Chinese Named Entity Recognition with a Multi-Phase Model

Zhou Junsheng
State Key Laboratory for Novel Software Technology, Nanjing University, China
Deptartment of Computer Science, Nanjing Normal University, China
`zhoujs@nlp.nju.edu.cn`

He Liang
State Key Laboratory for Novel Software Technology, Nanjing University, China
`hel@nlp.nju.edu.cn`

Dai Xinyu
State Key Laboratory for Novel Software Technology, Nanjing University, China
`dxy@nlp.nju.edu.cn`

Chen Jiajun
State Key Laboratory for Novel Software Technology, Nanjing University, China
`chenjj@nlp.nju.edu.cn`

Abstract

Chinese named entity recognition is one of the difficult and challenging tasks of NLP. In this paper, we present a Chinese named entity recognition system using a multi-phase model. First, we segment the text with a character-level CRF model. Then we apply three word-level CRF models to the labeling person names, location names and organization names in the segmentation results, respectively. Our systems participated in the NER tests on open and closed tracks of Microsoft Research (MSRA). The actual evaluation results show that our system performs well on both the open tracks and closed tracks.

1 Introduction

Named entity recognition (NER) is a fundamental component for many NLP applications, such as Information extraction, text Summarization, machine translation and so forth. In recent years, much attention has been focused on the problem of recognition of Chinese named entities. The problem of Chinese named entity recognition is difficult and challenging, In addition to the challenging difficulties existing in the counterpart problem in English, this problem also exhibits the following more difficulties: (1) In a Chinese document, the names do not have "boundary tokens" such as the capitalized initial letters for a person name in an English document. (2) There is no space between words in Chinese text, so we have to segment the text before NER is performed.

In this paper, we report a Chinese named entity recognition system using a multi-phase model which includes a basic segmentation phase and three named entity recognition phases. In our system, the implementations of basic segmentation components and named entity recognition component are both based on conditional random fields (CRFs) (Lafferty et al., 2001). At last, we apply the rule method to recognize some simple and short location names and organization names in the text. We will describe each of these phases in more details below.

2 Chinese NER with multi-level models

2.1 Recognition Process

The input to the recognition algorithm is Chinese character sequence that is not segmented and the output is recognized entity names. The process of recognition of Chinese NER is illustrated in figure 1. First, we segment the text with a character-level CRF model. After basic segmentation, a small number of named entities in the text, such as "山西队", "新华社", "福建省" and so on, which are segmented as a single word. These simple single-word entities will be labeled with some rules in the last phase. However, a great number of named entities in the text, such as "中国绿色照明工程办公室", "西柏坡纪念馆", are not yet segmented as a single word. Then, different from (Andrew et al. 2003), we apply three trained CRFs models with carefully designed and selected features to label person names, location names and organization names in the segmentation results, respectively. At last phase, we apply some rules to tag some names not recognized by CRFs models, and adjust part of the organization names recognized by CRFs models.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 213–216,
Sydney, July 2006. ©2006 Association for Computational Linguistics

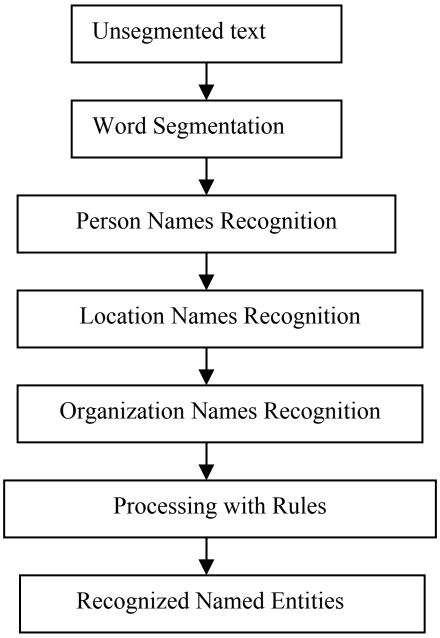
Unsegmented text
Word Segmentation
Person Names Recognition
Location Names Recognition
Organization Names Recognition
Processing with Rules
Recognized Named Entities

Fig1. Chinese NER process

2.2 Word segmentation

We implemented the basic segmentation component with linear chain structure CRFs. CRFs are undirected graphical models that encode a conditional probability distribution using a given set of features. In the special case in which the designated output nodes of the graphical model are linked by edges in a linear chain, CRFs make a first-order Markov independence assumption among output nodes, and thus correspond to finite state machines (FSMs). CRFs define the conditional probability of a state sequence given an input sequence as

$$P_{\mathrm{A}}(s\mid o)=\frac{1}{Z_o}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right)$$

Where $f_k(s_{t-1},s_t,o,t)$ is an arbitrary feature function over its arguments, and λ_k is a learned weight for each feature function.

Based on CRFs model, we cast the segmentation problem as a sequence tagging problem. Different from (Peng et al., 2004), we represent the positions of a *hanzi* (Chinese character) with four different tags: B for a *hanzi* that starts a word, I for a *hanzi* that continues the word, F for a *hanzi* that ends the word, S for a *hanzi* that occurs as a single-character word. The basic segmentation is a process of labeling each *hanzi* with a tag given the features derived from its surrounding context. The features used in our experiment can be broken into two categories: character features and word features. The character features are instantiations of the following templates, similar to those described in (Ng and Jin, 2004), *C* refers to a Chinese *hanzi*.

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

In addition to the character features, we came up with another type word context feature which was found very useful in our experiments. The feature captures the relationship between the *hanzi* and the word which contains the *hanzi*. For a two-*hanzi* word, for example, the first *hanzi* “连” within the word “连续” will have the feature WC0=TWO_F set to 1, the second *hanzi* “续” within the same word “连续” will have the feature WC0=TWO_L set to 1. For the three-*hanzi* word, for example, the first *hanzi* “梳” within a word “梳妆镜” will have the feature WC0=TRI_F set to 1, the second *hanzi* “妆” within the same word “梳妆镜” will have the feature WC0=TRI_M set to 1, and the last *hanzi* “镜” within the same word “梳妆镜” will have the feature WC0=TRI_L set to 1. Similarly, the feature can be extended to a four-*hanzi* word.

2.3 Named entity tagging with CRFs

After basic segmentation, we use three word-level CRFs models to label person names, location names and organization names, respectively. The important factor in applying CRFs model to name entity recognition is how to select the proper features set. Most of entity names do not have any common structural characteristics except for containing some feature words, such as “公司”, “学校”, “乡” , “镇” and so on. In addition, for person names, most names include a common surname, e.g. “张”, “王”. But as a proper noun, the occurrence of an entity name has the specific context. In this section, we only present our approach to organization name recognition. For example, the context information of organization name mainly includes the boundary words and some title words (e.g. 局长、董事长). By analyzing a large amount of entity name corpora, we find that the indicative intensity of different boundary words vary greatly. So we divide the left and right boundary words into two classes according to the indicative intensity. Accordingly we construct the four boundary words lexicons. To solve the problem of the selection and classification of boundary words, we make use of mutual Information *I(x, y).* If there is a genuine association between x and y, then I(x, y*)* >>0. If there is no interesting relationship between x and y, then I(x, y)≈0. If x and y are in complementary distribution, then I(x, y) << 0. By using mutual information, we compute the association between boundary word and the type of organization name, then select and classify the boundary words. Some example boundary words for organization names are listed in table 1.

Table 1. The classified boundary words for ORG names

Type	Class	Examples
Left boundary word	First-class	历任（6.0006）
	Second-class	接管（3.1161）
Right boundary word	First -class	管辖（5.4531）
	Second-class	规定（2.0135）

Based on the consideration given in preceding section, we constructed a set of atomic feature patterns, listed in table 2. Additionally, we defined a set of conjunctive feature patterns, which could form effective feature conjunctions to express complicated contextual information.

Table 2. Atomic feature patterns for ORG names

Atomic pattern	Meaning of pattern
CurWord	Current word
LocationName	Check if current word is a location name
PersonName	Check if current word is a person name
KnownORG	Check if current word is a known organization name
ORGFeature	Check if current word is a feature word of ORG name
ScanFeatureWord_8	Check if there exist a feature word among eight words behind the current word
LeftBoundary1_-2 LeftBoundary2_-2	Check if there exist a first-class or second-class left boundary word among two words before the current word
RightBoundary1_+2 RightBoundary2_+2	Check if there exist a first-class or second-class right boundary word among two words behind the current word

2.4 Processing with rules

There exists some single-word named entities that aren't tagged by CRFs models. We recognize these single-word named entities with some rules. We first construct two known location names and organization names dictionaries and two feature words lists for location names and organization names. In closed track, we collect known location names and organization names only from training corpus. The recognition process is described below. For each word in the text, we first check whether it is a known location or organization names according to the known location names and organization names dictionaries. If it isn't a known name, then we further check whether it is a known word. If it is not a known word also, we next check whether the word ends with a feature word of location or organization names. If it is, we label it as a location or organization name.

In addition, we introduce some rules to adjust organization names recognized by CRF model based on the labeling specification of MRSA corpus. For example, the string “阳城县李圪塔乡卫生院” is recognized as an organization name, but the string should be divided into two names: a location name (“阳城县”) and a organization name (“李圪塔乡卫生院”), according to label specification, so we add some rules to adjust it.

3 Experimental results

We participated in the three GB tracks in the third international Chinese language processing bakeoff: NER msra-closed, NER msra-open and WS msra-open. In the closed track, we constructed all dictionaries only with the words appearing in the training corpus. In the closed track, we didn't use the same feature characters lists for location names and organization names as in the open tracks and we collected the feature characters from the training data in the closed track. We constructed feature characters lists for location names and organization names by the following approach. First, we extract all suffix string for all location names and organization names in the training data and count the occurrence of these suffix strings in all location names and organization names. Second, we check every suffix string to judge whether it is a known word. If a suffix string is not a known word, we discard it. Finally, in the remaining suffix words, we select the frequently used suffix words as the feature characters whose counts are greater than the threshold. We set different thresholds for single-character feature words and multi-character feature words. Similar approaches were taken to the collection of common Chinese surnames in the closed track.

While making training data for segmentation model, we adopted different tagging methods for organization names in the closed track and in the open track. In the closed track, we regard every organization name, such as “内蒙古人民出版社”, as a single word. But, in the open track, we segment a long organization name into several words. For example, the organization name “内

蒙古人民出版社" would be divided into three words: "内蒙古", "人民" and "出版社". The different tagging methods at segmentation phase would bring different effect to organization names recognition. The size of training data used in the open tracks is same as the closed tracks. We have not employed any additional training data in the open tracks. Table 3 shows the performance of our systems for NER in the bakeoff.

Table 3: Named entity recognition outcome

Track	P	R	F	Per-F	Loc-F	Org-F
NER msra closed	88.94	84.20	86.51	90.09	85.45	83.10
NER msra open	90.76	89.22	89.99	92.61	90.99	83.97

For the separate word segmentation task(WS), the above NER task is performed first. Then we added several additional processing steps on the result of named entity recognition. As we all know, disambiguation problem is one of the key issue in Chinese words segmentation. In this task, some ambiguities were resolved through a rule-set which was automatically constructed based on error driven learning theory. The pre-constructed rule-set stored many pseudo-ambiguity strings and gave their correct segmentations. After analyzing the result of our NER based on CRFs model, we noticed that it presents a high recall on out-of-vocabulary. But at the same time, some characters and words were wrongly combined as new words which caused the losing of the precision of OOV and the recall of IV. To this phenomenon, we adopted an unconditional rule, that if a word, except recognized name entity, was detected as a new word and its length was more than 6 (Chinese Characters), and it should be segmented as several in-vocabulary words based on the combination of FMM and BMM methods. Table 4 shows the result of our systems for word segmentation in the bakeoff.

Table 4: Word segmentation outcome

Track	P	R	F	OOV-R	IV-R
WS msra open	0.975	0.976	0.975	0.811	0.981

4 Conclusion

We have presented our Chinese named entity recognition system with a multi-phase model and its result for Msra_open and mrsa_closed tracks. Our open and closed GB track experiments show that its performance is competitive. We will try to select more useful feature functions into the existing segmentation model and named entity recognition model in future work.

Reference

Aitao Chen. 2003. Chinese Word Segmentation Using Minimal Linguistic Knowledge. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing.

Andrew McCallum, Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. Proceedings of the Seventh CoNLL conference, Edmonton,

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML 01.

Ng, Hwee Tou and Jin Kiat Low. 2004. Chinese Part-of-Speech Taging: One-at-a-Time or All at Once? Word-based or Character based? In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Spain.

Peng, Fuchun, Fangfang Feng, and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields . In Proceedings of the Twentith International Conference on Computaional Linguistics.