Laos Organization Name Using Cascaded Model Based on SVM and CRF

Shaopeng Duan[1,2], Lanjiang Zhou[1,2] and Feng Zhou[1,2]

[1]School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2]The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

Abstract. According to the characteristics of Laos organization name, this paper proposes a two layer model based on conditional random field (CRF) and support vector machine (SVM) for Laos organization name recognition. A layer of model uses CRF to recognition simple organization name, and the result is used to support the decision of the second level. Based on the driving method, the second layer uses SVM and CRF to recognition the complicated organization name. Finally, the results of the two levels are combined, And by a subsequent treatment to correct results of low confidence recognition .The results show that this approach based on SVM and CRF is efficient in recognizing organization name through open test for real linguistics, and the recalling rate achieve 80.83％and the precision rate achieves 82.75％.

Keywords. organization name recognition, conditional random fields (CRF),support vector machine (SVM), cascaded model, laos.

1 Introduction

Named entity recognition (NER) is a basic requirement for the processing tasks of many natural languages, the recognition effect of NER has a direct impact on the deep processing of text information, and the recognition of organization name is one of the main tasks of NER. Compared with personal name and place name, Laos organization name has the characteristics of long and unstable, complex with large unlisted words, and nested structures etc. Hence the recognition is more difficult.

Based on existing material, there is few research on Laos entity name recognition, especially on the recognition of Laos organization name, therefore it has great research significance and value.

Rule-based approach were adopted for organization .In this paper, Laos organization names are divided into simple organization name and complicated organization name two categories. Simple organization name only has one word, like court, congress, republic,etc. Complicated9h organization name consists of mufti-letters, The complicated organization name of Laos is different from that of Chinese, which has the characteristic of post-modifier, feature words in the left boundary of the organization name. For example: the Confucius Institute in Laos. Therefore defined as S + P form, S is the feature words of organization name, like company , university etc and P is the rear word of organization name. Namely, Laos complicated organization name is composed of organization name feature word and one or more organization name rear word .me recognition in the early stage. Reference[1-2] builds a model of rules for university names, but the acquisition of those rules are mainly depend on specific area, which has become a bottleneck problem for this approach. Reference [3] adopted decision tree approach for NER, but the recognition accuracy is very low.

Reference [4-5] adopted HMM approach for NER, this model is based on a stringent independence assumption, but in fact most data could not be treated as a series of independent element. Reference [6] adopted SVM for NER. Reference [7-8] adopted CRF for organization name recognition, the result is more desirable, but still have room for improvements. Reference [9] established a role-tagging approach, but the inadequacy is the role set has a great impact on the recognition result, so repeated researches shall be done for the chosen of the right role set. Reference [10] combined machine learning and artificial knowledge for organization name recognition.

In this paper, Laos organization names are divided into simple organization name and complicated organization name two categories. Simple organization name only has one word, like court, congress, republic,etc. Complicate organization name consists of mufti-letters, The complicated organization name of Laos is different from that of Chinese, which has the characteristic of post-modifier, feature words in the left boundary of the organization name. For example: the Confucius Institute in Laos. Therefore defined as S + P form, S is the feature words of organization name, like company, university ,etc and P is the rear word of organization name. Namely, Laos complicated organization name is composed of organization name feature word and one or more organization name rear word.

2 Laos organization name using cascaded model based on svm and crf

2.1 The resources needed to recognition the laos organization name

Automatic extraction of each list organization name recognition required from the training corpus. The details are as follows:

(1)Feature word table D_f

Feature words refers to the organization name is characterized by significant words, such as "factory , University, company".

The organization name recognition of Laos is the first of left boundary, so the establishment of the list can be used as the trigger condition of organization name recognition.

(2)Rear word table D_b

Rear word refers to the words that in addition to the feature words of organization name, location names and common nouns larger proportion, but overall, the word is more complex, there is a strong randomness.

(3)The left and right boundary word table

The left boundary word is the first word of organization name, such as "representative ""admitted ";The right boundary word is the last word of organization name, such as "director" "host ".Different boundary word indicate different directions on the boundary of organization name. Therefore, .When statistical boundary word table, it is necessary to statistics the number of times as boundary word, and according to the number of times will be divided into different levels.

(4)Simple organization name table

Mainly used for simple organization name recognition, the words that in the vocabulary are considered to be the candidate words of simple organization name, for example: post office .

2.2 Laos organization name using cascaded model based on CRF and SVM

Laos organization name recognition model divided into two layers, the first layer use CRF to recognition simple organization name, and recognition results transmitted to the second layer; The second layer is based on the driven tagging method, which combines SVM and CRF to identify the complex organization name, that is to use the SVM to identify the name of the left boundary of the organization, the words which to be recognized as the left boundary word

backwards using CRF to rear marking. Then the recognition results of the two layer are combined. Figure 1 is an example of the organization name recognition is converted to a sequence annotation, figure 2 is the model structure.

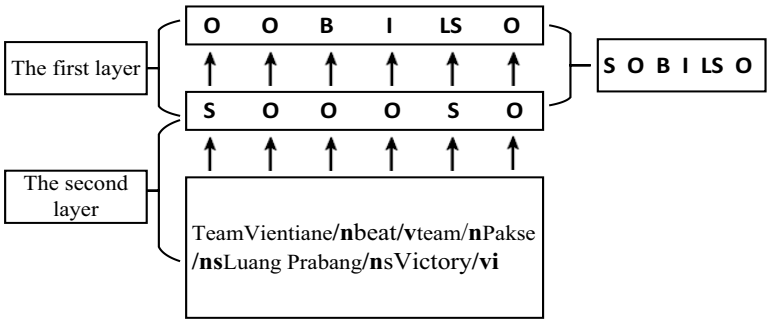

Fig.1 Recognition process of cascaded model

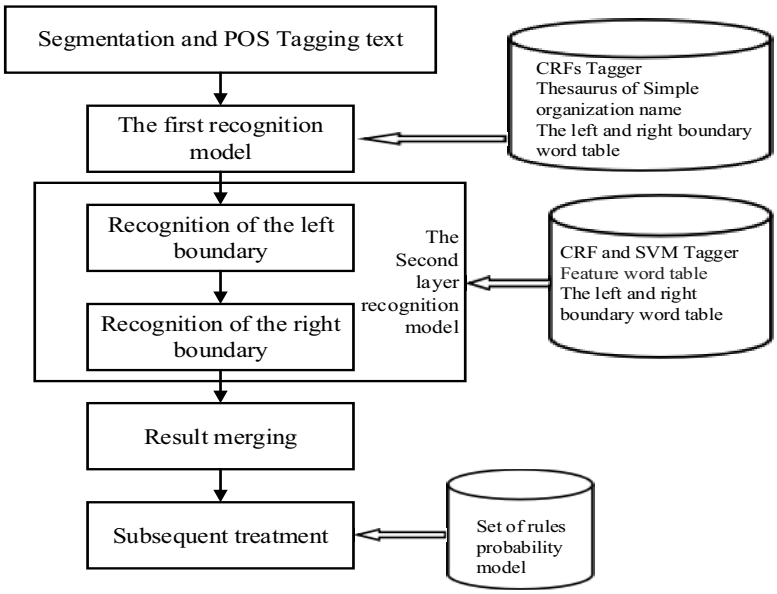

Fig.2 Architecture of Hybrid cascaded model

3 Svm and crf in combination with the identification of complex laos organization name

(1) SVM to determine the left boundary of the organization name

Left boundary is determined to be a two value classification problem, and SVM is an excellent two value classifier, so the SVM is used to determine the left boundary. To appear in the feature list of words were as the left border of the lookup, SVM was used to conduct screening determined whether indeed left word boundary of institutions . SVM also need to choose the appropriate features for a specific task, taking into account the two aspects of efficiency and recognition, select the part of speech and word form these two features. The definition of 11 dimensional vector format are as follows: $(S\ W(-2)\ \ P(-2)\ \ W(-1)\ \ P(-1)\ \ W(0)\ \ P(0)\ \ W(1)\ \ P(1)\ \ W(2)\ \ P(2))$,Where a is a class, in the left boundary of the task, $S=-1$ represents the word is not the left edge of the organization name, $S=+1$ represents the word is the organization name left boundary.

(2）CRF conduct rear label

After the left boundary is determined, use CRF to carry on the back label. Considering a smaller proportion of the organization name, use full labeled policy will cause a lot of waste of resources, decided to adopt the drive type annotation, namely the left boundary driven, only for the look-up annotation. Candidate words determine the rules as follows: assuming the longest organization name length is n, each determined a left boundary, the word directly labeled as "L", the words followed by N - 1 words become the organization name look-up, unless encounter punctuation, another left boundary or f the first of a line. Then tagging the word that Identified as candidate,in other words which are directly labeled as non organization name components .The use of this strategy, to a certain extent, reduces the training and tagging time, improve the recognition efficiency, and because of the reduction of redundant information, the recognition accuracy is also improved. The atomic features used here are also to be as follows in addition to the a used in the first layer.

This method is suitable for the complete identification of organization names, according to different materials need to make some adjustments in the way. If the text does not complete the organization name occupies a certain proportion, the use of two kinds of methods to identify, first in this paper, the second directly with CRF identification, then compare two recognition results, identification of different the confidence degree is higher as the final results.

3.1 Subsequent processing

Subsequent processing includes two parts, the first part for the construction of a probabilistic model, the recognition result confidence below the strings of a certain threshold calculating the confidence, a proper threshold is selected through experiments and reliability are higher than the threshold determination for organization name, or identified as non organization name. The Credibility T (org) of organization names including organization names feature words credibility T (S) and rear organization name word credibility T (P), is calculated as follows:

$$T(S)=\frac{\log(N_s+2)}{\sum_{y\in D_f}(N_y+2)}\ T(P)=\frac{\log(N_P+2)}{\sum_{y\in D_b}(N_y+2)}\ T(org)=\frac{C_n\times\sqrt[3]{n}\times(\sum_{i=1}^{n}T(F_i)+T(S))}{n+1}$$

Where N_s is the number of established institutions characteristics when S the feature word thesaurus; N_P is the number of the establishment of the former P of the rear table ; C_n is the adjustment factor, n is the number of rear word of the organization name. The second part for the construction of the rule model, mainly for identification of Incomplete organization name and concurrent mechanism, and the correction of recognition errors. Rules for example are as follows:

parallel relationship words (such as: and , versus , etc.) before and after the labeling should be consistent, inconsistent will mark the confidence degree is higher.

From the training corpus extraction framework of organization names, such as: (admitted to ,candidates to + Organization Name + (school, reading, work).And according to the number of times to streamline, Confidence is lower than a threshold of recognition and matching the results to determine the matching for organization name, or identified as non organization name.

4 Experiment analysis

In this paper, we use the data collected from the Lao language news website, We use word segmentation program to deal with the data,and in part by the Lao language experts and the Lao students to carry out a manual tagging, used as an experimental corpus .The remaining part is used as a test corpus. Test results taken three common evaluation

indicators,that is the correct rate (P), recall (R) and a comprehensive index F value (F) to the evaluation result of organization name recognition.

Testing method is as follows:

Testing results were evaluated by correct rate, recall rate, and F-measure .

(1) correct rate(P): The P measures the numbers of correct named entities in the answer file over the total number of named entities in the answer file.

$$\mathrm{P} = \frac{\text{The number of correct labeling of entities}}{\text{The total number of marked entities}} \times 100\%$$

(2) recall rate(R): The R measures the numbers of correct named entities in the answer file over the total number of named entities in the answer file.

$$\mathrm{R} = \frac{\text{The number of correct labeling of entities}}{\text{The correct total number of entities}} \times 100\%$$

(3) F-measure(F): The F is a weighted combination of precision and recall.

$$F = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \times 100\%$$

The experimental results of this method are shown in table 1.

Tab.1 Recognition result

Experiment	Correct rate(P)(%)	Recall rate (R)(%)	F-measure(F)(%)
Simple organization name recognition	83.62	85.19	84.40
Left boundary recognition	79.15	81.24	80.18
Complex organization name recognition	80.36	80.56	80.44
Global recognition	80.83	82.75	81.75

Experiments on complex organization names using different methods, the experiment is carried out with different methods, and the experimental results are compared as shown in Table 2.

Tab.2 Results of complicated organization name recognition

Recognition methods	Correct rate(P)(%)	Recall rate (R)(%)	F-measure(F)(%)
CRF	77.72	79.13	78.42
Driven tagging of SVM＋CRF	80.83	82.75	81.75

As can be seen from the experimental results, the recognition effect of driven tagging of SVM＋CRF is best, the training time is reduced due to the reduction of redundant information. But because the recognition of this paper is based on the correct word segmentation and part of speech tagging, the error of the word segmentation will decrease the recognition accuracy.

5 Conclusion

In this paper, a two layer model based on CRF and SVM is established to recognition the organization names,According to the different characteristics of the simple organization names and the complex organization names, at different levels, different methods are used to recognition.

Acknowledgements

Foundation item: China National Natural Science Foundation (61562049), (61662040).

References

1. Godeny, B., Rule Based Product Name Recognition and Disambiguation., Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, Brussels, 2012, pp: 858 - 860.
2. YanDanhui, BiYude., Identification of a rule-based Vietnamese named entity, Journal of Chinese Information Processing, 2014.9.
3. Sekine, S., Grishman, R., and Shinnou, H., A decision tree method for finding and classifying names in Japanese texts, in .6th Workshop on Very Large Corpora, Montreal, Canada, 1998.
4. Zheng Jiaheng, Zhang Hui, Automatic identification of Chinese organization names based on, Journal of Computer Applications, 2002 (11) .
5. Liu Qun, ZhangHuaping, YuHongkui, ChengXueqi., Based on cascaded hidden Markov model for Chinese lexical analysis, Journal of computer research and development., 2004 (08).
6. Xu-Dong Lin,Hong Peng,Bo Liu, Chinese Named Entity Recognition using Support Vector Machines., Machine Learning and Cybernetics, 2006 International Conference on, 2006, pp:4216- 4220.
7. Sisouvanh Vanthanavong, Choochart Haruechaiyasak., Lao Word Segmentation Based on Conditional Random Fields. Conference on Human Language Technology for Development, Alexandria, Egypt, 2011, pp: 2-5.
8. Su-Xiang Zhang, Guo-Yang Gao, Yin-Cheng Qi, and Wilks,Y, personal name and location name recognition based on conditional random fields(CRFs), Proceedings of the Eighth International Confreence on Machine Learning and Cybernetics, Baoding, 2009, pp:12-15.
9. Zhang Huaping, Liu Qun, Automatic Identification of Chinese Names Based Role Labeling, 2004, 27(1), pp:85-91.
10. Chanlekha, H., and Kawtrakul, A, Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information, The Specialty Research Unit of Natural Language Processing and Intelligent Information System Technology, 2011, pp:720-836.