Two Step Chinese Named Entity Recognition Based on Conditional Random Fields Models

Yuanyong Feng* **Ruihong Huang*** **Le Sun†**

*Institute of Software, Graduate University
Chinese Academy of Sciences
Beijing, China, 100080
{comerfeng,ruihong2}@iscas.cn

†Institute of Software
Chinese Academy of Sciences
Beijing, China, 100080
sunle@iscas.cn

Abstract

This paper mainly describes a Chinese named entity recognition (NER) system NER@ISCAS, which integrates text, part-of-speech and a small-vocabulary-character-lists feature and heristic post-process rules for MSRA NER open track under the framework of Conditional Random Fields (CRFs) model.

1 Introduction

The system NER@ISCAS is designed under the Conditional Random Fields (CRFs. Lafferty et al., 2001) framework. It integrates multiple features based on single Chinese character or space separated ASCII words. The early designed system (Feng et al., 2006) is used for the MSRA NER open track this year. The output of an external part-of-speech tagging tool and some carefully collected small-scale-character-lists are used as open knowledge. Some post process steps are also applied to complement the local limitation in model's feature engineering.

The remaining of this paper is organized as follows. Section 2 introduces Conditional Random Fields model. Section 3 presents the details of our system on Chinese NER integrating multiple features. Section 4 describes the post-processings based on some heuristic rules. Section 5 gives the evaluation results. We end our paper with some conclusions and future works.

2 Conditional Random Fields Model

Conditional random fields are undirected graphical models for calculating the conditional probability for output vertices based on input ones. While sharing the same exponential form with maximum entropy models, they have more efficient procedures for complete, non-greedy finite-state inference and training.

Given an observation sequence $o=<o_1, o_2, ..., o_T>$, linear-chain CRFs model based on the assumption of first order Markov chains defines the corresponding state sequence s' probability as follows (Lafferty et al., 2001):

$$p_\Lambda(\mathbf{s}\mid\mathbf{o}) = \frac{1}{Z_\mathbf{o}}\exp(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)) \qquad (1)$$

Where Λ is the model parameter set, Z_o is the normalization factor over all state sequences, f_k is an arbitrary feature function, and λ_k is the learned feature weight. A feature function defines its value to be 0 in most cases, and to be 1 in some designated cases. For example, the value of a feature named "MAYBE-SURNAME" is 1 if and only if s_{t-1} is OTHER, s_t is PER, and the t-th character in $\boldsymbol{o}$ is a common-surname.

The inference and training procedures of CRFs can be derived directly from those equivalences in HMM. For instance, the forward variable $\alpha_t(s_i)$ defines the probability that state at time t being s_i at time t given the observation sequence $\boldsymbol{o}$. Assumed that we know the probabilities of each possible value s_i for the beginning state $\alpha_0(s_i)$, then we have

$$\alpha_{t+1}(s_i) = \sum_{s'}\alpha_t(s')\exp(\sum_{k}\lambda_k f_k(s', s_i, \mathbf{o}, t) \qquad (2)$$

In similar ways, we can obtain the backward variables and Baum-Welch algorithm.

3 Chinese NER Using CRFs Model Integrating Multiple Features

Besides the text feature(TXT), simplified part-of-speech (POS) feature, and small-vocabulary-character-lists (SVCL) feature, which use in the early system (Feng et al., 2006), some new features such as word boundary, adjoining state bigram – observation and early NE output are also combined under the unified CRFs framework.

The text feature includes single Chinese character, some continuous digits or letters.

POS feature is an important feature which carries some syntactic information. Unlike those in the earyly system, the POS tag set are merged into 9 categories from the criterion of modern Chinese corpora construction (Yu, 1999), which contains 39 tags.

The third type of features are derived from the small-vocabulary-character lists which are essentially same as the ones used in last year except with some additional items. Some examples of this list are given in Table 1.

Value	Description	Examples
digit	Arabic digit(s)	1,2,3
letter	Letter(s)	A,B,C,...,a, b, c
Continuous digits and/or letters (The sequence is regarded as a single token)		
chseq	Chinese order 1	㈠, ⑴, ①, Ⅰ
chdigit	Chinese digit	1 , 壹, 一
tianseq	Chinese order 2	甲, 乙, 丙, 丁
chsurn	Surname	李, 吴, 郑, 王
notname	Not name	将, 对, 那, 的, 是, 说
loctch	LOC tail character	区, 国, 岛, 海, 台, 庄, 冲
orgtch	ORG tail character	府, 团, 校, 协, 局,
	ter	办, 军
other	Other case	情, 规, 息, ！, 。

Table 1. Some Examples of SVCL Feature

The fourth type of feature is word boundary. We use the B, I, E, U, and O to indicate Begining, Inner, Ending, and Uniq part of, or outside of a word given a word segmentation. The O case occurs when a token, for example the charactor "&", is ignored by the segmentator. We do not combine the boundary information with other features because we argue it is very limited and may cause errors.

The last type of features is bigram state combined with observations. We argue that observatoin (mainly is of named entity derived by early system or character text itself) and state transition are not conditionally independent and entails dedicate considerings.

Each token is presented by its feature vector, which is combined by these features we just discussed. Once all token feature (Maybe including context features) values are determined, an observation sequence is feed into the model.

Each token state is a combination of the type of the named entity it belongs to and the boundary type it locates within. The entity types are person name (PER), location name (LOC), organization name (ORG), date expression (DAT), time expression (TIM), numeric expression (NUM), and not named entity (OTH). The boundary types are simply Beginning, Inside, and Outside (BIO).

All above types of features are extracted from a varying length window. The main criteria is that wider window with smaller feature space and narrow window when the observation features are in a large range.

The main feature set is shown the following.

Character Texts(TXT): TXT_{-2}, TXT_{-1}, TXT_0, TXT_1, TXT_2, $TXT_{-1}TXT_0$, TXT_1TXT_0, TXT_1TXT_2
simplified part-of-speech (POS): unigram: $POS_{-4} \sim POS_4$

small-vocabulary-character-lists (SVCL): unigram: $SVCL_{-2}$ ~ $SVCL_7$ bigram:$SVCL_0SVCL_1$, $SVCL_1SVCL_2$
Word Boundary (WB): WB_{-1},WB_0,WB_1
Named Entity (NE): unigram: NE_{-4} ~ NE_4 bigram:$NE_{-2}NE_{-1}$, $NE_{-1}NE_0$, NE_0NE_1, NE_1NE_2
State Bigram (B) – Observation: B, B-TXT_0, B-NE_{-1}, B-NE_0, B-NE_1

Table 2. The Main Feature Set

4 Post Processing on Heuristic Rules

Observing from the evaluation, our model has worse performance on ORG and PER than LOC. Furthermore, the analysis of the errors tells us that they are hard to be tackled with the improvement of the model itself. Therefore, we decided to do some post-process to correct certain types of tagging errors of the unified model mainly concerning the two kinds of entities, ORG and PER.

At the training phrase, we compare the tagging output of the model with the correct tags and collect the falsely tagged instances. To identify the rules used in the post-process, we categorize the errors into several types, discriminate the types and encode them into the rules according to two principles:

1) the rules are applied on the tagged sequences output by the unified model.
2) The rules applied shouldn't introduce more other errors.

As a result, we have extracted eight rules, seven for ORG, one for PER. Generally, the rules work only on the local context of the examined tags, they correct some type of error by changing some tags when seeing certain pattern of context before or after the current tags in a limited distance. We want to give one rule as one example to explain the way they function.

Example: {<LOC>}+<ORG> ==> <ORG>,

After this rule is applied, one or more locations followed by a organization name will be tagged ORG. This is the case where there are a location name in a organization name. Besides, we can see since the location and latter part of the organization name are tagged separately in the unified model, we may only resort to the post-process to get the right government boundary.

5 Evaluation

5.1 Results

The evaluations in training phrase tell us the post-process can improve the performance by one percent. We are satisfied since we just applied eight rules.

The formal eveluation results of our system are shown in Table 3.

	R	P	F
Overall	86.74	90.03	88.36
PER	90.83	92.16	91.49
LOC	89.89	91.66	90.77
ORG	77.99	85.16	81.41

Table 3. Formal Results on MSRA NER Open

5.2 Errors from NER Track

The NER errors in our system are mainly of as follows:

- Abbreviations

Abbreviations are very common among the errors. Among them, a significant part of abbreviations are mentioned before their corresponding full names. Some common abbreviations has no corresponding full names appeared in document. Here are some examples:

R[1]: 总后[嫩江基地 LOC]的先进事迹

K: [总后嫩江基地 LOC]的先进事迹

R: [中 丹 LOC]兩國

K: [中 LOC][丹 LOC]兩國

In current system, the recognition is fully depended on the linear-chain CRFs model, which is heavily based on local window observation features; no abbreviation list or special abbreviation

[1] R stands for system response, K for key.

recognition involved. Because lack of constraint checking on distant entity mentions, the system fails to catch the interaction among similar text fragments cross sentences.

- Concatenated Names

For many reasons, Chinese names in titles and some sentences, especially in news, are not separated. The system often fails to judge the right boundaries and the reasonable type classification. For example:

R:将[瓦西里斯 LOC]与[奥纳西斯 PER]比较

K:将[瓦西里斯 PER]与[奥纳西斯 PER]比较

- Hints

Though it helps to recognize an entity at most cases, the small-vocabulary-list hint feature may recommend a wrong decision sometimes. For instance, common surname character "王" in the following sentence is wrongly labeled when no word segmentation information given:

R:[希腊 LOC]船[王 康斯坦塔科普洛斯 PER]

K:[希腊 LOC]船 王[康斯坦塔科普洛斯 PER]

Other errors of this type may result from failing to identify verbs and prepositions, such as:

R:全国保护明天行动组委会 举行表彰会

K:[全国保护明天行动组委会 ORG]举行表彰会

6 Conclusions and Future Work

We mainly described a Chinese named entity recognition system NER@ISCAS, which integrates text, part-of-speech and a small-vocabulary-character-lists feature for MSRA NER open track under the framework of Conditional Random Fields (CRFs) model. Although it provides a unified framework to integrate multiple flexible features, and to achieve global optimization on input text sequence, the popular linear chained Conditional Random Fields model often fails to catch semantic relations among reoccurred mentions and adjoining entities in a catenation structure.

The situations containing exact reoccurrence and shortened occurrence enlighten us to take more effort on feature engineering or post processing on abbreviations / recurrence recognition.

Another effort may be poured on the common patterns, such as paraphrase, counting, and constraints on Chinese person name lengths.

From current point of view, enriching the hint lists is also desirable.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China under grant #60773027, #60736044 and by "863" Key Projects #2006AA010108.

References

Chinese 863 program. 2005. Results on Named Entity Recognition. *The 2004HTRDP Chinese Information Processing and Intelligent Human-Machine Interface Technology Evaluation*.

Yuanyong Feng, Le Sun, Yuanhua Lv. 2006. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields Models. *Proceedings of SIGHAN-2006, Fifth SIGHAN Workshop on Chinese Language Processing,* Sydney, Australia,.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML*.

Shiwen Yu. 1999. Manual on Modern Chinese Corpora Construction. Institute of Computational Language, Peking Unversity. Beijing.