Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields Models

Yuanyong Feng **Le Sun** **Yuanhua Lv**

Institute of Software, Chinese Academy of Sciences, Beijing, 100080, China

{yuanyong02, sunle, yuanhua04}@ios.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 for MSRA NER open track under the framework of Conditional Random Fields (CRFs) model. The techniques used for the close NER and word segmentation tracks are also presented.

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., 2005) 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 outer knowledge.

The close word segmentation and named entity recognition tracks are also based on this system by some adjustments.

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 features extraction for close track. 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):

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

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 ***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 ***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 for Open Track

In our system the text feature, part-of-speech (POS) feature, and small-vocabulary-character-lists (SVCL) feature are combined under a unified CRFs framework.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 181–184,
Sydney, July 2006. ©2006 Association for Computational Linguistics

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

POS feature is an important feature which carries some syntactic information. Our POS tag set follows the criterion of modern Chinese corpora construction (Yu, 1999), which contains 39 tags.

The last feature is based on lists. We first list all digits and English letters in Chinese. Then most frequently used character feature in Chinese NER are collected, including 100 single character surnames, 100 location tail characters, and 40 organization tail characters. The total number of these items in our lists is less than 600. The lists altogether make up a list feature (SVCL). 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, 办, 军
other	Other case	情, 规, 息, ！，。

Table 1. Some Examples of SVCL Feature

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).

4 Feature Extraction for Close Tracks

In close tracks, only character and word list features which are extracted from training data are applied for word segmentation. In NER track we also include a named entity list extracted from the training data.

To extract the list feature, we simply search each text string among the list items in maximum length forward way.

Taking the word segmentation task for instance, when a text string $c_1c_2\ldots c_n$ is given, we tag each character into a BIO-WL style. If $c_ic_{i+1}\ldots c_j$ matches an item I of length j-i+1 and no other item I' of length k (k>j-i+1) in the list matches $c_ic_{i+1}\ldots c_j\ldots c_{k+i-1}$, then the characters are tagged as follows:

c_i	c_{i+1}	…	c_j
B-WL	I-WL	…	I-WL

If no item in the list matches head subpart of the string, then c_i is tagged as 0.

The tagging operation iterates on the remaining part until all characters are tagged.

5 Evaluation

5.1 Results

The system for our MSRA NER open track submission has some bugs and was trained on a much smaller training data set than the full set the organizer provided. The results are very low, see Table 2:

Accuracy	96.28%
Precision	83.20%
Recall	67.03%
FB1	74.24%

Table 2. MSRA NER Open

When we fixed the bug and retrained on the full training corpus, the result comes out to be as follows:

Accuracy	98.24%
Precision	89.38%
Recall	83.07%
FB1	86.11%

Table 3. MSRA NER Open (retrained)

All the submissions on close tracks are trained on 80% of the training corpora, the remaining 20% parts are used for development. The results are shown in Table 4 and Table 5:

Measure	Corpus			
	UPUC	CityU	CKIP	MSRA
Recall	0.922	0.952	0.939	0.933
Precision	0.912	0.954	0.929	0.942
FB1	0.917	0.953	0.934	0.937
OOV Recall	0.680	0.747	0.606	0.640
IV Recall	0.945	0.960	0.954	0.943

Table 4. WS Close

Measure	MSRA	CityU	LDC
Accuracy	92.44	97.80	93.82
Precision	81.64	92.76	81.43
Recall	31.24	81.81	59.53
FB1	45.19	86.94	68.78

Table 5. NER Close

The reason for low measure on MSRA NER track exists in that we chose a much smaller training data file encoded in CP936 (about 7% of the full data set). This file may be an incomplete output when the organizer transfers from another encoding scheme.

5.2 Errors from NER Track

The NER errors in our system are mainly 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]:针对大陆人民申请进入 金 妈 地区，[内政部警政署入出境管理局 ORG][金门 GPE]、[妈祖 GPE]服务站定于明天……

K:针对大陆人民申请进入 [金 GPE][妈 GPE]地区，[内政部警政署入出境管理局 ORG][金门 GPE]、[妈祖 GPE]服务站定于明天……

R: 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 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:身边还有[张龙 赵虎 PER]王朝[马汉 PER] 四个卫士

K:身边还有[张龙 PER][赵虎 PER][王朝 PER][马汉 PER] 四个卫士

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:[中共中央 致 中国致公党十一大 ORG]的贺词……向[致公党 ORG]的同志们……

K:[中共中央 ORG]致[中国致公党十一大 ORG]的贺词……向[致公党 ORG]的同志们……

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

R:包公 赶驴

K:[包公 PER] 赶驴

- Other Types:

R:特别助理 由喜贵 等也同机抵达。

K:特别助理[由喜贵 PER]等也同机抵达。

R:脸谱上还有 日 月 的图案

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

K:脸谱上还有[日 LOC][月 LOC]的图案

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.

Acknowledgment

This work is supported by the National Science Fund of China under contract 60203007.

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 and Junlin Zhang. 2005. Early Results for Chinese Named Entity Recognition Using Conditional Random Fields Model, HMM and Maximum Entropy. *IEEE Natural Language Processing & Knowledge Engineering*. Beijing: Publishing House, BUPT. pp. 549~552.

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.