Chinese Named Entity Recognition and Word Segmentation Based on Character

He Jingzhou, Wang Houfeng
Institution of Computational Linguistics
School of Electronics Engineering and Computer Science,
Peking University, China, 100871
{hejingzhou, wanghf}@pku.edu.cn

Abstract

Chinese word segmentation and named entity recognition (NER) are both important tasks in Chinese information processing. This paper presents a character-based Conditional Random Fields (CRFs) model for such two tasks. In The SIGHAN Bakeoff 2007, this model participated in all closed tracks for both Chinese NER and word segmentation tasks, and turns out to perform well. Our system ranks 2nd in the closed track on NER of MSRA, and 4th in the closed track on word segmentation of SXU.

1 Introduction

Chinese word segmentation and NER are two of the most fundamental problems in Chinese information processing and have attracted more and more attentions. Many methods have been presented, of which, machine learning methods have obviously competitive advantage in such problems. Maximum Entropy (Ng and Low, 2005) and CRFs (Hai Zhao et al. 2006, Zhou Junsheng et al. 2006) come to good performance in the former SIGHAN Bakeoff.

We consider both tasks as sequence labeling problem, and a character-based Conditional Random Fields (CRFs) model is applied in this Bakeoff. Our system used CRF++ package Version 0.49 implemented by Taku Kudo from sourceforge[1].

2 System Description

The system is mainly based on CRFs, while different strategies are introduced in word segmentation task and NER task.

2.1 CRFs

CRFs are undirected graphical models which are particularly well suited to sequence labeling tasks, such as NER & word segmentation. In these cases, CRFs are often referred to as linear chain CRFs.

CRFs are criminative models, which allow a richer feature representation and provide more natural modeling.

[1] http://www.sourceforge.net/

CRFs define the conditional probability of a state sequence given an input sequence as

$$p(\mathbf{s}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})}\exp\left(\sum_{k}\lambda_k F_k(\mathbf{s},\mathbf{o})\right)$$

Where F is a feature function set over its arguments, λk is a learned weight for each feature function, and Z is the partition function, which ensures that p is appropriately normalized.

2.2 Word Segmentation Task

Similar to (Ng and Low, 2005), a Chinese character comes into four different tags, as in Table 1.

Tag	Meaning
S	Character that occurs as a single-character word
B	Character that begins a multi-character word
I	Character that continues a multi-character word
E	Character that ends a multi-character word

Table 1. Word segmentation tag set

(Ng and Low, 2005) presented feature templates as:

(a) Cn(n = −2,−1, 0, 1, 2)
(b) CnCn+1(n = −2,−1, 0, 1)
(c) C−1C1
(d) Pu(C0)
(e) T(C−2)T(C−1)T(C0)T(C1)T(C2)

In order to find the effect of different features on the result, some experiments are conducted on these templates, with 90% of the training data provided by The SIGHAN Bakeoff 2007 for training, leaving 10% for testing. Based on our results, some templates are adjusted as follows.

First, the character-window is reduced to (-1, 0, 1) in (a).

Second, feature template (d) is not used. Instead, original sentences are split into clauses ended with punctuations “，”, “。”, “：”, “？”, “；” and “！”. There are two advantages of this processing: (1) the template is simplified but little performance is lost; (2) shorter sentences make CRFs training quicker.

Third, template (e) is separate it to three items: T(C-1), T(C0) and T(C1), in which five types of the characters are considered: N stands for numbers, D for dates, E for English letters, S for punctuations and C for other characters. Besides, this feature template does not always contribute to the segmentation result in our experiments, so it will be a tradeoff whether to use it or not according to experiments.

Finally, we use the following feature templates:

(a) Cn(n = −1, 0, 1)
(b) CnCn+1(n = −1, 0)
(c) C−1C1
(d) T(Cn) (n = −1, 0, 1)

We only took part in word segmentation closed track, so no additional corpora, dictionary or linguistic resources are introduced.

1.1 NER Task

Many Chinese NER researches are based on word segmentation and even Part-Of-Speech (POS) tagging. In fact these steps are not necessary. The relationship of them is described in Figure 1.

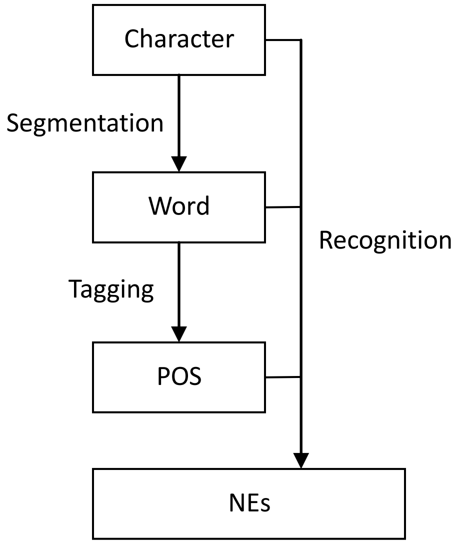

Figure 1. NER model achitecture

In closed track of both MSRA and CITYU, a character-based CRFs model is used in our system. There two reasons as follows:

First, no word-level information is provided in training data of NER tasks in closed track, so it's hard to perform word segmentation with good accuracy.

Second, we had done some experiments on Chinese NER, and found that character-based method outperformed word segmentation and word segmentation + POS, if only character sequence is given. Table 2 shows the comparison results.

Feature Level	Integrated F-measure
Character	0.8760
Word	0.8538
POS	0.8635

Table 2. Comparison result among different NER models[3]

In our NER system, a Chinese character can be labeled as one of four different tags, as in Table 3.

Tag	Meaning
B	First character of a NE
I	Character in a NE but neither the first nor the last one
E	Last character of a NE except a single-character one
O	Character not in a NE

Table 3. NER tag set

It's similar to the standard of The SIGHAN Bakeoff 2007 NER track except for an additional tag "E". Unlike the tag set used in word segmentation task, there is no "S" tag for single-character NEs. This kind of entities is usually surname of a Chinese person. In this case, the tag "B" will handle it as well.

There are actually 3 types of NEs in MSRA and CITYU corpora: PER, LOC and ORG, so the tag set is further divided into 10 sub tags: B-PER, I-PER, E-PER, B-LOC, I-LOC, E-LOC, B-ORG, I-ORG, E-ORG and O.

The feature template is similar to the one used in word segmentation task except that here a character-window of (-2,-1, 0, 1,2) is applied:

(a) Cn(n = −2,−1, 0, 1, 2)
(b) CnCn+1(n = −2,−1, 0, 1)
(c) C−1C1
(d) T(Cn) (n = −1, 0, 1)

For CRFs, the precision is usually high while recall is low. To solve this problem, a set of feature templates (only differ in window size, or punctuations) are used to train several different models, and finally achieve a group of results. Merge them as in Table 4 (for the same Chinese character string in result A and B).

[3] Train with Annotation Corpora of People's Daily 199801 and test with 199806

A	B	Result
Is a NE	Isn't a NE	Refer to A
Isn't a NE	Is a NE	Refer to B
Isn't a NE	Isn't a NE	Refer to A or B
Is a NE	Is the same NE type as A	Refer to A or B
Is a NE	Is a NE but not the same type as A	Choose A or B according to predefined rules

Table 4. Merge strategy of results

With a slight loss of precision, an improvement is achieved on recall rate.

In open track of MSRA, an additional segmentation system is used on the corpora and some NEs are retrieved based on several predefined rules. It was merged with closed track result to form open track result.

3 Evaluation Results

Our word segmentation system is evaluated in closed track on all 5 corpora of CITYU, CKIP, CTB, NCC and SXU. Table 5 shows our results on the best RunID. Columns R, P, and F show the recall, precision, and F measure, respectively. Column BEST shows best F-measure of all participants in the track.

Our NER system is evaluated in closed track on both MSRA and CITYU corpora, and open track on MSRA corpora only. Table 6 shows our official results on best RunID. Columns R, P, and F show the recall, precision, and F measure, respectively. Column BEST shows best F-measure of all participants in the track.

Track (all closed)	R	P	F	BEST
CITYU	0.9421	0.9339	0.938	0.951
CKIP	0.9369	0.927	0.9319	0.947
CTB	0.9487	0.9514	0.95	0.9589
NCC	0.9278	0.925	0.9264	0.9405
SXU	0.9543	0.9568	0.9556	0.9623

Table 5. Evaluation results on word segmentation

Track	R	P	F	BEST
CITYU closed	0.7608	0.8751	0.814	0.8499
MSRA closed	0.8862	0.9304	0.9078	0.9281
MSRA open	0.9135	0.9321	0.9227	0.9988

Table 6. Evaluation results on NER

4 Conclusion

In this paper, a character-based CRFs model is introduced on both word segmentation and NER. Experiments are done to form our feature templates, and approaches are used to further improve its performance on NER. The evaluation results show its competitive performance in The SIGHAN Bakeoff 2007. We'll launch more research and experiments on feature picking-up methods and combination between character-based model and other models in the future.

References

Hai Zhao, Chang-Ning Huang and Mu Li. An Improved Chinese Word Segmentation System with Conditional Random Field. 2006. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.*

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. A maximum Entropy Approach to Chinese Word Segmentation. 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.*

Wang Xinhao, Lin Xiaojun, Yu Dianhai, Tian Hao,Wu Xihong. Chinese Word Segmentation with Maximum Entropy and N-gram Language Model. 2006. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.*

Zhou Junsheng, Dai Xinyu, He Liang, Chen Jiajun. Chinese Named Entity Recognition with a Multi-Phase Model. 2006. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.*

Acknowledgement

This paper is supported by National Natural Science Foundation of China (No. 60675035), National Social Science Foundation of China (No. 05BYY043) and Beijing Natural Science Foundation (No. 4072012).