked by + and then prefix are delimited by ^ marker. Moreover, spaces between words are separated by _ marker.

[ကျန်းမာ]+ခြင်း_သည်_[လာဘ်]_တစ်+ပါး_[ဖြစ်]+သည်_။

Fig. 3: Example of stemming result.

Figure 4 shows the word segmentation result. In this figure, spaces between words are delimited by _ marker.

ကျန်းမာခြင်း_သည်_လာဘ်_တစ်ပါး_ဖြစ်သည်_။

Fig. 4: Example of word segmentation result.

4. Experiments

We investigate the comparison between different deep neural models on our joint task. And we examine the main factors that influence to system accuracy, such as inference algorithm, optimizers.

4.1. Data set

There is no standard corpus published for Myanmar language. So, to evaluate the proposed joint model, we use a training set selected from manually segmented 12,000 sentences that are collected for News Data. We divide the training corpus into two sets, the first 80% of the data to training and 10% each to development and test set. There are 5 different labels (11 with BIO prefix included). The dataset statistics are shown in Table1.

TABLE I: Statistics of datasets

Dataset	Train	Dev	Test	Label
@Sent	10,000	1,147	1,050	11
@Syllable	422k	67k	49k	

4.2. Pretrained embedding

Word embedding provides a good generalization to unseen words since they can capture general syntactic as well as semantic properties of words. Representing words as dense vectors, usually with 100-300 dimensions, is a widely used technique in NLP and it can significantly increase performance. Character embedding capture character level information, especially pre- and suffixes of words, can contain valuable information for linguistic sequence labelling tasks such as word segmentation, named entity recognition etc. In this joint model, we use Learning Word Vectors for 157 Languages that trained on 3 billion words from Wikipedia and Common Crawl using CBOW 300-dimension [6] for both word and character embedding.

4.3. Hyperparameters

Table 2 shows the hyperparameters used in our experiments, which mostly follow Ma and Hovy (2016), including the learning rate = 0.015 for LSTM models. For syllable CNN based model we take learning rate = 0.005 with epochs 100.

TABLE II: Hyperparameters

Parameters	Value	Parameters	Value
char emb size	300	word emb size	300
char hidden	50	syllable hidden	50
CNN window	4	LSTM window	1
batch size	20	dropout rate	0.5
L_2 regularization	le-8	learning rate decay	0.05
biLSTM	True	epochs	100

4.4. Network settings

Our joint word segmentation and stemming neural sequence labelling framework that contains three layers, i.e., character sequence representation, syllable sequence representation and an inference layer. In character sequence representation, we model three different neural structures and compare the performance through CNN, LSTM or GRU.

Similarly, on the syllable level, we investigate CNN, LSTM and GRU models for our joint sequence labelling tasks.

4.5. Classifiers

We evaluate the two options for last layer of the network that takes the extracted syllable sequence representations as features and assigns labels to the syllable sequences. SoftMax classifier maps the layer scores into a probability distribution for the possible tags. The tag with the highest probability is selected. In this approach, each token in a sentence is considered independently and correlations between tags in a sentence cannot take into account. CRF classifier captures label dependencies by adding transition scores between neighbouring labels. During the decoding process, the Viterbi algorithm is used to search the label sequence with the highest probability.

To simplify the description, we use “CCNN”, “CLSTM”, “CGRU” represent character structure and “SCNN”, “SLSTM”, “SGRU” represent syllable structure, respectively. Table 3 shows the experimental results on different architecture of networks under same hyperparameters with CRF inference layer. Training is done by stochastic gradient descent (SGD) optimizer with fixed learning rate.

TABLE III: Comparison and analysis of different architecture of network under the same hyperparameters

Model	Precision	Recall	F1
NoChar+SCNN+CRF	85.30	83.01	84.14
NoChar+SLSTM+CRF	83.52	81.60	82.55
NoChar+SGRU+CRF	84.06	82.66	83.36
CCNN+SCNN+CRF	85.39	83.39	84.22
CCNN+SLSTM+CRF	84.39	81.94	83.15
CCNN+SGRU+CRF	84.53	82.37	83.43
CLSTM+SCNN+CRF	86.08	83.43	84.74
CLSTM+SLSTM+CRF	83.80	82.41	83.10

CLSTM+SGRU+CRF	83.86	80.74	82.27
CGRU+SCNN+CRF	85.47	83.65	84.55
CGRU+SLSTM+CRF	84.95	83.39	84.16
CGRU+SGRU+CRF	84.01	82.11	83.05

We evaluate the performance difference between character representation and syllable representation with different model. In the table, most work focus on SCNN+CRF structure with different character representation. NoChar+SCNN+CRF model also give the comparable performance even though there is no character representation. CLSTM+SCNN+CRF model gives best result.

Table 4 shows the experimental results on different architecture of networks under same hyperparameters with Softmax inference layer. In most case, models with CRF inference layer can slightly improve sequence labelling models. But character information is not use in the model; Softmax based models give slightly better accuracies while the difference is not very significant. It is quite surprising that syllable-based CNN approach gives preferable performance based on our observations. CGRU+SCNN model gives the best result.

TABLE IV: Comparison and analysis of different architecture of network under the same hyperparameters

Model	Precision	Recall	F1
NoChar+SCNN	85.59	83.69	84.63
NoChar+SLSTM	82.48	81.38	81.93
NoChar+SGRU	83.74	83.79	83.77
CCNN+SCNN	84.72	81.90	83.28
CCNN+SLSTM	83.33	82.58	82.95
CCNN+SGRU	83.34	82.24	82.79
CLSTM+SCNN	85.00	83.22	84.10
CLSTM+SLSTM	83.96	81.81	82.87
CLSTM+SGRU	83.08	81.77	82.42
CGRU+SCNN	86.40	84.63	85.50
CGRU+SLSTM	84.83	82.88	83.84
CGRU+SGRU	84.35	82.84	83.58

4.6. Optimizers

The optimizer is responsible for minimization of the objective function of the neural network. A commonly selected optimizer is stochastic gradient descent (SGD), which provide itself as an efficient and effective optimization method for a large number of published machine learning systems. However, SGD can be quite sensitive towards the selection of the learning rate. Choosing a too large rate can cause the system to diverge in terms of the objective function, and choosing a too low rate results in a slow learning process. To eliminate the short comings of SGD, we explored other more sophisticated optimization algorithms such as Adagrad (Duchi et al., 2011), Adadelta (Zeiler,2012), RMSProp (Tieleman and Hinton, 2012) and Adam (Kingma and Ba, 2014). The experimental result can be found in Table 5.

TABLE V: Comparison and Analysis of Different Optimizers

Model	SGD	Adam	Adagrad	Adadelta	RMSProp
CCNN+SLSTM+CRF	83.15	73.00	82.45	80.30	69.28
CLSTM+SCNN+CRF	84.74	84.40	83.54	79.60	84.25
CGRU+SGRU+CRF	83.05	82.34	85.42	82.27	76.41

In Table 5, our results show that most of the optimizers such as SGD, Adam and RMSProp get the better result in SCNN based model but Adagrad outperforms all other optimizers in CGRU+SGRU+CRF model.

On the other hand, RMSProp gives the worst results in CCNN+SLSTM+CRF model. Adagrad and SGD have the competitive results but the difference to Adagrad and SGD are not very significant. SGD and Adagrad were on average the best optimizer. To be concluded, SGD and Adagrad are on average the best optimizer.

5. Conclusion

In this research, we consider stemming as a typical sequence tagging problem over segmented word, while segmentation also can be modelled as a syllable-level tagging problem via predicting the labels that identify the word boundaries. Our new approach proposed a simple and effective neural sequence labelling model for joint Myanmar word segmentation and Stemming. We investigate the several experiments to demonstrate joint model. Experiments show that SCNN based model gives the better result than other models. In most case, CRF inference layer outperform than Softmax layer but CGRU+SCNN+Softmax model gives the best result with SGD optimizer. During our experiments, we investigate on different optimizers. We observed that Adagrad and SGD optimizers give comparable result with CRF inference layer.

In the future work, we would like to increase the size of the manually segmented corpus. We intent to extend our joint model to perform similar task on different pre-trained word embedding and we will examine hyperparameter optimization like dropout rate, number of LSTM layers etc.

6. References

[1] Jie Yang and Yue Zhang. NCRF++: An Open-source Neural Sequence Labeling Toolkit. arXiv:1806.05626v2 [cs.CL] 17 Jun 2018.

[2] Pa. W.P., N.L.: "Myanmar Word Segmentation using Hybrid Approach.", presented at ICCA, Yangon, pp.166-170, 2008.

[3] W.P.Pa, Y.K.Thu, A.Finch and E.Sumita, "Word Boundary Identification for Myanmar Text Using Conditional Random Field", Springer, Switzerland, 2016

[4] Jie Yang, Shuailong Liang, and Yue Zhang. "Design challenges and misconceptions in neural sequence labeling". In COLING, 2018.

[5] Thu, Y.K., Finch, A., Sagisaka, Y., Sumita, E.: "A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation". In Proceedings of 12th International Conference on Computer Applications, Yangon, Myanmar, pp.167-179, 2014.

[6] Edouard Grave, Piotr Bojanowski, Prakhar Gupta1, Armand Joulin and Tomas Mikolov, "Learning Word Vectors for 157 Languages" arXiv:1802.06893v2, 28 Mar 2018

[7] G. Lev Ratinov and Dan Roth, "Design challenges and misconceptions in named entity recognition." In CoNLL, pages 147–155, 2009.

[8] Nils Reimers and Iryna Gurevych. "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks." 2017a,,arXiv preprint arXiv:1707.06799.

[9] Kingma.Diederik P and Ba.Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[10] Zeiler.Matthew D . Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[11] Xuezhe Ma and Eduard Hovy. "End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF". In ACL. volume 1, pages 1064–1074, 2016.

[12] Duchi, J., Hazan, E., and Singer, Y. "Adaptive subgradient methods for online learning and stochastic optimization". The Journal of Machine Learning Research, 2011.

[13] Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 4, 2012.