Chapter 23
Khmer Named Entity Recognition Based on LSTM-CRF Model

Lei Teng, Xin Yan, Jun Xie, Feng Zhou, Guangyi Xu, and Yuanyuan Mo

Abstract Named Entity Recognition(NER), as a significant research content of information extraction, is broadly used in Natural Language Processing fields such as information retrieval, machine translation, question answering system, and so on. Named Entity Recognition in Khmer plays a supportive part in the study of Chinese–Khmer bilingual understanding. However, the differences between languages make it hard to transfer the common Chinese and English Named Entity Recognition into Khmer. Aimed at the problem that the output of long short-term memory neural network model does not consider the sequence of tags, the output of long short-term memory neural network model and Khmer entity feature is used to input features of conditional random field model to extract Khmer corpus entity. The result of experiment shows that the precision of the conditional random field model is increased by 1.32%, recall rate by 2.55% and the $F1$ value which is used to measure the precision and recall rate by 1.92% after using long short-term memory neural network as the inputted feature of conditional random field model, which unites the output of neural network model with the entities in Khmer. The outcome of experiment proves the effectiveness of this method.

L. Teng · X. Yan (✉) · J. Xie · F. Zhou
Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
e-mail: Kg_yanxin@sina.com

Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China

G. Xu
Yunnan Nantian Electronic Information Industry Co., Ltd., Kunming 650040, China

Y. Mo
School of Southeast and South Asia Languages and Culture, Yunnan Minzu University, Kunming 650500, China

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.
L. C. Jain et al. (eds.), *Wireless Technology, Intelligent Network Technologies, Smart Services and Applications*, Smart Innovation, Systems and Technologies 258,
https://doi.org/10.1007/978-981-16-5168-7_23

23.1 Introduction

As a significant research content of information extraction, Named Entity Recognition is broadly used in information retrieval, machine translation, question answering system, and so forth. The study of Named Entity Recognition in Khmer plays an essential part in Chinese–Khmer bilingual understanding. In task of Khmer Named Entity Recognition, the lack of corpus resources and experts in Khmer linguistics makes the language processing research technology for the language lag behind. From the historical factors of Khmer language, we can see that Khmer language also has a variety of word formation problems, which makes its entity composition no single. Therefore, Named Entity Recognition in Khmer task is a more complex research work.

In research of Named Entity Recognition task, the main research methods are divided into rule-based, statistical model-based, and neural network-based methods.

Rule-based method: The main feature is to analyze corpus in a field, then ask experts in the language field to learn the corpus of the language, then summarize some characteristics of the language learned and some unique language rules of the language, and use rule matching and other methods to realize the Named Entity Recognition of the language at last. Example 1: Alfred et al. [1] presented a method on the basis of rules for Named Entity Recognition of Malay language. In this essay, they take part-of-speech tagging features method to select candidate entities from the corpus, and then identify entities according to the entity tagging in the manually tagged dictionary. Yan and Bi [2] also proposed a method on the basis of rule, which is mainly used for Vietnamese Named Entity Recognition. Through analysis of some characteristics of Vietnamese entities, they add artificial rules, and then use these rules to named entity recognition of Vietnamese. This kind of method needs to be written by linguists. This kind of work is not only time-consuming and labor-consuming, but also prone to conflict between rules due to rule inconsistency. Therefore, this method of Named Entity Recognition on the basis of rules is not the better method in practical work.

Research method based on Statistics: The main characteristics of this method is to integrate some features of artificial learning into models of statistical learning to realize entity recognition. The most important advantage of this method compared with the manual rule-making named entity recognition method lies in the use of easier model and less artificial annotation corpus for named entity recognition. Moreover, this method of machine learning entity features can not only be applied to a single language, but also to other languages that we have not been exposed to, we only need to linguistic this language Xi, we could have the model of Named Entity Recognition of languages, without any needs for experts to write rules for the language, saving resources. In this kind of statistical methods, the models which used for Named Entity Recognition contain Hidden Markov Model, maximum entropy, and conditional random fields. For instance, Yu et al. [3] used two-layer hidden Markov model to realize Chinese entity recognition. Zhang et al. [4] presented a multi feature cross entropy Chinese Named Entity Recognition model. By quoting a variety of external

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

constraint features, the semantic information of the model prototype was added to obtain heuristic knowledge, and then this heuristic knowledge was combined with the maximum entropy model to obtain Chinese named entities. Pan et al. [5] presented a Named Entity Recognition method for Khmer language that integrates entity characteristics. This method firstly makes use of word segmentation model and part-of-speech tagging model to extract morphological and speech characteristics of Khmer language and then combines the unique language characteristics of Khmer language manually tagged. Then, the entity recognition of Khmer language is realized through conditional random field algorithm. This method has gotten good results in task of named entity recognition in Khmer. Pan et al. [6] presented a method for Vietnamese Named Entity Recognition on the basis of conditional random field. The method is to extract terms and lines of speech as feature templates according to the language characteristics of Vietnamese itself and then use conditional random field model to realize Vietnamese named entity recognition. However, this kind of learning method needs to understand the semantics to develop the characteristics of exercising model. The quality of feature setting directly affects the quality of Named Entity Recognition task results.

Method in the basis of neural network: The method solves the problem of using artificial feature engineering to train model and increase human labor. The method based on neural network can learn the semantic features of languages from the original corpus and then extract the information from the semantic features through neural network model to achieve Named Entity Recognition. The common neural network models for Named Entity Recognition include recurrent neural network (RNN) [7], long short-term memory (LSTM) [8, 9], and convolutional neural network (CNN) [10, 11]. Many scholars have used these methods in Named Entity Recognition research. Gregoire et al. [12] used cyclic neural network model to solve sequence annotation, and this model is more useful than the original statistical model in sequence annotation. Jin [13] proposed a method of biomedical Named Entity Recognition in the basis of recurrent neural network, which integrates sentence vector and word vector containing sentence information into bi-directional long short-term memory (Bi-LSTM) neural network. The result of experiment shows that the accuracy rate of the method is 1.40% higher than that of the traditional RNN model. Maimaitiayifu [14] proposed a method of Uyghur Named Entity Recognition on the basis of Bi-LSTM plus convolutional neural network and conditional random field (Bi-LSTM-CNN-CRF) model, which uses whirly neural network to train character features in Uyghur words, not only word vector and word vector are used, but also the part-of-speech features manually marked are input into neural network in the form of vector. Furthermore, it has gotten good outcome in Uighur Named Entity Recognition. Lample et al. [15] proposed a single-layer bi-directional cyclic neural network plus conditional random field to deal with the task of Named Entity Recognition and then achieved good results in datasets of multiple languages. Huang et al. [16] used the Bi-LSTM and conditional random field (Bi-LSTM-CRF) model to label sequences, connect the input context features and the spelling features of words directly with the output of the Bi-LSTM neural network and obtained good results through experiments.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

To sum up, we can see that the deep learning method can abandon the traditional artificial feature engineering, thus saving a lot of human costs. After the practice of the research methods in Named Entity Recognition, the traditional Bi-LSTM model never consider order between the output tags when it obtains the complete context information, resulting in the poor effect of entity recognition. It is necessary to add conditional random field (CRF) model which can obtain the constraint rule from exercising data to ensure the correctness of the prediction tags. This essay presents a method of Named Entity Recognition on the basis of Bi-LSTM-CRF neural network [17]. The Bi-LSTM neural network model output and the unique entity features of Khmer are taken as the input features in CRF model; then, entity recognition of Khmer is realized by using CRF model in order to strengthen the influence of entity recognition of Khmer.

23.2 Khmer Named Entity Recognition on the Basis of Bi-LSTM-CRF

23.2.1 Khmer Entity Features

Khmer originated in the Cambodia countries of Southeast Asia. Due to historical reasons, Khmer has been influenced by many languages [18] in the process of development, resulting in the unique language characteristics of Khmer, which have the following impacts on the work of named entity recognition: First of all, Khmer vocabulary is different from English with clear boundary features, such as spaces, which leads to a high error rate in the work of word segmentation in Khmer, thus affecting the named entity recognition rate. Secondly, Khmer lacks the part-of-speech transformation features that play a great part in the recognition of naming entities, such as the feature that naming entities in English start with capital letters, which brings higher demands on mining deeper semantic information of Khmer. For the above reasons, it is very difficult for Khmer language to identify the named entity. However, linguists have found that the named entity of Khmer language still has some unique characteristics as follows.

Different from the position of Chinese entity deixis, the position of Khmer entity deixis in entity is mostly in the position of prefix. For example, the location of "ខេត្ត(province)" in the entity "Yunnan Province" is in front of the provincial name "Yunnan", that is to say, its composition is "ខេត្ត(province) យូណាន(Yunnan)";

The writing method of place names is similar to that in English. The place names and place names appear in the order of small to large, and the place names can appear next to each other. For example, the Khmer corresponding to "Pu'er, Yunnan Province" is "ទីក្រុង(city) Pu'er ខេត្ត(province) យូណាន(Yunnan)";

Different from the writing order of organization names in Chinese, organization names in Khmer are generally in the middle structure, and the attribute is in the post position, and the category of organization is in the position

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

of prefix. For example, the Khmer language corresponding to the organization name "Kunming University of technology" is " សាកលវិទ្យាល័យ(University) វិទ្យាសាស្ត្រនិងបច្ចេកវិទ្យា(technology) គុនមីង(Kunming)";

In the Khmer name representation, the appellation words should be placed in front of the name, for example, "Auntie", "Li". The title indicating the position should be placed before the name and the space should be added before the position and the name. For example, " សាស្ត្រាចារ្យលី" means "Professor Li". People's names cannot appear next to each other. They have to be divided due to spaces, punctuations, and no notional words, such as: in the sentence of " ហ៊ុនសែនព្រះមហាក្សត្រអ្នកដឹកនាំសំខាន់នៅក្នុងប្រទេសកម្ពុជា(Hun Sen and Sihamoni are important leaders in Cambodia)", the names of " ហ៊ុនសែន" and " ព្រះមហាក្សត្រ(Sihamoni)" are separated by spaces;

The place names usually need to be separated by commas before and after, but if the place names are embedded in the organization structure names, they can appear next to each other. For example, the organization name "Kunming charity" corresponds to Khmer language " គុនមីងសប្បុរសធម៌".

23.2.2 Bi-LSTM-CRF Model

The model in this paper for Named Entity Recognition is Bi-LSTM-CRF model (see Fig. 23.1). It adds a layer of conditional random field model after the original hidden layer of Bi-LSTM so that model can not only obtain context information of a single word, but still take full advantages of the annotation information before or after the sentence. Because complex entity structure of Khmer language is to enhance the influence from entity recognition of Khmer language, this essay uses the output of Bi-LSTM neural network and entity features of Khmer language as the input features of CRF model, which is similar to the use of maximum entropy features. Finally, the named entity recognition of Khmer language is realized by learning CRF.

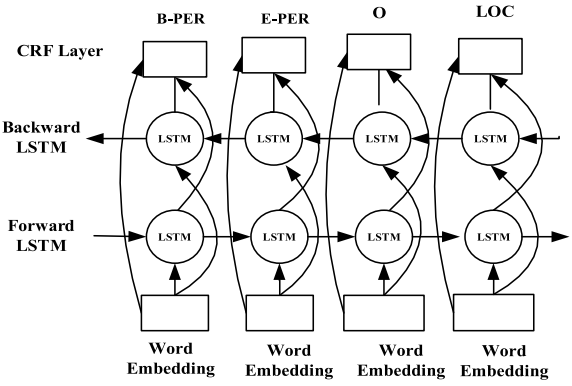

Fig. 23.1 Bi-LSTM-CRF mode

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

The process of Named Entity Recognition in the basis of Bi-LSTM-CRF network model is as follows:

23.2.3 Bi-LSTM Layer

In this essay, input layer of Bi-LSTM layer in Bi-LSTM-CRF network model is word vector, which can be expressed as $e = (e_1, e_2, \ldots, e_n)$. The hidden state of sequence in hidden layer of Bi-LSTM layer can be expressed as $H = \{h_1, h_2, \ldots, h_n\}$. According to the Bi-LSTM network structure chart, the hidden state sequence here is the forward hidden state sequence $\left(\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n}\right)$ and backward hidden state sequence obtained by reverse encoding $\left(\overleftarrow{h_n}, \overleftarrow{h_{n-1}}, \ldots, \overleftarrow{h_1}\right)$. In the forward representation sequence $\left(\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n}\right)$, $\overrightarrow{h_t}$ represents the forward notation of the t time of vector. The calculation is as follows:

$$\overrightarrow{h_t} = \mathrm{LSTM}\left[(e_n, e_{n-1}, \ldots, e_t)\right] \tag{23.1}$$

where LSTM$(\cdot)$ represents the calculation of the LSTM model, namely e $= (e_n, e_{n-1}, .., e_1)$, obtaining the backward notation sequence by backward coding $\left(\overleftarrow{h_n}, \overleftarrow{h_{n-1}}, \ldots, \overleftarrow{h_1}\right)$, $\overleftarrow{h_t}$ represents the forward of the t time of vector. The calculation is as follows:

$$\overleftarrow{h_t} = LSTM\left[(e_n, e_{n-1}, \ldots, e_t)\right] \tag{23.2}$$

The backward representation sequence receives the reverse sequence of the input vector as the input, where LSTM$(\cdot)$ represents the calculation of the LSTM model.

After the above sequence input, the hidden state representation sequence h_t of the hidden layer of Bi-LSTM network structure is composed of forward representation and backward representation: $h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]$, at this time, the hidden state sequence of the hidden layer can be obtained as $\{h_1, h_2, \ldots, h_n\} \in \mathbb{R}^{n \times m}$, where [;] is the connection.

23.2.4 CRF Layer

In Bi-LSTM-CRF network model mentioned in this essay, the CRF is mainly to assign markers to each word; then calculate the score of the whole marker sequence of the input sequence; obviously the score of the marker sequence is composed of two parts—one is the score of the word marker, the other is the score of the transfer

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

between the marker sequences; finally, select the final marker sequence with high score as the final part of this paper tag results.

Firstly, from the description of the above Bi-LSTM layer, we can see that the hidden state sequence output by the hidden layer of the Bi-LSTM network model can be shown as $\sqrt{(h_1, h_2, \ldots, h_n)}$. Then a linear layer needs to be connected to the hidden layer. The main function of linear layer maps the hidden state representation sequence from m dimension to k dimension, where k represents the number of labels in the annotation set used in this paper. In this paper, the annotation method of BIOES is used. The specific content is shown in Table 23.1.

Khmer corpus has been pre-processed by word segmentation, part-of-speech tagging, and so on. Taking the basic feature of word form and part-of-speech as one of the input features of conditional random field can enrich the semantic information of input features and improve the effect of Khmer entity recognition. The basic feature template designed in this paper is shown in Table 23.2.

The basic feature template simply describes the part-of-speech or morphology of the core word and its up and down words, and its semantic information is very limited.

Table 23.1 BIOES tagging set

Entity tag	Start tag	Middle	End	Single entity tag
Person name	B-PER	I-PER	E-PER	S-PER
Organization	B-ORG	I-ORG	E-ORG	S-ORG
Place	B-LOC	I-LOC	E-LOC	S-LOC
Non-entity mark	O	O	O	O

Table 23.2 Basic feature template

Index	Template form	Template meaning
1	Current word(0)	Current word
2	Current word(−1)	The first word on the left of the current word
3	Current word(−2)	The second word on the left of the current word
4	Current word(1)	First word on the right of current word
5	Current word(2)	The second word on the right of the current word
6	Current POS(0)	The part-of-speech of current words
7	Current POS(−1)	The part-of-speech of the first word on the left of the current word
8	Current POS(−2)	The part-of-speech of the second word on the left of the current word
9	Current POS(1)	The part-of-speech of the first word on the right of the current word
10	Current POS(2)	The part-of-speech of the second word on the right of the current word

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

According to the entity characteristics of Khmer entity mentioned above, the entity indicators are collected manually and the word list of them is constructed, as shown in Table 23.3. These entity indicators can provide rich entity information for Khmer named entity recognition, so as to enhance the impact of Khmer Named Entity Recognition.

Based on the rich entity features in Khmer provided above and the entity indicators listed in Table 23.3, the feature templates of entity information are established.

Then, the $n \times k$ fraction matrix output by the Bi-LSTM network structure through the linear layer is defined as P, where $P_{i,j}$ in the matrix P represents the fraction marked as the jth label in the ith word of a sentence.

After that, the output matrix P in Bi-LSTM neural network and Khmer entity features is taken as the input in CRF model. And annotation information of input sentence sequence is obtained through CRF model learning. In this case, for the input sequence: $(x_1, x_2, \ldots, x_n)$, the possibility of prediction tag: $(y_1, y_2, \ldots, y_n)$. Conditional probability is expressed as:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_k \lambda_k f_k(y, x)\right) \tag{23.3}$$

Among them,

$$Z(x) = \sum_y \exp\left(\sum_k \lambda_k f_k(y, x)\right) \tag{23.4}$$

Table 23.3 Khmer entity demonstrative words list

Entity classification	Pointer word
Person name	ព្រឹទ្ធបុរស (dean), សាស្ត្រាចារ្យ(professor), វេជ្ជបណ្ឌិត(doctor), គ្រូស្រី(female teacher), គ្រូប្រុសប្រុស(male teacher), មិត្តរួមថ្នាក់(classmate), ពូ(uncle) មីង(aunt) ជីតា(grandfather), យាយ(grandmother), ប្រធានាធិបតី(president), នាយករដ្ឋមន្ត្រី(prime Minister)
Place name	ខេត្ត(province), ប្រទេស(country), ទីក្រុង(city), ខោនធី(country), ស្រុក(district), ភូមិ(village), ផ្លូវហាយវ៉ែ(highway), ភ្នំ(mountain), ទន្លេ(river), បឹង(lake), សមុទ្រ(sea)
Organization name	សាកលវិទ្យាល័យ(university), វិទ្យាល័យ(high school), វិទ្យាល័យ(junior high school), សាលាបឋមសិក្សា(primary school), អង្គការ(organization), ក្រុមហ៊ុន(company), មន្ទីរពេទ្យ(hospital), គ្លីនិក(clinic), ស្ថាប័ន(organization)

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

Table 23.4 Khmer entity information feature template

Index	Template form	Meaning of templates
11	CUR_PER_SUF	Is current word a person name demonstrator
12	NEXT_PER_SUF	Does current word contain a person name indicator
13	CUR_LOC_SUF	Is current word a place name indicator
14	NEXT_LOC_SUF	Does two words on the left of the current word contain the place name indicator
15	CUR_ORG_SUF	Is current word an organization name indicator
16	NEXT_ORG_SUF	Does two words on the left of the current word contain the organization name indicator
17	CUR_PER_NAME	Is current word a common person name
18	CUR_LOC_NAME	Is current word a common place name
19	CUR_ORG_NAME	Is the current word a common organization name

$Z(x)$ expressed as normalization factor. λ_k get through training data. $f_k(y, x)$ represents the sum of transfer features and state features of each location.

Finally, the maximum likelihood estimation is used to calculate the parameter λ, and the Viterbi algorithm is to generate the best output sequence.

23.3 Experiments

23.3.1 Datasets

Khmer corpus needed for the experiment includes 138,396 Khmer words after segmentation and manual tagging. A Khmer corpus published by PAN localization Khmer (PCL), which contains 73,127 Khmer words and has been segmented and part-of-speech tagging, but the corpus does not have entity tagging. That is to say, the entities in the corpus need to be labeled manually. The remaining 66,269 Khmer words are divided tools. Finally, the Khmer language experts are invited to label all the corpus named entities to get corpus of Khmer named entity recognition. The annotation set for manual annotation mentioned in this essay is as same as shown in Table 23.4. 80% of the corpus is used as training set and rest of 20% as testing set. Proportion of datasets is shown in Table 23.5.

23.3.2 Evaluation Metric and Parameter Setting

The evaluation metric used in this paper is accuracy P, recall R and F_1 severally. This is the specific formula:

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

Table 23.5 Corpus of Khmer named entity recognition

Corpus classification	Total no. of words	No. of person names	No. of place names	No. of organization names
Total corpus	138,396	2672	3485	3027
Training set	110,717	2137	2788	2421
Test set	27,679	535	697	606

Table 23.6 Influence on different dropout parameters

Model	Accuracy (%)	Recall (%)	F_1 value (%)
Bi-LSTM-0.3	73.05	73.13	73.08
Bi-LSTM-0.5	73.03	73.16	73.09
Bi-LSTM-0.6	73.01	73.19	74.10

$$P = \frac{\text{Number of named entities correctly recognized}}{\text{Total number of named entities identified}} \times 100\% \tag{23.5}$$

$$R = \frac{\text{Number of named entities correctly recognized}}{\text{Total number of named entities in Corpus}} \times 100\% \tag{23.6}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{23.7}$$

The super parameter setting in the neural network model has a great impact on the implementation in named entity recognition task. Then, the text sets the super parameters as follows:

In this paper, word embedding is used to train fixed dimension word vectors. This method can capture semantic and syntactic information between word and can maintain semantic relationship between words. The dimension of word vector is set to 100. The word vector is trained by the word2vec tool published by Google. The selected optimization algorithm is the most widely used stochastic gradient descent (SGD) [19] on Bi-LSTM neural network, with the learning rate set to 0.01, and dropout parameter set to 0.3, 0.5, and 0.6 for comparison. The specific experimental data is shown in Table 23.6. In this paper, the dropout parameter is set to 0.6.

23.3.3 Experimental Analysis

To check the rationality of proposed model and the necessity of each module in model effectively, this paper conducts four groups of comparative experiments on the Khmer named entity recognition task, and these four experiments are performed in the corpus composed of above. The evaluation method is still the above-mentioned accuracy, recall rate, and F_1 value.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

Experiment 1: The main objective of this experiment is to use named entity recognition result of CRF model as a standard, then to verify the performance difference of recurrent neural network model contrasted with the traditional statistical model. In experiment, CRF ++ is used as the tool to realize CRF model, and the basic features composed of morphology and part-of-speech are used as input features. The model accuracy is 70.85%, recall rate is 71.98%, and value of F_1 is 71.41%.

Experiment 2: This experiment is to prove the influence of CRF model on the recognition of Bi-LSTM neural network model. Main function of CRF is to optimize output result label of neural network according to the relationship between adjacent labels. The specific operation of the experiment is to add a CRF layer after Bi-LSTM neural network model to output optimal tag sequence. The specific experiment data is shown in Table 23.7.

From Table 23.7, the recognition effect of neural network model is significantly improved after adding CRF. For the Bi-LSTM neural network model, after adding CRF model, its F_1 value is increased by 1.28%. Compared with the basic experiment 1, its F_1 value increased by 6.48%. Therefore, it can be concluded that adding the linear CRF module is helpful for improving the recognition rate of the model.

Experiment 3: The objective of experiment is to verify the effect of Khmer entity features on Khmer entity recognition of Bi-LSTM-CRF neural network model. The specific method is to take output of Bi-LSTM neural network model and Khmer entity features as the input features of CRF model for Khmer named entity recognition. The outcome of this experiment is shown in Table 23.8.

From the above table, the CRF model with entity features significantly improves the recognition influence of the neural network model. For the Bi-LSTM-CRF neural network model used in this essay, after using Bi-LSTM neural network, the output of the neural network model and the entity features in Khmer language are taken as the input features of the CRF model; accuracy rate, recall rate, and F_1 value are increased by 1.32%, 2.55%, and 1%, respectively, 1.92%. Compared with the basic experiment 1, accuracy, recall, and F_1 values of this method are increased by 7.83%, 8.99%, and 8.4%, respectively. Therefore, it can be concluded that adding Khmer entity features to CRF model can enhance the recognition rate of the model.

Table 23.7 Effect of adding CRF model on model performance

Model	Accuracy (%)	Recall (%)	F_1 value (%)
Bi-LSTM	75.67	77.58	76.61
Bi-LSTM-CRF	77.36	78.42	77.89

Table 23.8 Effect of adding Khmer entity features to CRF model on model performance

Model	Accuracy (%)	Recall (%)	F_1 value (%)
Bi-LSTM-CRF	76.89	78.03	77.46
Bi-LSTM + CRF	78.68	80.97	79.81

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

23.4 Conclusion

In the chapter, the Bi-LSTM-CRF model is proposed for Named Entity Recognition of Khmer. Among them, on the basis of Bi-LSTM-CRF model, the output of Bi-LSTM neural network and entity features of Khmer are taken as the inputted features of CRF model, which enriches the input of CRF model and improves effectiveness of model in the recognition of Khmer named entities. However, the training time is too long due to the effect of internal parameters of the model. How to optimize the model and shorten the training time is the next problem to be studied. The research results of this paper are not only applied to the study of Named Entity Recognition in Khmer, but also have an important reference for other languages.

Acknowledgements This work is supported by National Nature Science Foundation under Grant: No. 61562049 and No. 61462055.

References

1. Alferd, R., Leong, L.C., On, C.K., et al.: Malay named entity recognition based on rule-based approach. Int. J. Mach. Learn. Comput. **4**(3), 300–306 (2014)
2. Yan, D.H., Bi, Y.D.: Rule-based recognition of vietnamese named entities. J. Chin. Inf. Process. **28**(5), 198–205 (2014)
3. Yu, H.K., Zhang, H.P., Qun, L., Lv, X.Q., Shi, S.C.: Chinese named entity identification using cascaded hidden Markov model. J. Commun. **27**(2), 87–94 (2006)
4. Zhang, Y.J., Xu, Z.T., Xue, X.Y.: Fusion of multiple features for Chinese named entity recognition based on maximum entropy model. J. Comput. Res. Dev **45**(6), 1004–1010 (2008)
5. Pan, H.S., Xin, Y., Yu, Z.T., Guo, J.Y.: A Khmer named entity recognition method by fusing language characteristics. J. Chin. Inf. Process (1), 4003–4007 (2014)
6. Pan, Q.Q., Zhou, F., Yu, Z.T., Guo, J.Y., Xian, Y.T.: Recognition method of Vietnamese named entity based on conditional random fields. J. Shandong Univ. (Nat. Sci.) **49**(1), 76–79 (2014)
7. Rafal, J., Wojciech, Z.: An empirical exploration of recurrent network architectures. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 2342–2350 (2015)
8. Hu, Z.T., Ma, X.Z., Liu, Z.Z., Eduard, H., Eric, P.: Harnessing deep neural networks with logic rules. In: Proceedings of ACL-2016, pp. 21–30 (2016)
9. Wu, Y.G., Jiang, M., Lei, J.B., et al.: Named entity recognition in Chinese text using deep neural network. Studies in Health Technology and Information, pp. 624–628 (2015)
10. Santos, C.D., Guimaraes, V., Niteroi, R.J., Janeiro, R.D.: Boosting named entity recognition with neural character Embedding's. In: Proceedings of NEWS 2015 The Fifth Named Entities Workshop, pp. 25–30 (2015)
11. Zhang, J., Song, Y., Zhang, C., Liu, S.: Evolutionary hierarchical Dirichlet processes for multiple correlated time varying corpora. In: Proceeding of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, pp. 1079–1088 (2010)
12. Gregoire, M., Yann, D.H., et al.: Using recurrent neural networks for slot filling in spoken language understanding EEE/ACM. Trans. Audio, Speech, Lang. Process 530–540 (2015)
13. Jin, L.K.: Named entities recognition based on recurrent neural network in biomedical literature. Dalian University of Technology, pp. 1–76 (2016)

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.

14. Maimaitiayifu, Silamu, W., Muhetaer, P., Yang, W.Z.H.: Uyghur named entity recognition based on Bi-LSTM-CNN-CRF model. Comput. Eng. (44), 230–236 (2018)
15. Lample, G., Ballesteros, M., Subramanian, S., et al.: Neural architectures for named entity recognition. ArXiv preprint arXiv:1603.01360 (2016)
16. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. ArXiv preprint arXiv:1508.01991 (2015)
17. Chen, G.Q.: On the evolution of Wa prepositions in Khmer. Minority Lang. Chin. (4), 32–37 (1999)
18. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. ArXiv preprint arXiv:1310.4546 (2013)
19. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, Physica-Verlag HD, pp. 177–186 (2010)

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.