Chinese Named Entity Recognition Based on Multiple Features

Youzheng Wu, Jun Zhao, Bo Xu
National Laboratory of Pattern Recognition
Institute of Automation, CAS
Beijing, 100080, China
(yzwu,jzhao,bxu)@nlpr.ia.ac.cn

Hao Yu
Fujitsu R&D Center Co., Ltd
Beijing 100016, China
yu@frdc.fujitsu.com

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 consistence 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-

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 427–434, Vancouver, October 2005. ©2005 Association for Computational Linguistics

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.

Recently, approaches for NER are a shift away from handcrafted rules[Grishman, et al. 1995] [Krupka, et al. 1998][Black et al. 1998] towards machine learning algorithms, i.e. unsupervised model like DL-CoTrain, CoBoost[Collins, 1999, 2002], supervised learning like Error-driven [Aberdeen, et al. 1995], Decision Tree [Sekine, et al. 1998], HMM[Bikel, et al. 1997] and Maximum Entropy[Borthwick, et al. 1999][Mikheev, et al.1998].

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 end with the specific words like "省/province" and "公司/company". However, data sparseness is very serious when using word features. So 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):

$$W/T = w_1/t_1 \cdots w_i/t_i \cdots w_n/t_n \quad (1)$$

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 WC^*/TC^* by splitting, combining and classifying the sequence of (1).

$$WC^*/TC^* = wc_1/tc_2 \cdots wc_i/tc_i \cdots wc_m/tc_m \quad (2)$$

where $wc_i = \lfloor w_j \cdots w_{j+l} \rfloor$, $tc_i = \lfloor t_j \cdots t_{j+l} \rfloor$, $m \le n$.

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^*/TC^* 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).

$$W = w_1\ w_2 \cdots w_i \cdots w_n \quad (3)$$

$$T = t_1\ t_2 \cdots t_i \cdots t_n \quad (4)$$

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

$$WC^* = argmax_{wc} P(WC)P(W \mid WC) \quad (5)$$

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

$$TC^* = argmax_{TC} P(TC)P(T \mid TC) \quad (6)$$

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

$$\begin{aligned}
&(WC^*, TC^*) \\
&= argmax_{(WC,TC)} P(WC, TC \mid W, T) \\
&= argmax_{(WC,TC)} P(WC, TC, W, T) P(W, T) \\
&= argmax_{(WC,TC)} P(WC, TC, W, T) \\
&\approx argmax_{(WC,TC)} P(W \mid WC) P(WC) [P(T \mid TC) P(TC)]^{\xi}
\end{aligned} \quad (7)$$

where factor ζ > 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.

$$P(WC) = \prod_{i=1}^{m} P(wc_i \mid wc_{i\ \ 2}\ wc_{i\ \ 1}) \quad (8)$$

$$P(TC) = \prod_{i=1}^{m} P(tc_i \mid tc_{i\ \ 2}\ tc_{i\ \ 1}) \quad (9)$$

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

Tag	Description
Sur	Surname of CPN
Dgb	First character of Given Name of CPN
Dge	Last character of Give Name of CPN
Bfn	First character of EPN
Mfn	Middle character of EPN
Efn	Last character of EPN
RBfn	First character of RPN
RMfn	Middle character of RPN
REfn	Last character of RPN
JBfn	surname of JPN
JMfn	Middle character of JPN
JEfn	Last character of JPN
Bol	First word of LN
Mol	Middle word of LN
Eol	Last word of LN
Aloc	Single character LN
Boo	First word of ON
Moo	Middle word of ON
Eoo	Last word of ON

Table 1 Sub-classes in Entity Model

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).

$$\begin{aligned}&P\left(w_{wc_{i1}} \cdots w_{wc_{ik}} \mid wc_i\right)\\&= P\left(w_{wc_{i1}} \cdots w_{wc_{ik}} \mid BNe\ \overbrace{MNe \cdots MNe}^{k-2}\ ENe \right)\\&\cong P\left(w_{wc_{i1}} \mid BNe\right) \times \prod_{l=2}^{k-1} P\left(w_{wc_{il}} \mid MNe, w_{wc_{i(l-1)}}\right)\\&\times P\left(w_{wc_{ik}} \mid ENe, w_{wc_{i(k-1)}}\right)\end{aligned} \tag{10}$$

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 (11).

$$\begin{aligned}&P\left(w_{wc_i\ \ start} \cdots w_{wc_i\ \ end} \mid wc_i\right)\\&= P\left(wc_{wc_{i1}} \cdots wc_{wc_{il}} \cdots wc_{wc_{ik}} \mid BNe\ \overbrace{MNe \cdots MNe}^{k\;\;2}\ ENe \right)\\&= P\left(wc_{wc_{i1}} \mid BNe\right) P\left(w_{wc_{i1}\ \ start} .. w_{wc_{i1}\ \ end} \mid wc_{wc_{i1}}\right)\\&\times \prod_{l=2}^{k\;\;1} P\left(wc_{il} \mid MNe, wc_{i(l\;\;1)}\right) P\left(w_{wc_{il}\ \ start} \cdots w_{wc_{il}\ \ end} \mid wc_{wc_{il}}\right)\\&\times P\left(wc_{wc_{ik}} \mid ENe, wc_{wc_{i(k\;\;1)}}\right) P\left(w_{wc_{ik}\ \ start} \cdots w_{wc_{ik}\ \ end} \mid wc_{ik}\right)\end{aligned} \tag{11}$$

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 mil-lion 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).

$$P\left(w_i \mid ALoc\right) = \frac{C\left(w_i, ALoc\right)}{C(\,ALoc\,)} \tag{12}$$

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).

$$\begin{aligned}&P\left(t_{tc_i\ \ start} \cdots t_{tc_i\ \ end} \mid tc_i\right)\\&= P\left(tc_{tc_{i1}} \cdots tc_{tc_{il}} \cdots tc_{tc_{ik}} \mid BNe\ \overbrace{MNe \cdots MNe}^{k\;\;2}\ ENe \right)\\&= P\left(tc_{tc_{i1}} \mid BNe\right) P\left(t_{tc_{i1}\ \ start} .. t_{wc_{i1}\ \ end} \mid tc_{tc_{i1}}\right)\\&\times \prod_{l=2}^{k\;\;1} P\left(tc_{il} \mid MNe, tc_{i(l\;\;1)}\right) P\left(t_{tc_{il}\ \ start} \cdots t_{tc_{il}\ \ end} \mid tc_{tc_{il}}\right)\\&\times P\left(tc_{tc_{ik}} \mid ENe, tc_{tc_{i(k\;\;1)}}\right) P\left(t_{tc_{ik}\ \ start} \cdots t_{tc_{ik}\ \ end} \mid tc_{ik}\right)\end{aligned} \tag{13}$$

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

$$P\left(t_i \mid ALoc\right) = \frac{C\left(ti, ALoc\right)}{C(\,ALoc\,)} \tag{14}$$

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 person name recognition.

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

3. Entity Length Restriction: Person name cannot 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 prepositions): Words in the list usually is followed by a location name, such as "在/at", "去/go". If the 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 keyword 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 recognize the missed nested ONs in the statistical model. Some of these templates are as follows:

ON-->LN D OrgKeyWord*
ON-->PN D OrgKeyWord*
ON-->ON 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 parameters were never observed in training corpus, the model will back off to a less powerful model. The escape probability[Black, et al. 1998] was adopted to smooth the statistical model shown as (15).

$$\hat{p}(W_N|W_1\cdots W_{N-1}) = \lambda_N p(W_N|W_1\cdots W_{N-1}) + \lambda_{N-1} p(W_N|W_2\cdots W_{N-1}) + \cdots + \lambda_1 p(W_N) + \lambda_0 p_0 \qquad (15)$$

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

$$e_N = \frac{q(W_1W_2\cdots W_{N-1})}{f(W_1W_2\cdots W_{N-1})} \qquad (16)$$

$q(w_1w_2...w_{N-1})$ in (16) denotes the number of different symbol w_N that have directly followed the word sequence $w_1w_2...w_{N-1}$.

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 ζ and find the best value of ζ 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 without 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 ζ in equation (7) denotes the balancing factor of the Word Model and the POS Model. The larger ζ, the larger contribution of the POS Model. The smaller ζ, the larger contribution of the Word Model. So the task of this experiment is to find the best value of ζ. 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 ζ, the performances of recognizing PNs are shown in Fig.1.

Note that the left, middle and right point in abscissa respectively denote the performance of the

Word Model, the Hybrid Model and the POS Model.

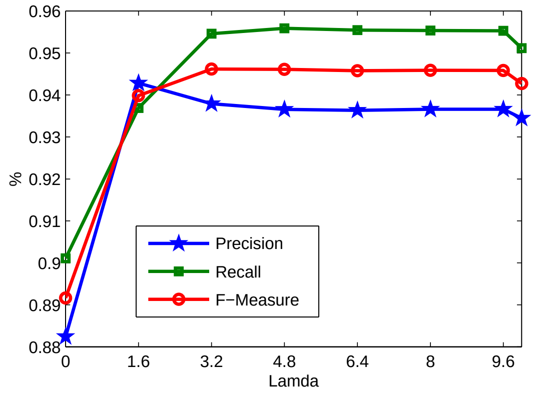

Fig.1 Performance of Recognizing LNs Impacted by ζ

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.

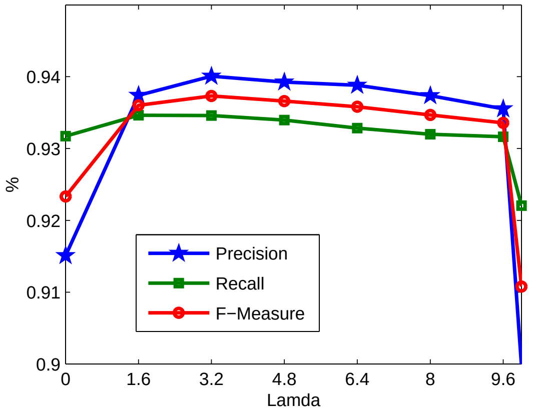

Fig.2 Performance of Recognizing LNs Impacted by ζ

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.

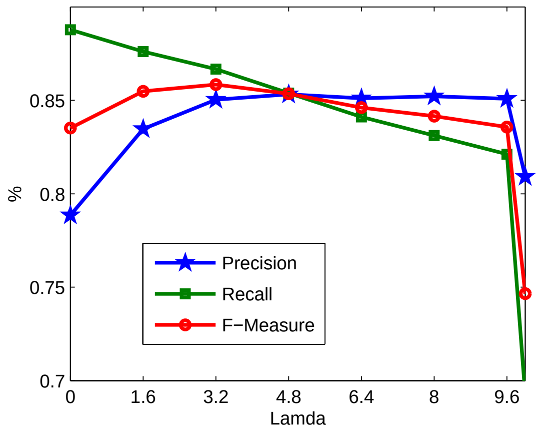

Fig.3 Performance of Recognizing LNs Impacted by ζ

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.

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

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 ζ.

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.

Model	Word Model	Hybrid Model	POS Model
PN	75.21%	**80.77**%	76.61%
LN	89.78%	**90.95**%	89.81%
ON	76.30%	**80.21**%	76.83%

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 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?

One of our ideas in this paper is that human knowledge can not only reduce the search space, but also improve the performance through avoiding generating the noise NEs. This experiment will be conducted to validate this idea. Table 4 shows the performances of models with and without human knowledge.

		P(%)	R(%)	F(%)
Model I	**PN**	91.81	70.65	79.85
	LN	79.47	88.83	83.89
	ON	64.95	80.63	71.95
Model II				
	PN	94.06	95.21	94.63
	LN	93.98	93.48	93.73
	ON	84.69	86.86	85.76

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 misrecognized 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

N.A. Chinchor: Overview of MUC-7/MET-2. In: Proceedings of the Seventh Message Understanding Conference (MUC-7), April. (1998).

Youzheng Wu, Jun Zhao, Bo Xu: Chinese Named Entity Recognition Combining Statistical Model with Human Knowledge. In: The Workshop attached with 41st ACL for Multilingual and Mix-language Named Entity Recognition, Sappora, Japan. (2003) 65-72.

Endong Xun, Changning Huang, Ming Zhou: A Unified Statistical Model for the Identification of English BaseNP. In: Proceedings of ACL-2000, Hong Kong. (2000).

Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang: Chinese Named Entity Identification Using Class-based Language Model. In: COLING 2002. Taipei, August 24-25. (2002).

Huaping Zhang, Qun Liu, Hongkui Yu, Xueqi Cheng, Shuo Bai: Chinese Named Entity Recognition Using Role Model. In: the International Journal of Computational Linguistics and Chinese Language Processing, vol.8, No.2. (2003) 29-60.

D.M. Bikel, Scott Miller, Richard Schwartz, Ralph Weischedel: Nymble: a High-Performance Learning Name-finder. In: Fifth Conference on Applied Natural Language Processing, (published by ACL). (1997) 194-201.

Borthwick .A: A Maximum Entropy Approach to Named Entity Recognition. PhD Dissertation. (1999).

Mikheev A., Grover C. and Moens M: Description of the LTG System Used for MUC-7. In: Proceedings of 7th Message Understanding Conference (MUC-7), 1998.

Sekine S., Grishman R. and Shinou H: A decision tree method for finding and classifying names in Japanese texts. In: Proceedings of the Sixth Workshop on Very Large Corpora, Canada, 1998.

Aberdeen, John, et al: MITRE: Description of the ALEMBIC System Used for MUC-6. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), November. (1995) 141-155.

Ralph Grishman and Beth Sundheim: Design of the MUC-6 evaluation. In: 6th Message Understanding Conference, Columbia, MD. (1995)

Krupka, G. R. and Hausman, K. IsoQuest: Inc.: Description of the NetOwl TM Extractor System as Used for MUC-7. In Proceedings of the MUC-7, 1998.

Black, W.J.; Rinaldi, F, Mowart, D: FACILE: Description of the NE System Used for MUC-7. In Proceedings of the MUC-7, 1998.

Michael Collins, Yoram Singer: Unsupervised models for named entity classification. In Proceedings of EMNLP. (1999)

Michael Collins: Ranking Algorithms for Named Entity Extraction: Boosting and the Voted Perceptron. In: Proceeding of ACL-2002. (2002) 489-496.

S.Y.Yu, et al: Description of the Kent Ridge Digital Labs System Used for MUC-7. In: Proceedings of the Seventh Message Understanding Conference, 1998.

H.H. Chen, et al: Description of the NTU System Used for MET2. In: Proceedings of the Seventh Message Understanding Conference.

Tat-Seng Chua, et al: Learning Pattern Rules for Chinese Named Entity Extraction. In: Proceedings of AAAI'02. (2002)

Maosong Sun, et al: Identifying Chinese Names in Unrestricted Texts. Journal of Chinese Information Processing. (1994).

Jiahen Zheng, Xin Li, Hongye Tan: The Research of Chinese Names Recognition Methods Based on Corpus. In: Journal of Chinese Information Processing. Vol.14 No.1. (2000).

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

IEER. http://www.nist.gov/speech/tests/ie-er/er99/er99.htm

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