A Khmer NER method based on conditional random fields fusing with Khmer Entity Characteristics Constraints

Yu Nuo[1], Xin Yan[1], Zhengtao Yu[1] , Shuhui Huang[1],Jianyi Guo[1]

1.School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650504,China

E-mail: kg_yanxin@sina.com

Abstract. In order to improve the performance of Khmer named entity recognition(NER), a NER method based on conditional random field (CRF) model fusing with Khmer entity characteristics constraints is proposed in this paper. First of all, we carried out analyses on the Khmer entity characteristics, summarized the constraint on these entity characteristics and introduced into the CRF; then solved the labeling sequence by integer linear programming integrated entity characteristic constraint and obtained a model of CRF integrated with constraints based on Khmer entity characteristics. Based on a contrastive experiment, CRF model of the constraint has a better performance than traditional CRF model when carrying out the Khmer NER.

Keywords: CRF integrated with constraints; Khmer; Entity characteristic; Integer linear programming; Named entity recognition

1 Introduction

With the increasing frequency of Sino-Cambodia cooperation and further development of relations, the Khmer NER has become extremely significant. The NER is the basic task in the natural language processing field and even more the precondition for researches in the natural language application field. At present, the NER technology in English field has been mature relatively; as for Khmer, however, relatively scare corpus resources, lower degree of popularization and less Khmer experts result in less processing research work for Khmer and comparatively laggard technology, especially the research on the NER becomes much weaker. Accordingly, it is urgently required to put more efforts into researches on the Khmer NER based on existing Khmer processing task, so as to improve performances of the NER and facilitate subsequent natural language processing researches.

Named entity is an important information-carried language unit of documents, the recognition and analysis of named entity plays a vital role in the extract of network information and the acquisition of bilingual comparable corpus. For researches on the NER, scholars at home and abroad have made a plenty of work, mainly involving two methods: the first is based on artificial rules, Farmakiotou [6] et al carried out researches on the NER on Greek finance documents by dictionary resource acquired artificially during research. Alfred [7] et al proposed a NER algorithm based on rules and utilized Malay POS and context features to recognize three kinds of named entities (PER, LOC and ORG) in Malay documents. In English field, the most typical NER systems based on rule include LaSIE-II system [1] of Sheffield University and LTG system [2] of Edinburgh University, as well as NTU system [3], OKI system [4] and ACILE system [5], etc. But this method has its disadvantages, including high cost, performances of the system mainly depend on experienced philologists, huge amounts of works and difficult transfer. The second method is based on machine learning (ML), including decision tree (DT), hidden markov model (HMM), maximum entropy (ME), support vector machine (SVM) and conditional random field (CRF), etc. Zhou [8] et al built up a block labeler for NER by HMM, Isozaki et al [9] converted the multi-classification problem such as NER into a binary classification problem and Chieu [10] utilized the information of the whole document to realize a ME-based English NER system. Zhang et al [11] put forward a role-labeling method for recognition of a named entity so as to recognize Chinese named entities. Qin et al[12] regarded the task of NER as a classification problem and adopted the decision tree to recognize the named entity. Goh et al [13] made use of SVM to automatically recognize Chinese names and ORGs. Lu et al [14] used CRF model to evaluate the recognition performance through cross validation and guided the selection of features by recognition performance and validated the advantage of CRF model over the product entity recognition. Qiu et al [15] proposed a method for Chinese NER research based on CRF model fusing with the feature which is integrated with POS and word boundary. Yan et al [16] proposed a method for NER in Chinese medical records based on cascaded CRFs of which the first layer is simple NER and the feature set of the second layer is together with the character features, word boundary features and context features in a sentence, compared to single layer CRF method,

This work is supported by National Nature Science Foundation of China(NSFC) via grant 61462055,61562049,61662040

a significant increase in overall performance. In recent years, NER has become more mature. As for languages with scarce labeling corpus, some scholars attempted to help them by incorpusting cross-lingual features and knowledge bases from English using cross-lingual links, such as Kazama [17] et al acquired named entity knowledge by Wikipedia multi-language entry links, Fu [18] et al utilized large-scale English-Chinese bilingual parallel corpus and transferred English named entity labels to Chinese words by word alignment technology, filtered noise by rules to obtain large-scale Chinese named entity training corpus and achieve a good performance. For the recognition of PER, LOC and ORG in Khmer, because of its inherent features of Asian language family, Khmer is lack of obvious characteristics of named entity, named entity labeling corpus is scarce and expensive and it is difficult to obtain enough bilingual resource to assist Khmer NER; additionally, the existing CRF model fails to sufficiently make use of the logical relations of Cambodian entity formations. Aiming at conditions above, in order to improve the recognition performances and relieve the corpus insufficiency, the NER is performed by analysis of language characteristics and conformation of CRF model integrated with constraints based entity characteristics. Because the Vitervi algorithm for traditional CRF model decoding is only able to deal with simple constraints during decoding, we adopt the integer linear programming method for decoding with the entity constraint. The comparative experiment shows that the model proposed in this paper is able to effectively improve the performance of Khmer NER.

2 CRF Model Integrated with Khmer entity characteristics constraints

2.1 Analysis of Khmer Entity Characteristics

Khmer named entities have their own characteristics as follow:

(1) The locations of LOCs appear in the ascending sequence, names of small and big places can appear adjacently. For example, កុនមីង (Kunming City) appears before ខេត្តយុនណាន (Yunnan Province) in the sentence កុនមីងខេត្តយុនណាន(Kunming City, Yunnan Province), where ខេត្ត (Province) is before យុនណាន(Yunnan);

(2) PER and LOC can not appear adjacently: when PER and LOC appear simultaneously, they must be separated by a single space, punctuation mark or non-entity word, for example, PER ហ ុនសែន (Hun Sen) and LOC រាជធានីភ្នំខេញ (Phnom Penh) do not appear adjacently in the sentence ហ ុនសែនខៅរាជធានីភ្នំខេញ (Hun Sen is in Phnom Penh), an entity word ខៅ (in) exists between them;

(3) PER and ORG can not appear adjacently: when PER and ORG appear simultaneously, they must be separated by a single space, punctuation mark or non-entity word, for example, PER ចាន់ មុននី (Zhan Moni) and ORG តុ ុលាការប្រជាជនកប្ិតុ្ឋម (the Intermediate People's Court) do not appear simultaneously in the sentence ចាន់ មុននីក្តងការងារតុ ុលាការប្រជាជនក្ឋិត ុ្ឋម (Zhan Moni works for the Intermediate People's Court).

(4) PERs can not appear adjacently: when PERs are enumerated, they shall be separated by single space, or they are separated by non-entity word when several PERs appear simultaneously, for example, PERs ហ ុនសែន (Hun Sen) and នរោត្តមសីហនុ (Norodom Sihanouk) are separated by single space in the sentence ហ ុនសែន នរោត្តមែ ីហនុបានចូលរួមក្តងកុ ិចចប្រ ជុំខនេះ (Hun Sen and Norodom Sihanouk attend the meeting);

(5) ORGs can not appear adjacently: when ORGs are enumerated, they shall be separated by single space, or they must be separated by non-entity word when several ORGs appear simultaneously, for example, ORGs សាកលវិទ្យាល័យឡ៊ីប្ ុងឌរ៉ែកាំង (Peking University) and សាកលវិទ្យាល័យប្ ុងសា ងហាហ (Shanghai University) are separated by single space in the document សាកលវិទ្យាល័យឡ៊ីប្ ុងឌរ៉ែកាំង សាកលវិទ្យាល័យប្ ុងសា ងហាហ(Peking University, Shanghai University);

(6) LOC and ORG can not appear adjacently: when LOC and ORG appear simultaneously, they must be separated by a single space, punctuation mark or non-entity word, for example, មនទីរខេទ្យប្រជាជនគុនមីញ(Kuiming People's Hospital) and ខេត្តយូណាន(Yunnan Province) are separated by non-entity word ក្តង (in) in the sentence មនទីរខេទ្យប្រជាជនគុនមីញក្តងខេត្តយូ ណាន(Kuiming People's Hospital in Yunnan Province).

Under the circumstance of lacking of a large amount of training corpus, it is difficult to acquire the above language characteristics through CRF model training. When recognizing named entities in a Khmer document, if the characteristic of

entity information above is integrated into the recognition model of named entities as a constraint, we can adequately make use of the Language characteristics of Khmer to reduce error labels of recognition model and improve the recognition performance to some extent. In order to fully utilize these characteristics, we shall consider the improvement of decoding algorithm to raise the recognition performance of named Khmer entities.

2.2 CRF Model Integrated with Constraint

Usually, the NER is converted into sequence labeling problem, while the CRF has better performances when solving the sequence labeling problem, it can overcome such disadvantages as labeling bias of maximum entropy (ME) during labeling and independent assumption of hidden markov model (HMM), etc, so CRF model is adopted in this paper to recognize named entities in Khmer. When recognizing named entities in document in specific languages and fields, traditional CRF model is difficult to fully utilize this language and the particular characteristic or constraint in this field. Considering the sufficient utilization of Khmer entity characteristics, therefore, a CRF model integrated with Khmer NER constraints is employed in this paper.

Linear chain CRF is defined, G=<X, Y>, where, X indicates observation sequence, Y indicates labeling sequence. A logical relation exists between Yi and Yj, namely constraint based on entity characteristic, so we call the CRF model integrated with entity information constraints as CRF model for constraint.

The probability distribution for given observation sequence x=x1, x2, x3…xn and constraint condition c and labeling sequence y=y1, y2, y3…yn is as follows:

$$p(y \mid x) = \frac{1}{Z(x,c)} \exp\left(\sum_{i=1}^{N} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, c, i)\right) \quad (1)$$

where, fk indicates feature function, constraint condition c refers to the constraint based on Khmer entity characteristics, the normalization factor Z(x, c) is expressed as follow:

$$Z(x,c) = \sum_{y} \exp\left(\sum_{i=1}^{N} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, c, i)\right) \quad (2)$$

When estimating parameters of this model, the maximum likelihood estimation method is adopted to estimate the weight of characteristic function θ= {λk}.

2.3 Decoding

A traditional CRF model is usually decoded by Viterbi algorithm, while the existing Viterbi algorithm is only used to deal with simple constraints during decoding. For some complex constraint relations, for example, the decoding of constraints based on entity information, which have been added in entities, is difficult to achieve on the basis of Viterbi algorithm. In order to sufficiently make use of these constraints for Khmer NER document, we build Constrained Conditional random field (CCRF) replace Veterbi algorithm for decoding with the integer linear programming method, during which Khmer entity information constraint is introduced.

In fact, Viterbi decoding process of CRF is just to solve the shortest path sequence in the graph. We regard the inference of CRF as the shortest path problem and then the integer linear programming method can be adopted for decoding CRF, with specific method as follows: First, a graph G is constructed, where n indicates the number of observation sequence x, m indicates the number of labels y for each element in the observation sequence x, each node is recorded as vij (1≤i≤n, 1≤j≤m), which means that yj is the label for xi, directed edge xi,jj′(1≤i≤n, 1≤j,j′≤m) indicates a directed edge from node v(i-1)j to vij′, and an additional starting point(start) and an additional end point(end) of the path are added to the head and tail of the sequence. In the integer linear programming process of finding the shortest path sequence, the indegree sum of all nodes(except start node and end node) equals to the outdegree sum of all nodes(except start node and end node), while the outdegree sum of start node equals to 1 and the indegree sum of end node is 1, and the value of all the nodes can only be 0 or 1, as shown in formula (3), (4), (5) and (6).

$$\min - \sum_{\substack{1 \le i \le n \\ 1 \le j, j' \le m}} lb(M_i(j, j')) x_{i,jj'}$$

$$\text{subject to:}$$

$$\sum_{1 \le j_1 \le m} x_{i-1,j_1 j} - \sum_{1 \le j_2 \le m} x_{i-1,jj_2} = 0 \ (1 \le i \le n, 1 \le j \le m); \quad (3)$$

$$\sum_{1 \le j \le m} x_{1,0j} = 1; \quad (4)$$

$$\sum_{1 \le j \le m} x_{n+1,j(n+1)} = 1; \quad (5)$$

$$x_{1,0j}, x_{i,j_1 j}, x_{n+1,j(n+1)} \in \{0,1\} \ (1 \le i \le n, 1 \le j, j_1 \le m). \quad (6)$$

From literature [19], there must be an integer solution for the shortest path problem, and interior-point algorithm can be used for solving the integer programming within the polynomial time. If any other constraint is never integrated during the calculation, the shortest path labeling sequence can be got from decoding by Viterbi algorithm. When some complex constraints are fused, however, the path labeling sequence is obtained difficultly by Viterbi algorithm, so integer linear programming is used to decode the CRF fusing with Khmer entity characteristics constraints in this paper.

In order to integrate the above entity constraints into the integer linear programming well, first of all, these characteristics are expressed as Boolean expression, and then the Boolean expression is converted into a corresponding linear set of inequalities which is introduced into the integer linear programming for solution. The Boolean expression correspondent with the entity information constraint and linear set of inequalities correspondent with the integer linear programming are listed in table 1, and respective explanations of these sets of inequalities are presented below. Here, we only set out some common corresponding characteristic constraint inequalities, however, these inequalities are not only limited to those set out in Table 1 and also expanded later.

Boolean variables a, b, c respectively indicates whether PER label La, LOC label Lb and ORG label Lc appears or not, a, b, c∈{0,1}, as shown in formula (7)

$$a=\begin{cases}1, & L_a \text{ apears}\\ 0, & L_a \text{ does not apear}\end{cases} \quad b=\begin{cases}1, & L_b \text{ apears}\\ 0, & L_b \text{ does not apear}\end{cases} \quad c=\begin{cases}1, & L_c \text{ apears}\\ 0, & L_c \text{ does not apear}\end{cases} \tag{7}$$

Table 1. Boolean expressions corresponding to the Khmer entity characteristics

number	Boolean expression	meaning	Linear inequality
1	$v(b_1)<v(b_2)$	If different LOCs with no space separated appears successively, the LOC with the minimum value is arranged before that with the maximum value	$\left[v(b_1)<v(b_2)\right]\geq \sum_{1\leq e\leq m} x_{i,eb_1}+\sum_{1\leq j\leq m} x_{i+2,b_2j}+s(b_2)-2$ $(2\leq i\leq n)$
2	$a_1 \rightarrow \overline{a}_2$	In the absence of space-separated, the two PER labels （La1,La2）can not appear continuously	$m\left[x_{i,a_1e}+s(a_1)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i-2,ja_2})$ $(1\leq e\leq m,3\leq i\leq n)$
3	$a \rightarrow \overline{b}$	In the absence of space-separated, Lb can not appear in front of La	$m\left[x_{i,ae}+s(a)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i-2,jb})$ $(1\leq e\leq m,3\leq i\leq n)$
		In the absence of space-separated, Lb can not appear at the back of La	$m\left[x_{i,ea}+t(a)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i+2,bj})$ $(1\leq e\leq m,2\leq i\leq n)$
4	$c_1 \rightarrow \overline{c}_2$	In the absence of space-separated, the two PER labels （Lc1,Lc2）can not appear continuously	$m\left[x_{i,c_1e}+s(c_1)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i-2,jc_2})$ $(1\leq e\leq m,3\leq i\leq n)$
5	$a \rightarrow \overline{c}$	In the absence of space-separated, Lc can not appear in front of La	$m\left[x_{i,ae}+s(a)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i-2,jc})$ $(1\leq e\leq m,3\leq i\leq n)$
		In the absence of space-separated, Lc can not appear at the back of La	$m\left[x_{i,ea}+t(a)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i+2,cj})$ $(1\leq e\leq m,2\leq i\leq n)$
6	$b \rightarrow \overline{c}$	In the absence of space-separated, Lc can not appear in front of Lb	$m\left[x_{i,be}+s(b)-1\right]\leq \sum_{1\leq j\leq m}(1-x_{i-2,jc})$ $(1\leq e\leq m,3\leq i\leq n)$

		In the absence of space-separated, Lc can not appear at the back of Lb	$m\left[x_{i,eb}+t(b)-1\right]\le\sum_{1\le j\le m}(1-x_{i+2,cj})$ $(1\le e\le m,2\le i\le n)$

At the same time, LOC in Khmer can be roughly divided into four categories, namely, provincial, municipal, district and village names, then we assign each type of LOC element with value, of which the provincial name value is the maximum, following municipal name value, district name value, and the village name value is the minimum, so as to create a LOC level list. In Table 1, a1 and a2 are Boolean variables which represent the two PERs appear or not, b1 and b2 are Boolean variables which represent the two LOCs appear or not, similarly as c1 and c2.

Where, Boolean expression of constraint 1 is equivalent to "if places with different levels appear successively with no space separated, the place with the minimum level is arranged before that with the maximum level", namely, places without single space bar appear according to the ascending sequence. Its Boolean expression can be transferred as $b_1\wedge b_2\wedge s(b_2)\rightarrow[v(b_1)>v(b_2)]$ because $x_4\ge x_1+x_2+x_3-2$ can be derived by logical expression $x_1\wedge x_2\wedge x_3\rightarrow x_4$, the logical expression corresponding to the integer linear programming can be expressed as $[v(b_1)>v(b_2)]\ge b_1+b_2+s(b_2)-2$, accordingly the corresponding linear inequality of constraint 1 in Table 1 can be derived.

Boolean expression of constraint 2 is equivalent to "if L_{a_1} appears, no single space is available before L_{a_1} in the original document, L_{a_2} must not appear in the adjacent position before L_{a_1}", namely, PERs can not appear adjacently, and multiple PERs shall be separated by space. Its corresponding Boolean expression is $x_{i,a_1e}\wedge s(a_1)\rightarrow\neg x_{i-2,ja_2}$, where s (a) indicates whether there is any single space before the word labeled as La in the original document, s (a) = 0 if any single space is available and s (a) =1 if no single space is available. Additionally, $x_3\ge x_1+x_2-1$ can be known from the logical expression $x_1\wedge x_2\rightarrow x_3$, so the corresponding linear inequality of constraint 2 in Table 1 can be derived.

Boolean expression of constraint 3 is equivalent to "In the absence of space-separated, Lb can not appear in the adjacent position before or after La.", so it can be equivalent to "Lb can not be in the adjacent position before La in the case of no space" and "Lb can not be in the adjacent position after La in the case of no space.". Where the corresponding Boolean logical expression of "Lb can not be in the position before La in the case of no space" in the integer linear programming can be expressed as $x_{i,ae}\wedge s(a)\rightarrow\neg x_{i-2,jb}$, and $x_{i,ae}\wedge s(a)\rightarrow 1-x_{i-2,jb}$ can be known from $\overline{x}=1-x$ because $x_3\ge x_1+x_2-1$ can be derived by logical expression $x_1\wedge x_2\rightarrow x_3$, accordingly corresponding inequality in Table 1 can be known; while "Lb can not be in the adjacent position after La in the case of no space" can be expressed as Boolean logical expression $x_{i,ea}\wedge t(a)\rightarrow\neg x_{i+2,bj}$, where t (a) indicates whether there is any single space after the word labeled as La in the original document, t (a) =1 if no single space is available, otherwise t (a) =0. Similarly the corresponding inequality can be derived.

Boolean expression of constraint 4 is equivalent to "if L_{c_1} appears and no single space is available before L_{c_1}, L_{c_2} can not appear in the adjacent position before L_{c_1}", namely, ORGs can not appear adjacently, and multiple ORGs shall be separated by space. Its corresponding Boolean logical expression is expressed as $x_{i,c_1e}\wedge s(c_1)\rightarrow\neg x_{i-2,jc_2}$, $x_3\ge x_1+x_2-1$ can be derived by $x_1\wedge x_2\rightarrow x_3$, accordingly the corresponding linear inequality of constraint 4 in Table 1 can be derived.

The corresponding Boolean expression of constraint 5 is equivalent to "if La appears, Lc can not appear in the adjacent position before and after La in the absence of space-separated", so it can be equivalent to "Lc can not in the adjacent position before La in the absence of space-separated" and "Lc can not be in the adjacent position after La in the absence of space-separated". The conversion process of corresponding linear inequality is similar to that of constraint 3, which can be solved according to the method for constraint 3.

The corresponding Boolean expression of constraint 6 is equivalent to "if Lb appears, Lc can not appear in the adjacent position before and after Lb in the absence of space-separated", so it can be equivalent to "Lc can not be in the

adjacent position before Lb in the absence of space-separated" and "Lc can not be in the adjacent position after Lb in the absence of space-separated". The conversion process of corresponding linear inequality is similar to that of constraint 3, which can be solved according to the method for constraint 3.

The linear set of inequalities above is introduced into the integer linear programming inequality for solving the labeling sequence, so as to realize the constraints based on Khmer entity characteristics and finally obtain a CRF model integrated with constraints based on entity characteristics.

3 Experiment and Result Analysis

3.1 Experimental Data

Two corpus are required for the experiment, the corpus includes training data and testing data. Firstly we obtain Khmer web documents through crawling, secondly CRF model proposed by Pan [20] et al is utilized for labeling segment and POS in these documents, the result is corrected by native Khmer speakers, and then this corpus and a open corpus KCorpus (Khmer Tagged Corpus) released by the PLC (PAN Localization Cambodia)[1] constitute a whole corpus, finally we ask these natives to label the named entities of this corpus to become Khmer entity corpus. This corpus contains 121316 words and 5861 named entities, including 1414 PERs, 2905 LOCs and 1536 ORGs. The testing data is 20% of Khmer entity corpus which is chosen randomly. The rest of Khmer entity corpus is the training data, and we only choose a half of the training data randomly in Baseline Experiment. The experimental feature template of CRF is as shown in Table 2

3.2 Evaluation Metrics

precision (P for short), recall (R for short) and F1 value are taken as evaluation metrics for NER, which are defined as follows: given A indicates the number of designated data elements really existing in these documents, B indicates the number of data elements labeled correctly in these documents and C shows that labeled by error, the calculation formulas of precision, recall and F1 value are as follows:

$$Precision = \frac{\text{the number of data elements labeled correctly}}{\text{the number of data elements labeled}} = \frac{B}{B+C} \times 100\%$$

$$Recall = \frac{\text{the number of data elements labeled correctly}}{\text{the number of all designated data elements}} = \frac{B}{A} \times 100\%$$

$$F_1 \ value = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

3.3 Experimental design and analysis

In order to evaluate more objectively the effect on of these constraints, this paper sets up three groups of contrast experiments as follow:

Experiment 1, Comparison of the performance of NER by using CRF model and CCRF model integrated with the entity characteristics

In order to verify the feasibility and effectiveness of the proposed identification model in the Khmer NER, we set up a set of comparative experiments as follow:

(1) Baseline Experiment: Using a half of the training data (80% of Khmer entity corpus) .

(2) CRF model: Using the whole training data (80% of Khmer entity corpus) as training data trained by CRF model. Compared with the Baseline experiments, it is better to quantify the performance of different size training data in the recognition of named entities.

(3) CCRF model: Using (2)'s corpus as the training data trained by CCRF model integrated with entity characteristic. Compared with experiment (2), verify the feasibility of entity feature constraint in NER. Compared with the Baseline experiment, verify and quantify the performance of Khmer entity corpus and the CCRF model integrated with entity characteristic for recognizing the named entity.

With the same testing data, the experimental results are shown in Table 3.

Base on the experimental results, the experiment using CRF model is trained with a larger size corpus which has a variety of entities, so that it's entity words and context information are much richer and the precision and the recall are slightly improved compared with the Baseline experiment. In the case of using the same training data, recognizing the named entity, integrated with the Khmer specific entity characteristics, has a significant improvement in the precision and recall of the experiment.

Experiment 2, The Comparison of the performances on the recognition of different entity categories by using CCRF model integrated with entity characteristics

In the Khmer NER, categories of entities can be roughly divided into three categories: PER, LOC, ORG. In order to analyze the differences of the

[1] http://www.panl10n.net/pan-localization-cambodia-plc-of-idrc-cambodia/#comment-10

methods proposed in this paper, the precision and the recall and F1 value of the three categories of entities are shown in table 4.

.

Table 2. Feature template of CRF

number	Template format	Template meaning
1	CurrentWord(0)/CurrentPOS(0)	The current word and its POS
2	CurrentWord(0)/CurrentWord(1)	The current word and its first word on the right
3	CurrentWord(0)/CurrentWord(-1)	The current word and its first word on the left
4	CurrentWord(-1)/CurrentPOS(0)	The current word's first word on the left and its POS
5	CurrentWord(0)/CurrentPOS(1)	The current word and the POS of its first word on the right
6	CurrentWord(-1)/CurrentPOS(-1)	The first word on the left of the current word and the POS of its first word on the left
7	CurrentWord(-1)/CurrentWord(-2)	The first word on the left of the current word and its second word on the left
8	CurrentWord(-2)/CurrentPOS(-2)	The second word on the left of the current word and the POS of its second word on the left
9	CurrentWord(1)/CurrentWord(2)	The first word on the right of the current word and its second word on the right
10	CurrentWord(-1)/CurrentWord(1)	The first word on the left of the current word and its first word on the right
11	CurrentWord(1)/CurrentPOS(0)	The first word on the right of the current word and the POS of it
12	CurrentPOS(-2)/CurrentPOS(-1)	The POS of the second word on the left of the current word and the POS of its first word on the left
13	CurrentPOS(-2)/CurrentPOS(0)	The POS of the second word on the left of the current word and the POS of it
14	CurrentPOS(-1)/CurrentPOS(0)	The POS of the first word on the left of the current word and the POS of it
15	CurrentPOS(-1)/CurrentPOS(1)	The POS of the first word on the left of the current word and the POS of its first word on the right
16	CurrentPOS(0)/CurrentPOS(1)	The POS of the current word and the POS of its first word on the right
17	CurrentPOS(0)/CurrentPOS(2)	The POS of the current word and the POS of its second word on the right
18	CurrentPOS(1)/CurrentPOS(2)	The POS of the first word on the right of the current word and the POS of its second word on the right

Table 3. Experimental results for the model integrated with different constraint

Named entity recognition model	Precision（%）	Recall（%）	F1 value（%）
Baseline experiment	65.31	62.84	64.05
CRF model	67.16	64.85	65.98
CCRF model	71.04	68.64	69.81

Table 4. The recognition results for different types of entity

Entity type	Precision（%）	Recall（%）	F1 value（%）
PER	69.2	66.4	67.8
LOC	75.1	72.5	73.8
ORG	65.1	63.4	64.2

From the above, it can be known that the recognition performance of the LOC has been in the lead in the three kinds of entities by using CCRF model integrated with entity characteristics. The reason is that the LOC's constitute is the most stable of the three kinds of entities' and the prefix indicator of the LOC provides rich semantic information to the LOC entity recognition. But the precision and the recall of it are lower than those of the PER and the LOC when recognizing the ORG in Khmer. And the word of ORG is very extensive and the OOV occupy a considerable proportion, such as most of the enterprise name. In addition, the word-formation of ORGs is more complicated than other named entities, which contains a large number of PER, LOC and other proprietary names. And the recognition performance of PERs, LOCs and other entities also limits the performance of ORG recognition. In this paper,

we reduce appropriately the error recognition caused by the internal entity nesting, so as to improve the recognition performance to a certain extent by using the constraint on Khmer entity characteristics to recognizing the ORG. But the recognition performance of the ORG is worse than that of the PER and the LOC.

4 Conclusion

In this paper, a conditional random field model, integrated with Khmer entity constraints, is proposed for recognizing the named entity. In the CRF model decoding process, we calculate the shortest path labeling sequence by using integer linear programming method in conjunction with the constraint on the logic relation of the Khmer entity word's context, and ultimately realize the recognition of named entity. The experimental results show that the method proposed in this paper can effectively improve the precision and recall of the NER. In the process of recognizing entities, we find that recognition of the ORG is a difficult problem, so in the future we will focus on the recognition of ORG.

References

1. Humphreys K, Gaizauskas R, Azzam S, et al. University of Sheffield: Description of the LaSIE-II system as used for MUC-7[C]//Proceedings of the Seventh Message Understanding Conferences (MUC-7). 1998.
2. Mikheev A, Grover C, Moens M. Description of the LTG system used for MUC-7[C]//Proceedings of 7th Message Understanding Conference (MUC-7). Fairfax, VA, 1998: 1-12.
3. Chen H H, Ding Y W, Tsai S C, et al. Description of the NTU System Used for MET2[C]//Proceedings of 7th Message Understanding Conference. 1998.
4. Fukumoto J, Masui F, Shimohata M, et al. Oki electric industry: Description of the oki system as used for muc-7[C]//Proceedings of the 7th Message Understanding Conference. 1998.
5. Ciravegna F, Gilardoni L, Lavelli A, et al. Flexible text classification for financial applications: the FACILE system[C]//ECAI. 2000: 696-700.
6. Farmakiotou D, Karkaletsis V, Koutsias J, et al. Rule-based NER for Greek financial texts[C]//Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000). 2000: 75-78.
7. Alfred R, Leong L C, On C K, et al. Malay Named Entity Recognition Based on Rule-Based Approach[J]. International Journal of Machine Learning & Computing, 2014, 4(3).
8. Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]//proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 473-480.
9. Isozaki H, Kazawa H. Efficient support vector classifiers for NER[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7.
10. Chieu H L, Ng H T. Named entity recognition: a maximum entropy approach using global information[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7.
11. Zhang Huaping, Liu Qun. Automatic recognition of Chinese personal name based on Role tagging. Chinese Journal of Computers, 2004,27(1):85-91.
12. Qin Wen, Yuan Chunfa. Identification of Chinese unknown word based on decision tree. Journal of Chinese information, 2004, 18(1): 14-19.
13. Goh C L, Asahara M, Matsumoto Y. Chinese unknown word identification using Character-based tagging and chunking. Proceedings of the 4lst Annual Meeting of the Association for Computational Linguistics,SaPPoro,2003:197-200.
14. LU Wei, Ju Yuan, Zhang Xiaojuan, et al. Research on product named entity feature selection and recognition. Document,Information & Knowledge, 2012, 3: 4-34.
15. Qiu Sha, Wang Fuyang, Shen Haoru, et al. Chinese named entity recognition based on part of speech feature with edges . Computer Engineering, 2012, 38(13): 128-130.
16. Yan Yang, Wen Dunwei, Wang Yunji et al. Named entity recognition in Chinese medical records based on cascaded conditional random field. Journal of Jilin University(Engineering and Technology Edition) , 2014, 6: 048.
17. Kazama J, Torisawa K. Exploiting Wikipedia as external knowledge for NER[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007: 698-707.
18. Fu R, Qin B, Liu T. Generating Chinese Named Entity Data from a Parallel Corpus[C]//IJCNLP. 2011: 264-272.
19. Dong Yongquan, Li Qingzhong, Ding Yanhui et al. Constrained conditional random fields for semantic annotation of web data. Journal of Computer Research and Development, 2012, 49(2): 361-371.
20. Pan Huashan, Yan Xin, Yu Zhengtao et al. A Khmer named entity recognition method by fusing language characteristics[C]//Proceedings of the 26th Chinese Control and Decision Conference. 2014.