CRFs-Based Named Entity Recognition Incorporated with Heuristic Entity List Searching

Fan Yang
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
fyang@nlpr.ia.ac.cn

Jun Zhao
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
jzhao@nlpr.ia.ac.cn

Bo Zou
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
bzou@nlpr.ia.ac.cn

Abstract

Chinese Named entity recognition is one of the most important tasks in NLP. Two kinds of Challenges we confront are how to improve the performance in one corpus and keep its performance in another different corpus. We use a combination of statistical models, i.e. a language model to recognize person names and two CRFs models to recognize Location names and Organization names respectively. We also incorporate an efficient heuristic named entity list searching process into the framework of statistical model in order to improve both the performance and the adaptability of the statistical NER system. We participate in the NER tests on open tracks of MSRA. The testing results show that our system can performs well.

1 Introduction

Named Entity Recognition (NER) is one of the most important tasks in NLP, and acts as a critical role in some language processing applications, such as Information Extraction and Integration, Text Classification etc. Many efforts have been paid to improve the performance of NER.

NER task in Chinese has some differences from in English as follows. 1) There is no space between Chinese characters, which make boundary recognition more difficult. 2) In English, a capitalized letter at the beginning position of a word implies that the word is a part of a named entity. However, this kind of characteristic does not exist in Chinese.

In the paper, we will focus on two kinds of problems. 1) How to improve the performance of Chinese NER in one corpus, which contains boosting precision rate, recall rate and F-measure rate. 2) How to enhance the adaptability of a Chinese NER system, which means that a system can get a good performance on a testing set which has many differences from the training set. To solve the first problem, we should select a good model and adjust parameters carefully. But there is no framework that can solve the second problem completely.

Our goal is to find a way to solve these two problems. We select a language model to recognize Person names, and two CRFs models are used to recognize Location and Organization separately. We also try to incorporate a large-scale named entity list into the statistical model, where a heuristic searching method is developed to match the entities in the list quickly and efficiently.

2 Framework of NER System

The Input of the system is a raw text. We will apply some pre-processing such as code transformation. Then the heuristic searching will be executed to find the appearance of the entities in the named entity list. After that, two CRFs that have been trained before will be used to recognize Location and Organization based on the result of word segmentation, and a language model will be used to find Person names. All the results will be integrated at last.

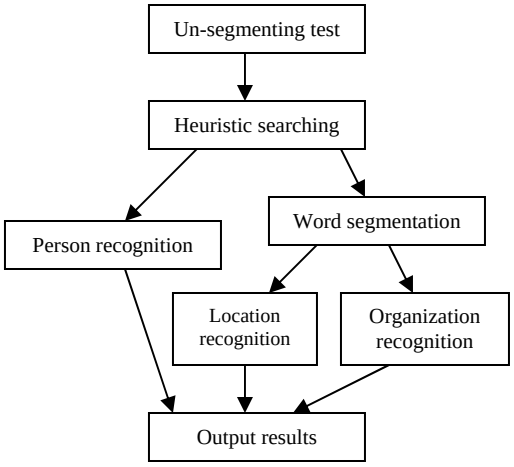

Figure 1. System Frameworks

3 System Details

3.1 Using heuristic method to search entity list

NER task meets many difficulties which come from the complexity of the construction of named entities. Named entities have flexible internal styles and external environments. Building a good model to describe the condition precisely will have many troubles. The statistical models we commonly used have some shortcomings, especially when they are adapted to a corpus of new domains or styles. We try to use an improved searching method to make up the relatively poor adaptability of the statistical models. The heuristic searching method is more flexible especially in the following two aspects. 1) Abbreviation can be matched. 2) Suffix in location and organization is universal, but it should not be taken in count when we search named entities.

The framework of the algorithm can be briefly described as follows:

1) Building an inverse index using Chinese characters as key term;
2) Using the text as a query to search for entities;
3) When comes terminal condition, a heuristic function is invoked to determine whether the character sequence is an entity;
4) When comes creation condition, a heuristic function is invoked to judge whether a new entity is created;
5) The labeled sequence is output.

Table 1. Heuristic Searching Method

One advantage of heuristic searching method is that the heuristic function can be set to fit a special corpus. The heuristic searching method we used in Bakeoff-4 is as follows:

- Ignoring the suffix key word in Location and Organization names. For example "同方/Tongfang 公司/Corporation" and "同方/Tongfang" will get same score under this heuristic rule.
- Ignoring the Location name as a prefix in an Organization name. For example, "美国/American 通用/General Motors " and "通用/General Motors" will get same score under this heuristic rule
- Taking Abbreviation rules in consideration. For example"北京/Peking 大学/University" can be abbreviated as "北/Bei 大/Da" rather than "北京/Peking" or "大学/University"

Heuristic searching method also has such advantages as follows:

It is easy to be expanded to a corpus of new domain or style. We only need to add the entities in the new domain into list

Searching method will improve the recall performance remarkably

But the precision will be reduced for the ambiguities, i.e. whether a sequence that matches an entity in the list really constructs an entity in the text. We will disambiguate it using statistical models.

3.2 Conditional Random Fields Model

Conditional Random Fields (CRFs) is an undirected graphical model that encodes a conditional probability distribution using a given set of features. Currently it is widely used as a discriminate model for sequence labeling:

$$P(Y \mid X) = \frac{1}{Z}\exp\left(\sum_{c \in C}\sum_{i=1}^{k} \lambda_i f_i(y_{c1}, y_{c2}, X)\right) \quad (1)$$

CRFs is considered to be a very effective model to resolve the issue of sequence labeling for the following characteristics:

Because it uses a non-greedy whole sentence joint labeling method, high accuracy rate can be guaranteed and bias labeling can be avoided.

Any types of features can be integrated in the model flexibly.

Over-fitting can be avoided to some extent by integrating a priori with training data.

As a discriminate model, CRFs inherits the advantages of both Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM) as well.

3.3 Person names recognize

We use language model to recognize Personal names. We use character-based rather than word-based model to avoid the word segmentation errors. We construct a context model and an entity model to respectively describe external and internal features of Personal names. The details of the model are as follows:

We use a tri-gram model as the context model:

$$P(WC) \approx \prod_{i=1}^{m} P\left(wc_i \mid wc_{i-2}\ wc_{i-1}\right) \qquad (2)$$

Entity model:

$$P\left(w_{wc_{i1}} \cdots w_{wc_{ik}} \mid wc_i\right) = P\left(w_{wc_{i1}} \cdots w_{wc_{ik}} \mid B_1 \overbrace{M_2 \cdots M_{k-1}}^{k-2} E_k\right) \qquad (3)$$

$$\cong P\left(w_{wc_{i1}} \mid B_1\right) \times \prod_{l=2}^{k-1} P\left(w_{wc_{il}} \mid M_l, w_{wc_{i(l-1)}}\right) \times P\left(w_{wc_{ik}} \mid E_k, w_{wc_{i(k-1)}}\right)$$

Where *B* means the beginning of the entity, *M* means middle, *E* means end.

Some expert knowledge is employed to assist the recognition process of language model.

- A Chinese family name list (476) and a Japanese family name list (9189) are used to restrict and select the generated candidates.
- A list of commonly used character in Russian and European name.
- Constrain of name length: A Chinese name cannot contain more than 8 characters.

3.4 Location names recognition

Location names have some composition characters. 1) There may be some key words as suffix, such as: "市/Shi, 镇/Zheng, 湖/Hu, 山/Shan" etc. 2) Other parts of Location names are always OOV words, such as "大岗村/Dagang Village, 夫子庙/Fuzi Temple". So the right boundary of Location can be determined easily. The mainly problem in Location recognition is on abbreviation, such as "晋冀鲁豫/JinJiLuYu" is the combination of four location abbreviations. In our system, CRFs model can be supported by the heuristic searching method because it can match the abbreviation of entity in list. Using the searching method can boost the recall rate of location recognition significantly. We construct the recognition model based on the word-segmented texts.

To construct a CRFs model, we select the following features:

- A list of key word suffix is used to trigger the recognition processing.
- Using a list of indication words to restrict the boundary.
- Heuristic searching method is used to assist Location recognition.

The features we used in CRFs model is followed:

W_0	Current Word
W_{-1}, W_{-2}, W_1, W_2	Two words before and behind
$W_{-1}W_0$, W_0W_1	Bi-gram Features
POS_0	POS tag
PRE_{-1}, PRE_0, PRE_1	Pre-Position reference words
SRE_{-1}, SRE_0, SRE_1	Suf-Position reference words
Key	Has Key suffix
DIC	In Dictionary

Table 2. Features used in Location recognition

Statement:

The indication words used in Location recognition include "for-indicate" and "back-indicate" words, where "for-indicate" denotes the indicating words that occur as the left neighbor of the candidate Location named entity, while "back-indicate" denotes the indicating words that occur as the right neighbor. "for-indicate" and "back-indicate" words are got from the training corpus. We calculate the mutual information between neighbor words and location entity, and get the top N words as indication words.

$$MI(x,y) = p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad (4)$$

We select 1216 for-indicate words and 1227 back-indicate words. We also get 607 key words as location name suffix.

3.5 Organization names recognition

Organization name recognition is the most difficult part in NER task. The difficulties are as follows. 1) The composition of Organization name is very complex. For example: "大连/Dalian 实德/Shide 集团/Group", the first words in the entity is a location name. The second is a phonetic name

which is also an OOV word and the last one is a key word as suffix.2) The boundary of organization name is hard to be classified, and the length of organization names is dynamitic. 3) Organization names are easily confused as Location names. We must use contextual information to determine its type. 4) To recognize the abbreviation of an organization is also a difficult task. So we choose the following features to solve the above problems.

- A list of key word suffix is used to trigger the recognition processing.
- Using indication words to define the boundary of organization.
- Heuristic searching method is used to assist Location recognition.

The features we used in the CRFs model are the same as used in Location model. We use the mutual information to select 513 for-indicate words and 1195 back-indicate words from training corpus. The number of key suffix words is 3129.

4 Experiments

We participate in the SigHAN Microsoft Research Asia (MSRA) corpus in open track. The table 3 is the official result of NER by our system.

	R	P	F
Person	0.9657	0.9574	0.9615
Location	0.9593	0.9769	0.968
Organization	0.8778	0.9338	0.9049
overall	0.9377	0.9603	0.9489

Table 3. SigHAN MSRA corpus test results

The training corpora we used comes from 1) 1998 People's Daily corpus; 2) the training corpus supplied by MSRA for SigHAN bakeoff 4. These two corpora have many difference and we focus on how to get a good performance both on training corpus and testing corpus. We select some general features and get assistance from the heuristic searching method. A good list is very important, which has been proved by the experimental data. We collect nearly 1 million personal names, 40 thousand location names and more than 300,000 organization names.

5 Conclusion

In the paper, we give a presentation to our Chinese Named Entity Recognition System. It uses a language mode to recognize personal names, and two CRFs models to find Location and organization separately. We also have a flexible heuristic searching method to match entity in named entity list with text characters sequence. Our system achieves a good result in the open NER track of MSRA corpus.

Acknowledgement

The work is supported by the National High Technology Development 863 Program of China under Grants no. 2006AA01Z144, the National Natural Science Foundation of China under Grants No. 60673042, the Natural Science Foundation of Beijing under Grants no. 4052027, 4073043.

References

J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data,* In Proc. ICML 2001.

F. Sha, F. Pereira. *Shallow parsing with conditional random fields.* In Proc. NAACL 2003

HP Zhang, HK Yu, DY Xiong, Q Liu and Liu Qun. *HHMM-based Chinese Lexical Analyzer ICTCLAS.* In Proc. Second of SIGHAN Workshop on Chinese Language Processing 2003.

Youzheng Wu, Jun Zhao, Bo Xu. *Chinese Named Entity Recognition Model Based on Multiple Features.* In Proceedings of HLT/EMNLP 2005

Youzheng Wu, Jun Zhao, Bo Xu. *Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge.* In Proceedings of ACL2003 Workshop on Multilingual and Mixed-language Named Entity Recognition