Abstract

This paper proposes a hybrid Chinese named entity recognition model based on multiple features. It differentiates from most of the previous approaches mainly as follows. Firstly, the proposed Hybrid Model integrates coarse particle feature (POS Model) with fine particle feature (Word Model), so that it can overcome the disadvantages of each other. Secondly, in order to reduce the searching space and improve the efficiency, we introduce heuristic human knowledge into statistical model, which could increase the performance of NER significantly. Thirdly, we use three sub-models to respectively describe three kinds of transliterated person name, that is, Japanese, Russian and Euramerican person name, which can improve the performance of PN recognition. From the experimental results on People's Daily testing data, we can conclude that our Hybrid Model is better than the models which only use one kind of features. And the experiments on MET-2 testing data also confirm the above conclusion, which show that our algorithm has consistency on different testing data.

1 Introduction

Named Entity Recognition (NER) is one of the key techniques in the fields of Information Extraction, Question Answering, Parsing, Metadata Tagging in Semantic Web, etc. In MET-2 held in conjunction with the Seventh Message Understanding Conference (MUC-7), the task of NER is defined as recognizing seven sub-categories entities: person (PN), location (LN), organization (ON), time, date, currency and percentage. As for Chinese NEs, we further divide PN into five sub-classes, that is, Chinese PN (CPN), Japanese PN (JPN), Russian PN (RPN), Euramerican PN (EPN) and abbreviated PN (APN) like "吴先生/Mr. Wu". Similarly, LN is split into common LN (LN) like "中关村/Zhongguancun" and abbreviated LN (ALN) such as "京/Beijing", "沪/Shanghai". The recognition of time (TM) and numbers (NM) is comparatively simpler and can be implemented via finite state automata. Therefore, our research focuses on the recognition of CPN, JPN, RPN, EPN, APN, LN, ALN and ON.

Compared to English NER, Chinese NER is more difficult. We think that the main differences between Chinese NER and English NER lie in: (1) Unlike English, Chinese lacks the capitalization information which can play very important roles in identifying named entities. (2) There is no space between words in Chinese, so we have to segment the text before NER. Consequently, the errors in word segmentation will affect the result of NER.

In this paper, we proposes a hybrid Chinese NER model based on multiple features which emphasizes on (1) combining fine particle features (Word Model) with coarse particle features (POS Model); (2) integrating human knowledge into statistical model; (3) and using diverse sub-models for different kinds of entities. Especially, we divide transliterated person name into three sub-classes according to their characters set, that is, JPN, RPN and EPN. In order to deduce the complexity of the model and the searching space, we divide the rec-
ognition process into two steps: (1) word segmentation and POS tagging; (2) named entity recognition based on the first step.

Trained on the NEs labeled corpus of five-month People's Daily corpus and tested on one-month People's Daily corpus, the Hybrid Model achieves the following performance. The precision and the recall of PN (including CPN, JPN, RPN, EPN, AP N), LN (including ALN) and ON are respectively (94.06%, 95.21%), (93.98%, 93.48%), and (84.69%, 86.86%). From the experimental results on People's Daily testing data, we can conclude that our Hybrid Model is better than other models which only use one kind of features. And the experiments on MET-2 testing data also confirm the above conclusion, which show that our algorithm has consistence on different testing data.

2 Related Work

On the impelling of international evaluations like MUC, CoNLL, IEER and ACE, the researches on English NER have achieved impressive results. For example, the best English NER system[Chinchor. 1998] in MUC7 achieved 95% precision and 92% recall. However, Chinese NER is far from mature. For example, the performance (precision, recall) of the best Chinese NER system in MET-2 is (66%, 92%), (89%, 91%), (89%, 88%) for PN, LN and ON respectively.


Similarly, the models for Chinese NER can also be divided into two categories: Individual Model and Integrated Model.

Individual Model[Chen, et al. 1998][Sun, et al. 1994][Zheng, et al. 2000] consists of several sub-models, each of them deals with a kind of entities. For example, the recognition of PN may be statistical-based model, while LN and ON may be rule-based model like [Chen, et al. 1998]. Integrated Model[Sun, et al. 2002] [Zhang, et al. 2003][Yu, et al. 1998][Chua, et al. 2002] deals with all kinds of entities in a unified statistical framework. Most of these integrated models can be viewed as a HMM model. The differences among them are the definition of state and the features used in entity model and context model.

In fact, a NER model recognizes named entities through mining the intrinsic features in the entities and the contextual features around the entities. Most of existing approaches employ either coarse particle features, like POS and ROLE[Zhang, et al. 2003], or fine particle features like word. The data sparseness problem is serious if only using fine particle features, and coarse particle features will lose much important information though without serious data sparseness problem. Our idea is that coarse particle features should be integrated into fine particle features to overcome the disadvantages of them. However, most systems do not combine them and especially ignore the impact of POS.

Inspired by the algorithms of identifying BaseNP and Chunk[Xun, et al. 2000], we propose a hybrid NER model which emphasizes on combining coarse particle features (POS Model) with fine particle features (Word Model). Though the Hybrid Model can overcome the disadvantages of the Word Model and the POS Model, there are still some problems in such a framework. Data sparseness still exists and very large searching space in decoding will influence efficiency. Our idea is that heuristic human knowledge can not only improve the time efficiency, but also solve the data sparseness problem to some extent by restricting the generation of entity candidates. So we intend to incorporate human knowledge into the statistical model to improve efficiency and effectivity of the Hybrid Model.

Similarly, for capturing intrinsic features in different types of entities, we design several sub-models for each kind of entities. For example, we divide transliterated person name into three sub-classes according to their characters sets, that is, JPN, RPN and EPN.

3 Chinese NER with Multiple Features

Chinese NEs have very distinct word features in their composition and contextual information. For example, about 365 highest frequently used surnames cover 99% Chinese surnames[Sun, et al. 1994]. Similarly the characters used for transliterated names are also limited. LNs and ONs often
We try to introduce coarse particle feature to overcome the data sparseness problem. POS features are simplest and easy to obtain. Therefore, our hybrid model combines word feature with POS feature to recognize Chinese NEs.

Given a word/pos sequence as equation (1):

\[ W / T = w_1 / t_1, \ldots, w_n / t_n \]  

where \( n \) is the number of words and \( t_i \) is the POS of word \( w_i \). The task of Chinese NE identification is to find the optimal sequence \( W C^*/T C^* \) by splitting, combining and classifying the sequence of (1).

\[ WC^*/T C^* = w_{c_1} / t_{c_1}, \ldots, w_{c_m} / t_{c_m} \]  

where \( w_{c_i} = [w_{i_{1}}, \ldots, w_{i_{m_i}}] \), \( t_{c_i} = [t_{i_{1}}, \ldots, t_{i_{m_i}}] \), \( m \leq n \).

Note that the definition of words in \( \{w_i\} \) set is that each kind of NEs (including PN, APN, LN, ALN, ON, TM, NM) is defined as a word and all the other words in the vocabulary are also defined as individual words. Consequently, \( \{w_i\} \) set has \( |V| + 7 \) words, where \( |V| \) is the size of vocabulary. The size of \( \{t_i\} \) set is 48 which include PKU POS tagging set1 and each kind of NEs.

Obviously, we could obtain the optimal sequence \( WC^*/T C^* \) through the following three models: the Word Model, the POS Model and the Hybrid Model.

The Word Model employs word features for NER, which is introduced by [Sun, et al. 2002]. The POS Model employs POS features for NER. This paper proposes a Hybrid Model which combines word features with POS features.

We will describe these models in detail in following section.

3.1 The Hybrid Model

For the convenience of description, we take apart equation (1) into two components: word sequence as equation (3) and POS sequence as (4).

\[ W = w_1, w_2, \ldots, w_n \]  

\[ T = t_1, t_2, \ldots, t_n \]  

The Word Model estimates the probability of generating a NE from the viewpoint of word sequence, which can be expressed in equation (5).

\[ WC^* = \arg\max_{w_i} P(WC)P(W | WC) \]  

The POS Model estimates the probability of generating a NE from the viewpoint of POS sequence, which can be expressed in equation (6).

\[ TC^* = \arg\max_{t_i} P(TC)P(T | TC) \]  

Our proposed Hybrid Model combines the Word Model with the POS Model, which can be expressed in the equation (7).

\[ (WC^*, TC^*) = \arg\max_{(w_i, t_i)} P(WC, TC | W, T) \]

\[ = \arg\max_{(w_i, t_i)} P(WC, TC, W, T)P(W, T) \]

\[ = \arg\max_{(w_i, t_i)} P(WC, TC, W, T) \]

\[ = \arg\max_{(w_i, t_i)} P(W | WC)P(WC)P(T | TC)P(TC)^\zeta \]

where factor \( \zeta > 0 \) is to balance the Word Model and the POS Model.

Therefore, the Hybrid Model consists of four sub-models: word context model \( P(WC) \), POS context model \( P(TC) \), word entity model \( P(W | WC) \) and POS entity model \( P(T | TC) \).

3.2 Context Model

The word context model and the POS context model estimate the probability of generating a word or a POS given previous context. \( P(WC) \) and \( P(TC) \) can be estimated according to (8) and (9) respectively.

\[ P(WC) = \prod_{i=1}^{m} P(w_{c_{1}}, w_{c_{2}}, \ldots, w_{c_{i}}) \]  

\[ P(TC) = \prod_{i=1}^{n} P(t_{c_{1}}, t_{c_{2}}, \ldots, t_{c_{i}}) \]  

3.3 Word Entity Model

Different types of NEs have different structures and intrinsic characteristics. Therefore, a single model can't capture all types of entities. Typical, character-based model is more appropriate for PNs, whereas, word-based model is more competent for LNs and ONs. Especially, we divided transliterated PN into three categories such as JPN, RPN and EPN.

For the sake of estimating the probability of generating a NE, we define 19 sub-classes shown as Table 1 according to their position in NEs.

1 http://icl.pku.edu.cn/nlp-tools/catetkset.html
### Table 1 Sub-classes in Entity Model

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sur</td>
<td>Surname of CPN</td>
</tr>
<tr>
<td>Dgb</td>
<td>First character of Given Name of CPN</td>
</tr>
<tr>
<td>Dge</td>
<td>Last character of Given Name of CPN</td>
</tr>
<tr>
<td>Bfn</td>
<td>First character of EPN</td>
</tr>
<tr>
<td>Mfn</td>
<td>Middle character of EPN</td>
</tr>
<tr>
<td>Efn</td>
<td>Last character of EPN</td>
</tr>
<tr>
<td>RBfn</td>
<td>First character of RPN</td>
</tr>
<tr>
<td>RMfn</td>
<td>Middle character of RPN</td>
</tr>
<tr>
<td>REfn</td>
<td>Last character of RPN</td>
</tr>
<tr>
<td>JBfn</td>
<td>Surname of JPN</td>
</tr>
<tr>
<td>JMfn</td>
<td>Middle character of JPN</td>
</tr>
<tr>
<td>JEfn</td>
<td>Last character of JPN</td>
</tr>
<tr>
<td>Bol</td>
<td>First word of LN</td>
</tr>
<tr>
<td>Mol</td>
<td>Middle word of LN</td>
</tr>
<tr>
<td>Eol</td>
<td>Last word of LN</td>
</tr>
<tr>
<td>Aloc</td>
<td>Single character LN</td>
</tr>
<tr>
<td>Boo</td>
<td>First word of ON</td>
</tr>
<tr>
<td>Moo</td>
<td>Middle word of ON</td>
</tr>
<tr>
<td>Eoo</td>
<td>Last word of ON</td>
</tr>
</tbody>
</table>

### 3.3.1 Word Entity Model for PN

For the class of PN (including CPN, APN, JPN, RPN and EPN), the word entity model is a character-based trigram model which can be expressed in equation (10).

$$P(w_{w_{ij}} \ldots w_{w_{ik}} \mid w_{i}) = P\left(w_{w_{ij}} \ldots w_{w_{ik}} \mid BNe \ MNe \ldots \ MNe \ ENe\right)$$

(10)

$$P(w_{w_{ij}} \mid BNe) \times \prod_{i=2}^{k-1} P(w_{w_{ij}} \mid MNe, w_{w_{ij-1}})$$

where, BNe, MNe and ENe denotes the first, middle and last characters respectively.

The word entity models for PN are estimated with Chinese, Japanese, Russian and Euramerican names lists which contain 15.6 million, 0.15 million, 0.44 million, 0.4 million entities respectively.

### 3.3.2 Word Entity Model for LN and ON

For the class of LN and ON, the word entity model is a word-based trigram model. The model can be expressed by equation (11).

$$P(w_{w_{ij}} \ldots w_{w_{ik}} \mid w_{i}) = P\left(w_{w_{ij}} \ldots w_{w_{ik}} \mid BNe \ MNe \ldots \ MNe \ ENe\right)$$

(11)

$$P(w_{w_{ij}} \mid BNe) \times \prod_{i=2}^{k-1} P(w_{w_{ij}} \mid MNe, w_{w_{ij-1}})$$

$$\times \prod_{i=2}^{k-1} P(w_{w_{ij}} \mid ENe, w_{w_{ij-1}})$$

The word entity models and the POS entity model for LN and ON are estimated with LN and ON names lists which respectively contain 0.44 million and 3.2 million entities.

### 3.3.3 Word Entity Model for ALN

For the class of ALN, we use word-based bi-gram model. The entity model for ALN can be expressed by equation (12).

$$P(w_{i} \mid ALoc) = \frac{C(w_{i} \mid ALoc)}{C(ALoc)}$$

(12)

where $w_{i}$ is the ALN which includes single and multiple characters ALN.

### 3.4 POS Entity Model

But for the class of PN, it’s very difficult to obtain the corpus to train POS Entity Model. For the sake of simplification, we use word entity model shown in equation (10) to replace the POS entity model.

For the class of LN and ON, POS entity model can be expressed by equation (13).

$$P(t_{i} \mid ALoc) = \frac{C(t_{i} \mid ALoc)}{C(ALoc)}$$

(13)

While for the class of ALN, POS entity model is shown as equation (14).

$$P(t_{i} \mid ALoc) = \frac{C(t_{i} \mid ALoc)}{C(ALoc)}$$

(14)

### 4 Heuristic Human Knowledge

In this section, we will introduce heuristic human knowledge that is used for Chinese NER and the
method of how to incorporate them into statistical
model which are shown as follows.

1. CPN surname list (including 476 items) and
JPN surnames list (including 9189 items): Only
those characters in the surname list can trigger per-
son name recognition.

2. RPN and EPN characters lists: Only those
consecutive characters in the transliterated charac-
ter list form a candidate transliterated name.

3. Entity Length Restriction: Person name can-
not span any punctuation and the length of CN
cannot exceed 8 characters while the length of TN
is unrestrained.

4. Location keyword list (including 607 items): If
the word belongs to the list, 2~6 words before the
salient word are accepted as candidate LNs.

5. General word list (such as verbs and preposi-
tions): Words in the list usually is followed by a
location name, such as "在/at", "去/go". If the cur-
current word is in the list, 2~6 words following it are
accepted as candidate LNs.

6. ALN name list (including 407 items): If the
current word belongs to the list, we accept it as a
candidate ALN.

7. Organization keyword list (including 3129
items): If the current word is in organization key-
word list, 2~6 words before keywords are accepted
as the candidate ONs.

8. An organization name template list: We
mainly use organization name templates to recog-
nize the missed nested ONs in the statistical model.

Some of these templates are as follows:
\[ \text{ON} \rightarrow \text{LN} \ D^* \text{OrgKeyWord} \]
\[ \text{ON} \rightarrow \text{PN} \ D^* \text{OrgKeyWord} \]
\[ \text{ON} \rightarrow \text{ON} \text{OrgKeyWord} \]
D and OrgKeyWord denote words in the middle
of ONs and ONs keywords. D* means repeating
zero or more times.

5 Back-off Model to Smooth

Data sparseness problem still exists. As some pa-
rameters were never observed in training corpus,
the model will back off to a less powerful model.
The escape probability[Black, et al. 1998] was ad-
opted to smooth the statistical model shown as (15).

\[
p(W_N | W_1 \cdots W_{N-1}) = \lambda_N p(W_N | W_1 \cdots W_{N-1}) + \lambda_{N-1} p(W_{N-1} | W_1 \cdots W_{N-2}) + \cdots + \lambda_1 p(W_1) + \lambda_0 p_0
\]

where \( \lambda_N = 1 \), \( \lambda_i = (1 - e_i) \sum_{e_i, 0 < i < N} \), and \( e_i \)
is the escape probability which can be estimated by
equation (16).

\[
e_N = \frac{q(w_1w_2\cdots w_{N+1})}{f(w_1w_2\cdots w_{N+1})}
\]

\[ q(w_1w_2\cdots w_{N+1}) \] in (16) denotes the number of dif-
ferent symbol \( w_N \) that have directly followed the
word sequence \( w_1w_2\cdots w_{N+1} \).

6 Experiments

In this chapter, we will conduct experiments to
answer the following questions.

Will the Hybrid Model be more effective than
the Word Model and the POS Model? To answer
this question, we will compare the performances of
models with different parameter \( \zeta \) and find the best
value of \( \zeta \) in equation (7).

Will the conclusion from different testing sets be
consistent? To answer this question, we evaluate
models on the MET-2 test data and compare the
performances of the Word Model, the POS Model
and the Hybrid Model.

Will the performance be improved significantly
after combining human knowledge? To answer this
question, we compare two models with and with-
out human knowledge.

In our evaluation, only NEs with correct
boundaries and correct categories are considered as
the correct recognition. We conduct evaluations in
terms of precision, recall and F-Measure. Note that
PNs in experiments includes all kinds of PNs and
LNs include ALNs.

6.1 Will the Hybrid Model be More Effective
Than the Word Model and POS Model?

The parameter \( \zeta \) in equation (7) denotes the balanc-
ing factor of the Word Model and the POS Model.
The larger \( \zeta \), the larger contribution of the POS
Model. The smaller \( \zeta \), the larger contribution of the
Word Model. So the task of this experiment is to
find the best value of \( \zeta \). In this experiment, the
training corpus is from five-month's People's Daily
tagged with NER tags and the testing set is from
one-month's People's Daily.

With the change of \( \zeta \), the performances of rec-
ognizing PNs are shown in Fig.1.

Note that the left, middle and right point in ab-
scissa respectively denote the performance of the
Word Model, the Hybrid Model and the POS Model.

From Fig. 1, we can find that the performances of recognizing PNs are improved with the increasing of ζ in the beginning stage but decline in the ending. This experiment shows that the Word Model and the POS Model can overcome their disadvantages, and it is a feasible approach to integrate the Word Model and the POS Model in order to improve the performance PNs recognition.

With the change of ζ, the performances of recognizing LNs are shown in Fig. 2.

As the Fig. 2 shows, the precision and recall of LNs are improved with the increasing of ζ and decreased in the later stage. This phenomenon also proves that the Hybrid Model is better for recognizing LN than either the Word Model or the POS Model.

Similarly, with the change of ζ, the performances of recognizing ONs are shown in Fig. 3.

Comparing Fig. 3 with Fig. 1 and Fig. 2, we find that the POS Model has different impact on recognizing ONs from that on recognizing PNs and LNs. Especially, the POS Model has obvious side-effect on the recall. We speculate that the reasons may be that the probability of generating POS sequence by POS entity model is lower than that by POS context model.

According to Fig. 1~Fig. 3, we choose the best value ζ = 2.8. And the performances of different models are shown in Table 2 in detail.

<table>
<thead>
<tr>
<th></th>
<th>P(%)</th>
<th>R(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ζ = 2.8) PN</td>
<td>94.06</td>
<td>95.21</td>
<td>94.63</td>
</tr>
<tr>
<td>LN</td>
<td>93.98</td>
<td>93.48</td>
<td>93.73</td>
</tr>
<tr>
<td>ON</td>
<td>84.69</td>
<td>86.86</td>
<td>85.76</td>
</tr>
<tr>
<td>Word Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PN</td>
<td>88.24</td>
<td>90.11</td>
<td>89.16</td>
</tr>
<tr>
<td>LN</td>
<td>91.50</td>
<td>93.17</td>
<td>92.32</td>
</tr>
<tr>
<td>ON</td>
<td>78.85</td>
<td>88.77</td>
<td>83.52</td>
</tr>
<tr>
<td>POS Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PN</td>
<td>93.44</td>
<td>95.11</td>
<td>94.27</td>
</tr>
<tr>
<td>LN</td>
<td>89.97</td>
<td>92.20</td>
<td>91.07</td>
</tr>
<tr>
<td>ON</td>
<td>80.90</td>
<td>69.29</td>
<td>74.65</td>
</tr>
</tbody>
</table>

Table 2 Performance of the Hybrid Model, the Word Model and the POS Model

From Table 2, we find that the F-Measures of the Hybrid Model for PN, LN, ON are improved by 5.4%, 1.4%, 2.2% respectively in comparison with the Word Model, and these F-Measures are improved by 0.4%, 2.7%, 11.1% respectively in comparison with the POS Model.
Conclusion 1: The experimental results validate our idea that the Hybrid Model can improve the performance of both the Word Model and the POS Model. However, the improvements for PN, LN and ON are different. That is, the POS Model has obvious side-effect on the recall of ON recognition at all times, while the recalls for PN and ON recognition are improved in the beginning but decreased in the ending with the increasing of $\zeta$.

6.2 Will the Conclusion from Different Testing Sets be Consistent?

We also conduct experiments on the MET-2 testing corpus to validate our conclusion from Exp.1, that is, the Hybrid Model could achieve better performance than either the Word Model or the POS Model alone. The experimental results (F-Measure) on MET-2 are shown in Table 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Word Model</th>
<th>Hybrid Model</th>
<th>POS Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN</td>
<td>75.21%</td>
<td>80.77%</td>
<td>76.61%</td>
</tr>
<tr>
<td>LN</td>
<td>89.78%</td>
<td>90.95%</td>
<td>89.81%</td>
</tr>
<tr>
<td>ON</td>
<td>76.30%</td>
<td>80.21%</td>
<td>76.83%</td>
</tr>
</tbody>
</table>

Table 3 F-Measure on MET-2 test corpus

Comparing Table 3 with Table 2, we find that the performances of models on MET-2 are not as good as that on People Daily's testing data. The main reason lies in that the NE definitions in People Daily's corpus are different from that in MET-2. However, Table 3 can still validate our conclude 1, that is, the Hybrid Model is better than both the Word Model and the POS Model. For example, the F-Measures of the Hybrid Model for PN, LN and ON are improved by 5.6%, 1.2% and 3.9% respectively in comparison with the Word Model, and these F-Measures are improved by 4.2%, 3.1% and 3.4% respectively in comparison with the POS Model.

Conclusion 2: Though the performances of the Hybrid Model on MET-2 are not as good as that on People's Daily's corpus, the experimental results also support conclusion 1, i.e. the Hybrid Model which combining the Word Model with the POS Model can achieve better performance than either the Word Model or the POS Model.

6.3 Will the Performance be Improved Significantly after Incorporating Human Knowledge?

<table>
<thead>
<tr>
<th></th>
<th>P(%)</th>
<th>R(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN</td>
<td>91.81</td>
<td>70.65</td>
<td>79.85</td>
</tr>
<tr>
<td>LN</td>
<td>79.47</td>
<td>88.83</td>
<td>83.89</td>
</tr>
<tr>
<td>ON</td>
<td>64.95</td>
<td>80.63</td>
<td>71.95</td>
</tr>
</tbody>
</table>

Model I

<table>
<thead>
<tr>
<th></th>
<th>P(%)</th>
<th>R(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN</td>
<td>94.06</td>
<td>95.21</td>
<td>94.63</td>
</tr>
<tr>
<td>LN</td>
<td>93.98</td>
<td>93.48</td>
<td>93.73</td>
</tr>
<tr>
<td>ON</td>
<td>84.69</td>
<td>86.86</td>
<td>85.76</td>
</tr>
</tbody>
</table>

Model II

Table 4 Performances Impacted by Human Knowledge

From Table 4, we find that F-Measure of model with human knowledge (Model II) is improved by 14.8%, 9.8%, 13.8% for PN, LN and ON respectively compared with that of the model without human knowledge (Model I).

Conclusion 3: From this experiment, we learn that human knowledge can not only reduce the search space, but also significantly improve the performance of pure statistical model.

7 Conclusion

In this paper, we propose a hybrid Chinese NER model which combines multiple features. The main contributions are as follows: ① The proposed Hybrid Model emphasizes on integrating coarse particle feature (POS Model) with fine particle feature (Word Model), so that it can overcome the disadvantages of each other; ② In order to reduce the search space and improve the efficiency of model, we incorporate heuristic human knowledge into statistical model, which could increase the performance of NER significantly; ③ For capturing intrinsic features in different types of entities, we design several sub-models for different entities. Especially, we divide transliterated person name into three sub-classes according to their characters set, that is, CPN JPN, RPN and EPN.

There is a lack of effective recognition strategy for abbreviated ONs such as 昆明机床(Kunming Machine Tool Co., Ltd), 凤 凰光学 (Phoenix Photonics Ltd) in this paper. And most of mis-
recognized ONs in current system belong to them. So in the future work, we will be focusing more on recognizing abbreviated ONs.

8 Acknowledgements

This research is carried out as part of the cooperative project with Fujitsu R&D Center Co., Ltd. We would like to thank Yingju Xia, Fumihito Nisino for helpful feedback in the process of developing and implementing. This work was supported by the Natural Sciences Foundation of China under grant No. 60372016 and 60272041, the Natural Science Foundation of Beijing under grant No. 4052027.

References


CoNLL. http://cnts.uia.ac.be/conll2004/


ACE. http://www.itl.nist.gov/iad/894.01/tests/ace/