# A Novel Dynamic Adaptive Method Based on Artificial Immune System in Chinese Named Entity Recognition

*Wei Jiang, Yi Guan, and XiaoLong Wang*
JiangWei@insun.hit.edu.cn
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

**Summary**
Named Entity recognition (NER), as a task of providing important semantic information, is a critical first step in Information Extraction and Question Answer system. The NER has been proved to be NP-hard problem, and the existing methods usually adopt supervised or unsupervised learning model, as a result, there is still a distance away from the required performance in real application, however the system can hardly improved with the model being applied. The paper proposes a novel method based on artificial immune system (AIS) for NER. We apply clonal selection principle and affinity maturation of the vertebrate immune response, where the secondary immune response has high performance than the primary immune response, and the similar antigens may have a good immunity. We also introduce the reinforcement learning method into our system to tune the immune response, and the context features are exploited by the maximum entropy principle. The experimental results indicate that our method exhibits a good performance and implements the dynamic learning behavior.

***Key words:***
*Named Entity Recognition, Artificial Immune System, Maximum Entropy Principle, Reinforcement Learning, QA system.*

## 1. Introduction

Named entity recognition (NER) is part of the common message understanding tasks. The objective is to identify and categorize all members of certain categories of "proper names". In MUC-7, there are seven categories: person, organization, location, date, time, percentage, and monetary amount. Named entities (NE) are broadly distributed in original texts from many domains, e.g. politics, sports, and economics. It can answer for us many questions: "who", "where", "when", "what", "how much", and "how long" [1]. So it is an essential process widely required in natural language understanding and many other text-based applications, such as question answering, information retrieval, and information extraction.

NER is also an important subtask of the Multilingual Entity Task (MET), as for NE in Chinese, we further divide PER into two sub classes: Chinese PER and transliterated PER on the basis of their distinct features. In this work, we only focus on those more difficult but commonly used categories: PER, LOC and ORG. Other NE such as times and quantities, in a border sense, can be recognized simply via finite state automata, less aided with a disambiguation algorithm.

There have been a number of conferences aimed at evaluating NER, e.g. MUC-6, MUC-7, CoNLL-2002 and CoNLL-2003, and on-going ACE (automatic content extraction) evaluations. Recent research on English NER has focused on the machine learning approaches[2], which include Hidden Markov Model (HMM) [3], Maximum Entropy Model (ME) [4], Conditional Random Fields (CRF) [5], or even the combining model [6]. Some other methods are also tried, such as SVM, AdaBoost, Memory based Learning, Robust Risk Minimization based learning, and Transformation based Learning [7]. So nearly all the available supervised methods related to NER task, have been tried and evaluated.

Compared with English NER, Chinese NER is more complicated and difficult. These approaches that are successfully applied in English cannot be simply transplanted into Chinese. Unlike English, there are no delimiters to mark word boundaries and no explicit definitions of words in Chinese. Generally speaking, Chinese NER has two sub tasks: locating the string of NE and identifying its category. The above machine learning algorithms also be applied to Chinese NER task, such as Role based HMM NER[1], Class based language model [8], Agent-based approach [9], ME and CRF [10], combining method [11]. In addition, Guo combines the ME and the Robust Risk Minimization (RRM) method to NER, so as to fuse multilevel linguistic features. Compared with supervised approaches, few unsupervised method is tried, for the natural language is complicated and there is more ambiguities in Chinese than those in English.

This paper proposes a novel reinforcement learning method for Chinese NER task, and we try to utilize the feedback information to further improve the NER performance. The extensive evaluation on NER systems in recent years (such as in CoNLL and ACE), indicate the following points. Firstly, the best statistical systems are typically achieved by using a linear classification algorithm [2], such as Maximum Entropy model, together with a vast amount of carefully designed linguistic features. Secondly, the combining model can not bring more improvement, e.g. there is 88.76%-88.31% = 0.45% improvement in term of F1 measure in CoNLL-2003.

Manuscript received April 25, 2006.
Manuscript revised April 30 , 2006.

Thirdly, many kinds of features have been tried and these delicate features do not bring more improvement. So the performance with supervised method seems not easy to achieve dramatic improvement, or even has neared the limitation of best possible perfomance. In this case, we try to address NER task from another angle of view, especially in large-scale web information processing.

We mainly focus on the following three respects: 1) Build a NER system based on Artificial Immune System, and take advantage of high recognition performance in the secondary immune response and in the similar antigens. 2) Present a reinforcement learning method to adjust our system by using the feedback information. Since the research on supervised method is not easy to bring further improvement, we attempt to exploit the feedback information, which may come from successive processing or the users, just as there is a dramatic success that the Information Retrieval uses the feedback information [12]. 3) We propose to extend the features words by combining the word cluster method with the thesaurus, so as to overcome the sparse data problem and make our system more robust. The experiments indicate that our method can effectively use the feedback information. And our work can also bring a good reference to other Asian language.

The paper is organized as follows. In the next section we present related works and basic theory. Section 3 presents our language model for Chinese NER, and Section 4 presents the feature extraction and feature extension algorithm. Section 5 gives a test for the new method, and Section 6 presents the conclusions.

## 2. Related Works

### 2.1 The clonal selection algorithm

When an animal is exposed to an antigen (Ag), some subpopulation of its bone marrow derived cells (B lymphocytes) respond by producing antibodies (Ab), Each cell secretes only one kind of antibody, which is relatively specific for the antigen. By binding to these antibodies (receptors), and with a second signal from accessory cells, such as the T-helper cell, the antigen stimulates the B cell to proliferate (divide) and mature into terminal (non-dividing) antibody secreting cells, called plasma cells. The various cell divisions (mitosis) generate a clone, i.e., a set of cells that are the progeny of a single cell. While plasma cells are the most active antibody secretors, large B lymphocytes, which divide rapidly, also secrete Ab, albeit at a lower rate [13].

Lymphocytes, in addition to proliferating and/or differentiating into plasma cells, can differentiate into long-lived B memory cells. When exposed to a second antigenic stimulus, Memory cells commence to differentiate into large lymphocytes capable of producing high affinity antibodies, pre-selected for the specific antigen that had stimulated the primary response. Fig. 1 depicts the clonal selection principle.

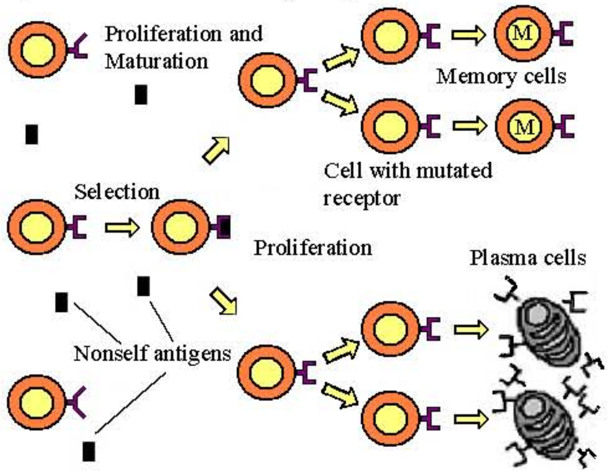


Fig. 1 The clonal selection principle

Affinity maturation is the whole mutation processes and selection of the variant offspring that better recognizes the antigen [14]. The two basic mechanisms of affinity maturation are those: hypermutation and receptor editing. In our task, the hypermutation is guided, since there are some principles in NE, such as the common character of person name or some the law to construct NE. And we adopt receptor editing to eliminate the bad candidate.

### 2.2 The reinforcement learning process

When we apply the feedback information to adjust our model, how much adjustment should do? Before answer this question, we should make sense of the following main questions: 1) which type is the available feedback information? Obviously, the whole feedback will make our adjustment easier, however, that is not easily met in real application. So we suppose the feedback type to be partial reception type. In this case, probability model and some common online learning algorithm cannot be applied in our task. 2) how to evaluate our adjustment? Since the entity distribution is sparse in the corpus or in real web data, the feedback is not real time, we will take a measure of the aggregate future rewards, namely we adopt reinforcement learning method to evaluate our behavior.

Partial reception makes our task harder than whole feedback. Different from the optimization problem, the population size of AIS in our task is unfixed, and the clonal process and some dynamic behavior should be evaluated by a long-term reward function. There are four main problems in our task: 1) How to obtain the feedback information? 2) How to implement a dynamic behavior? 3) How to fuse the context features? 4) How to decide the generalization performance?

In the research, our feedback information mainly comes from the answer template in Question Answer task or Information Extraction task, to acquire the feedback information [15]. For instance, a location name may comes from the answer template "which city is the biggest in China". This paper will focus on the latter three key problems, which will be detailed in the following sections.

## 3. NER based on reinforcement learning

### 3.1 The NER system structure

Fig. 2 illuminates the structure of our system, in which, we emphasize to exploit the feedback information.

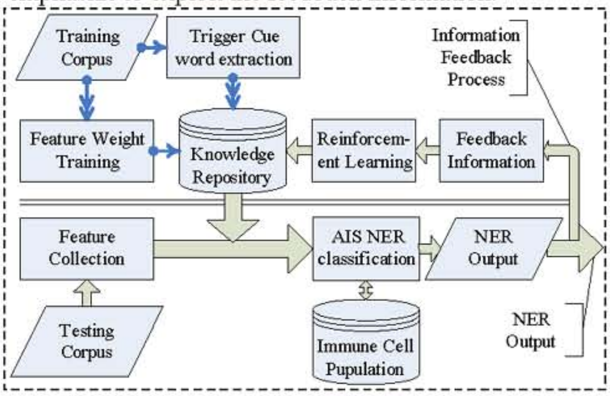


Fig. 2 Our AIS system structure

There are three main modules in our system: the linguistic features fusing, AIS NER classification module and feedback learning, the feature extraction.

### 3.2 Fusing linguistic features based on maximum entropy principle

We use $w_i$ ($i$=0,1,…n) to denote the input sequence, then every token $w_i$ should be assigned a tag $t_i$. We adopt B-I-O encoding, e.g., B-CPN, I-CPN is the beginning of Chinese person name and the continue part of person name respectively. Though NER is treated as sequential tagging task, a linear classification algorithm seems to have a better performance according to the evaluation in recent years [2]. Our evaluation in Chinese NER also complies with this empirical viewpoint, so we fuse all kinds of features based on Maximum Entropy principle.

Maximum Entropy(ME)[16] is a general technique to estimate probability distributions from the training data. The overriding principle in maximum entropy is that when nothing is known, the distribution should be as uniform as possible, that is, have maximal entropy. ME estimates the conditional distribution of the class label $t_i$ given $w_i$ and its context. It is defined over $H \times T$ in the classification task, where $H$ is the set of possible contexts about $w_i$ that will be tagged, and $T$ is the set of allowable tags. Then the model's conditional probability is defined as

$$p(t \mid h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')} \tag{1}$$

where $p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$ is a normalization factor.

The feature about $w_i$ is represented as

$$f_j(h,t) = \begin{cases} 1 & if((pd \text{ in } h)\, and\, (t = tag)) \\ 0 & \end{cases} \tag{2}$$

where $pd$ is the predicate extracted from the document, and the $tag$ represents the category tag.

The conditional entropy of a distribution p($t|h$) is

$$H(p) = -\sum_{t \in \mathrm{T}, h \in \mathrm{H}} \tilde{p}(h)\, p(t \mid h) \log p(t \mid h) \tag{3}$$

By means of maximizing the conditional entropy subject to certain constraints, $p(t|h)$ is estimated by maximum entropy theory, The constraints are defined as

$$P(p) = \{E_p f_j = E_{\tilde{p}} f_j, \forall f_j\} \tag{4}$$

$$\sum_t p(t \mid h) = 1 \tag{5}$$

$f_j$ is the feature function as shown in Eq. 2. $E_p f_j$ is the model's expectation of $f_j$. $E_{\tilde{p}} f_j$ is the empirical expectation of the feature $f_j$. They are defined as

$$E_p f_j = \sum_{t,h} \tilde{p}(h) p(t|h) f_j(t,h) \tag{6}$$

$$E_{\tilde{p}} f_j = \sum_{t,h} \tilde{p}(t,h) f_j(t,h) \tag{7}$$

And $\alpha$ in Eq. 1 is the multiplier parameter with respect to each feature function, which can be estimated by Improved Iterative Scaling algorithm [16]. When $\alpha$ being given, the classification function is log-linear, according to Eq. 1. Another advantage is that ME can relax feature independency assumption, which is suffered from by generative model, such as Hidden Markov Model.

### 3.3 The NER model based on Improved AIS

Table 1 The contrast between AIS and NER task

| AIS | NER task |
|---|---|
| Antigen | Named entity |
| Immune cell | Entity Information |
| Clonal process | Knowledge update in NER |
| Receptor editing | Eliminate malfunction entities |
| Memory cell | Long-term entity information |

Name entity recognition based on Artificial Immune System can be viewed as a process that AIS detect the nonself antigens. The main contrast between AIS and NER task is depicted in table 1.

The attractive advantages of AIS for NER at less include three points: (1) There is no central organ controlling the functioning of the immune system, and only a part of immune cells responses the antigen, so that the clonal selection has an effect on a few of immune cells. So the NER system can perform reinforcement learning and only response the feedback information; (2) long-term memory function, i.e. the secondary immune response nearly have a better performance than the performance in the primary immune response; (3) There is immunity to the similar antigens. This make the reinforcement learning in the NER have a good generalization performance.

There is a little percentage to the multi-classes in the similar fields, as demonstrated in Fig.3 (about the CPN).

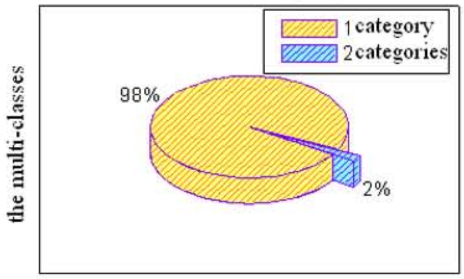


Fig. 3 The case of entity multi-classes

This result is done in Chinese People's Daily Newspaper in half year of 1998, which were annotated by Institute of Computational Linguistic of Peking University corpora. There is a similar case to other entity types. The entity distribution in the corpora is shown in Table 2.

Table 2 The entity distribution in People's Daily

| Named Enity | CPN | FPN | LOC | ORG |
|---|---|---|---|---|
| By entities | 27.54% | 8.86% | 41.53% | 22.07% |
| By corpus | 1.29% | 0.41% | 1.94% | 1.03% |
| Occur Count | 92941 | 29912 | 140162 | 74483 |

The lower multi-classes attribution and the lower occurrence in the corpora, will mainly interpret the antigen pattern. The former indicates that, when AIS as a core learning module, it is usually unnecessary that the antigen own a much complicated pattern. And the latter indicated that it is not absolutely necessary to deal with all the information in order to identify an entity.

In this paper, the immune cell pattern in an $L$-dimensional shape-space is described to a string $S = <s_1, s_2,\ldots,s_L>$, we adopt $S$ (where $s \in S^L$) to represent any immune cell or molecule. Each $s_i (i=1,2,\ldots,L)$ is a variable description to the entity property. For instance, we make use of the character in person name to denote $s_i$, while we may also use a sub-string in organization entity, since there are some kinds of generic structure to make up of organization name, such as "Location + a sub-string + organization suffix". So, $s_i$ is used in mutation process, and is useful to generalization performance.

The maturation mechanism of the immune response is another problem needed by AIS shown in fig. 1. Firstly, let us consider which factors affect the clonal proliferation. The somatic hypermutation should be relevant to the affinity, however, our task is different from the travelling salesman problem [14], because there is the explicit target measure function in the latter task, e.g. the shortest distance in [14]. No matter how to measure, a basic principle is important, it is that these new mutation cells should have more probability to increase the whole affinity of Artificial Immune System.

Let $C_i=\{c_1,c_2,\ldots,c_n\}$ represent the candidate set, which is corresponding to $s_i$ in the immune cell pattern, then we adopt the roulette wheel method to generate the mutation candidate, since our method can make the system have more chance to generate a new entity with a higher affinity. On the other hand, our model will have better performance gradually, in this case, the proliferation process may be bring more bad mutation entity.

After cloning and mutation processes, a percentage of the antibodies of the conflict population are eliminated, since we have exhibited the case about entity multi-classes in Fig. 3. This mechanism is also a vertebrate immune system mechanism. The probability to be eliminated is

$$\mathrm{P(immune\ cell}\,c) = \frac{\mathrm{Coflict\ Count}(c)}{\mathrm{Occurrence\ Count}(c)} \qquad (8)$$

These eliminated entities are selected by the roulette wheel way, and the selection probability is Eq. 8.

To evaluate the adjustment is complicated problem, especially to partial perception task. One basic principle to enhance the weight, if the current behavior is useful, on the contrary, to decrease the weight. Since the clone process need seek a balance between precision and recall, we specify a threshold, named generalization factor $\alpha$, which describes the percentage of the improper proliferation immune cells, and we may adopt a long-term reward value to evaluate the generalization process

$$V^{\pi}(s^t) = \sum_{i=0}^{h} r_{t+i} \qquad (9)$$

A window size $h$ is to measure the future cumulative reward, depicted in Eq.9. When the feedback signal isn't immediate, we have to evaluate the behavior according to the performance by a period of time. Obviously, we had better to decrease $\alpha$ according to an attenuation function.

## 4. Feature extraction and feature extension

### 4.1 The context features

The "ugly duckling theorem" has denoted that there is no generic feature extraction method suitable to all kinds of tasks. The basic feature template is shown as

Table 3 NER feature template[1]

| Type | Feature Template |
|---|---|
| one order feature | $w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$ |
| two order feature | $w_{i-1:i}, w_{i:i+1}$ |
| NER tag feature | $t_{i-1}$ |

In addition, in order to solve the unstable feature collection problem caused by no delimiter beside Chinese word, inspired by the term extraction in text classification [18], we construct a novel feature template of "word->tag" to extract the trigger features, which have a flexible distance between the two units. The *t* test and Pearson's chi-square test, which can be used to find correlations with fixed distance, can't be applied here.

Mutual information (MI) measure the independent between a trigger word and a NE type, it is defined as

$$\mathrm{MI}(W, C) = \log \frac{P(W \wedge C)}{P(W) \times P(C)} \qquad (10)$$

where P(W) represents the probability of the trigger word, P(C) is the probability of the corresponding NE category. However this method does not consider the influence when lacking one point. On contrast, average mutual information (AMI) is defined as:

$$\begin{aligned}\mathrm{AMI}(W, C) = {} & P(W, C)\log\frac{P(C \mid W)}{P(C)} \\ & + P(W, \overline{C})\log\frac{P(\overline{C} \mid W)}{P(\overline{C})} \\ & + P(\overline{W}, C)\log\frac{P(C \mid \overline{W})}{P(C)} \\ & + P(\overline{W}, \overline{C})\log\frac{P(\overline{C} \mid \overline{W})}{P(\overline{C})} \qquad (11)\end{aligned}$$

We can explain this formula from two sides: 1) the first expression in the Eq. 11 include MI which is a measure of reduction in uncertainty of C due to knowing about W. but AMI still include the measure of reduction in uncertainty of W due to knowing about other word. 2) MI in fact is point wise information, while AMI can be look as a Kullback-Leibler (KL) divergence:

$$\mathrm{AMI}(X, Y) = D(P(X, Y) \parallel P(X) \times P(Y)) \qquad (12)$$

Formula (12) measures how different two probability distributions between $P(X, Y)$ and $P(X) \times P(Y)$. But MI only is a point of the whole distributions.

Under this case, let m be the number of the possible categories count, the average mutual information is

$$\mathrm{AMI}_{avg}(W) = \sum_{i=1}^{m} P(C_i) \times \mathrm{AMI}(W, C_i) \qquad (13)$$

or another optional formula adopt in this paper:

$$\mathrm{AMI}_{max}(W) = \mathrm{MAX}_{i=1}^{m} \mathrm{AMI}(W, C_i) \qquad (14)$$

[1] $w_i$ – current word, $w_{i-1}$ – previous word, $t_i$ – current tag.

We select the top triggers with higher AMI value, and acquire the trigger words.

### 4.2 The entity features

Besides context features, entity features is also very important in the NER task, such as the suffix of Location or Organization. We perform statistic in the foreign resource, including the corpora and the collected entity name in the Internet. We build 8 kinds of dictionary [19]:

Table 4 NER resource dictionary[2]

| List Type | Lexicon | Example |
|---|---|---|
| Word list | Place lexicon | 北京, 纽约, 马家沟 |
| | Chinese surname | 张, 王, 赵, 欧阳 |
| String list | Prefix of PER | 老, 阿, 小 |
| | Suffix of PLA | 山, 湖, 寺, 台, 海 |
| | Suffix of ORG | 会, 联盟, 组织, 局 |
| Character list | Character for CPER | 军, 刚, 莲, 茵, 倩 |
| | Character for FPER | 科, 曼, 斯, 娃, 贝 |
| | Rare character | 滗, 胨, 薅 |

These entity features are a good hint to entity recognition. In addition, we identify factoid, including time, number, Email, web site, telephone etc, detail in [11].

### 4.3 The feature extension

Feature extension is used to overcome the sparse data problem and to increase robustness. In addition, semantic and pragmatic knowledge is useful to the language processing, e.g., if we know "教授" (professor) is a good hint to label person name, these similar words {老师(teacher),助教(assistant), 讲师(lecturer)}, should have the same effect. So, we build semantic class by combining word cluster and the thesaurus.

A vector for word *w* is derived from the close neighbors of *w* in the corpus. Close neighbors are all words that co-occur with *w* in a sentence or a larger context. The entry for word *v* in the vector for *w* records the number of times that word *v* occurs close to *w* in the corpus. We refer this vector space to as Word Space.

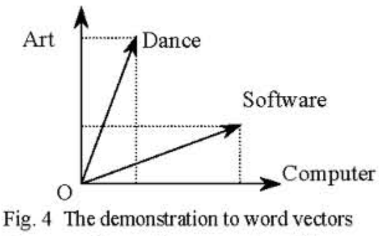


Fig. 4 The demonstration to word vectors

Fig. 4 gives a schematic example of two words being represented in a two-dimensional space. This vector

[2] Partial translation: 北京 BeiJing,纽约 New york;张 Zhang, 王 Wang; 老 Old;山 mountain,湖 lake;局 bureau.

representation captures the typical topic or subject matter of a word. By looking at the amount of overlap between two vectors, one can roughly determine how closely they are related semantically. This is because related meanings are often expressed by similar sets of words. Semantically related words will therefore cooccur with similar neighbors and their vectors will have considerable overlap.

We combine the basic semantic word in a thesaurus--HOWNET2005 with the TF-IDF algorithm, and use a frequency cutoff to select the 2000 words to serve as the dimensions of Word Space. Compared with the traditional TF-IDF method, our method increasing the taxonomy information, so our method can give a better measure to the word similarity [20].

After constructing word vectors, the similarity can be measured by the cosine between two vectors. The cosine is equivalent to the normalized correlation coefficient:

$$corr(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \quad (15)$$

The word cluster algorithm in the word vectors is used to measure the similarity by totaling the pragmatic knowledge from the corpora.

## 5. Experiments

In the following experiments: The precision P = the right count / the model count, the recall rate R = the right count / the corpus count, and F-measure = (2 * P * R) / (P + R).

### 4.1 Compared experiment

The experimental corpora come from Chinese People's Daily Newspaper in the first half-year in 1998, which were annotated by Institute of Computational Linguistic of Peking University, including CPN-92941, FPN-29912, LOC-140162, ORG-74483, and Other words-7212467.

Table 5 The comparison of several NER models[3]

| Model | Precision | Recall | F-measure |
|---|---|---|---|
| BaseLine | 68.99% | 73.54% | 71.19% |
| HMM | 79.20% | 79.96% | 79.58% |
| ME | 84.77% | 83.23% | 83.99% |
| HMM + Role | 83.68% | 85.20% | 84.43% |
| ME + Role | 87.95% | 84.62% | 86.25% |
| Our method | 88.31% | 86.17% | 87.23% |

ELUS word segmentation system [11] is used to re-segment the corpora, so as to make our evaluation more reasonable. The baseline result is obtained by selecting the NER tag that is most frequently associated with the current word. The role includes the entity prefix, infix and suffix. Table 5 indicates that our novel method is better than traditional supervised method, since we fuse the feedback information based on reinforcement learning.

Table 6 The comparison of several NER models

| Model | Precision | Recall | F-measure |
|---|---|---|---|
| HMM model + Role | 84.70 | 89.71 | 87.13 |
| ME model + Role | 89.16 | 87.54 | 88.34 |
| CRF model + Role | 82.81 | 65.56 | 73.18 |
| AIS-Primary response | 89.39 | 52.92 | 66.48 |
| AIS-Secondary response | 92.01 | 78.34 | 84.63 |

In order to evaluate in detail, Table 6 shows the comparison in detail in terms of CPN. While, don't fuse the context features based on maximum entropy principle, so at to well evaluate the AIS performance. Result shows that though CRF is good at overcoming label bias problem, CRF isn't better then ME, since the entity occurrence is sparse problem, and the CRF has the bad ability of capturing the locality of phenomena. Compared with the supervised models, AIS is based on reinforcement learning method, and may further improve the performance.

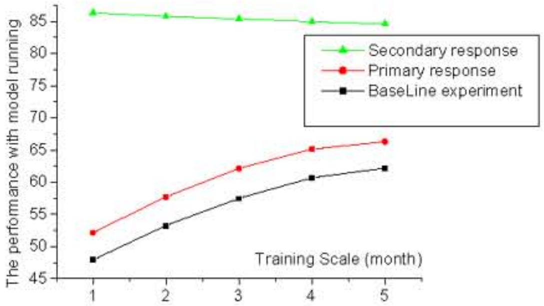


Fig. 5 The performance improving with the system running

This experiment indicates that the performance of AIS method itself can further improved with the system running, this is the main difference from the supervised method, when being used as large-scale web information processing task. Fig. 6 compares twice immune response.

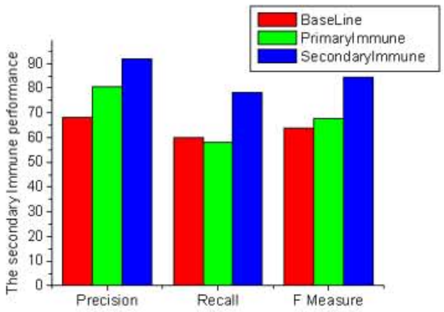


Fig. 6 The primary and the secondary immune response

The secondary immune response has a better

[3] In this experiment, HMM is one order model, and ME, CRF use the feature template: $W_{-2}, W_{-1}, W_0, W_1, W_2, W_{-1:0}, W_{0:1}, T_{-1}$.

performance than the primary immune response. This manifests the performance to multi-users or repetitious use in web application, such as Question Answer system.

### 4.2 Feature extraction and extension experiment

The corpora come from People's Daily Newspaper Corpus, including six months in 1998 and twelve months in 2000. And they also be segmented by ELUS system.

Table 7 Trigger pairs draw from corpora

| Pair | | AMI | | MI | |
|---|---|---|---|---|---|
| | | Value | Rank | Value | Rank |
| 同志 comrade | CPN | 3.9e-4 | 6 | 2.71 | 144 |
| 说 say | CPN | 2.3e-4 | 11 | 1.85 | 885 |
| 主任 director | ORG | 1.2e-4 | 23 | 2.63 | 181 |
| 会见 interview | CPN | 1.1e-4 | 27 | 2.43 | 269 |
| 举行 hold | LOC | 9.5e-5 | 34 | 1.61 | 1279 |
| 会议 meeting | ORG | 3.8e-5 | 83 | 1.39 | 1650 |
| 教授 professor | CPN | 3.1e-5 | 96 | 2.21 | 463 |

We select the top 500 trigger features. Table 8 shows the compared performance with trigger.

Table 8 The performance with AMI trigger

| Entity type | ME (%) | | | ME+AMI(%) | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| CPN | 84.54 | 77.71 | 80.98 | 86.36 | 82.41 | 84.34 |
| FPN | 73.27 | 53.21 | 61.65 | 78.50 | 56.90 | 65.97 |
| LOC | 86.95 | 76.53 | 81.41 | 87.57 | 77.62 | 82.30 |
| ORG | 74.87 | 55.29 | 63.61 | 74.08 | 60.95 | 66.88 |
| Overall | 82.81 | 69.74 | 75.71 | 83.60 | 72.97 | 77.92 |

Table 8 gives the detail comparison between ME and ME with AMI trigger features, training in first month corpus, and testing in the sixth month corpus. The overall improvement is 2.21% in terms of F-measure.

Then, we perform the word vectors method in a large-scale corpus, in 1998 and 2000 People's Daily Newspaper, the window of size $k$ =8 in this experiment.

Table 9 the proximity matrix[4]

| Case | Cosine of Vectors of Values | | | | | |
|---|---|---|---|---|---|---|
| | 学生 | 教授 | 副教授 | 导师 | 大学生 | 中学生 |
| 学生 | 1.000 | .352 | .280 | .288 | .433 | .331 |
| 教授 | .352 | 1.000 | .722 | .815 | .310 | .174 |
| 副教授 | .280 | .722 | 1.000 | .641 | .216 | .136 |
| 导师 | .288 | .815 | .641 | 1.000 | .226 | .139 |
| 大学生 | .433 | .310 | .216 | .226 | 1.000 | .674 |
| 中学生 | .331 | .174 | .136 | .139 | .674 | 1.000 |

The hierarchical cluster analysis or other cluster analysis methods can be used to obtain the word cluster result. Fig. 7 demonstrates the hierarchical cluster analysis result. We use a synonym dictionary "Word Forest Of The Synonym" to reduce the cluster space and increase the cluster precision. For instance, there are about 63 synonyms to the word "教授" (professor).

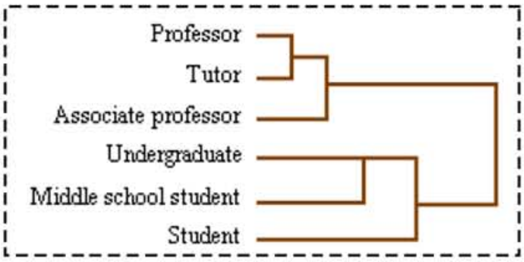


Fig. 7 The demonstration about hierarchical cluster

Since we use maximum entropy principle to fuse the context features, here, the above AMI and Word Space approach can also improve our new model performance.

## 6. Conclusion and Future work

In this paper, we proposed a novel dynamic adaptive method for Chinese Named Entity Recognition. With the growing number of text documents in the Web or in the real application, the performance of the traditional supervised learning method is not easy to achieve dramatic improvement. The evaluation of many researches on recent year has denoted that the improvement in the language model or the utilization in the new features can hardly bring a satisfying result, since the natural language has been proved to be NP-hard problem and it is more complicated for NER in Chinese than in English.

From a new angle of view, we exploit the feedback information and apply the reinforcement learning to improve our model. Different from the whole perception feedback, our model only requires the partial perception information, which is a more complicated task, but is requested in real application. In addition, we apply Average Mutual Information to acquire cue words and extend these words by combining the word cluster with the thesaurus. All the features are fused together, based on maximum entropy principle. The experiments show that our proposed methods are effective. Meanwhile, our work can also bring a good reference to other Asian language, such as the Japanese NER.

Still further research remains. Firstly, this study uses answer template in Question Answer task or Information Extraction task, to acquire the feedback information. Future works on more feedback source should be investigated. Secondly, we will apply text classification to reduce the multi-classes phenomenon. In addition, we will transplant our method into other proper name recognition, such as military proper name recognition.

---

[4] 学生 student, 教授 professor, 副教授 associate professor, 导师 tutor, 大学生 undergraduate, 中学生 middle school student.

**Acknowledgments**

This research was supported by National Natural Science Foundation of China (60435020, 60504021) and Key Project of Chinese Ministry of Education & Microsoft Asia Research Centre (01307620).

## References

[1] Huaping Zhang, Q Liu, HK Yu et al. Chinese Named Entity Recognition Using Role Model. Computational Linguistics and Chinese Language Processing. 2003: 8(2), p.29-60.

[2] Hongelei Guo, Jianmin Jiang, Gang Hu, et al. Chinese Named Entity Recognition Based on Multilevel Linguistic Features. First International Joint Conference on Natural Language Processing (IJCNLP-04), 2004: p.294-301.

[3] Dan Klein, Joseph Smarr. Huy Nguyen, and Christopher D. Manning. Named Entity Recognition with Character-level models. In Proceedings of CoNLL-2003. 2003: p.180-183.

[4] Hai Leong Chieu and Hwee Tou Ng. Named Entity Recognition with a Maximum Entropy Approach. In: Proceedings of CoNLL-2003. 2003: p.160-163.

[5] Andrew McCallum and Wei Li. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In: Proceedings of CoNLL-2003. 2003: p.188-191.

[6] Radu Florian, Abe Ittycheriah, Hongyan Jing et al. Named Entity Recognition through Classifier Combination. In: Proceedings of CoNLL-2003. 2003: p.168-171.

[7] Erik F. Tjong Kim Sang and Fien De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada, 2003: p.142-147.

[8] Jian Sun, Jianfeng Gao, Lei Zhang, et al. Chinese Named Entity Identification Using Class-based Language Model. CoNLL-2002. 2002: p. 967-973.

[9] Shiren Ye. An Agent-based Approach to Chinese Named Entity Recognition. CoNLL-2002. 2002: p.194-199.

[10] Zhao Jian. Research on Conditional Probabilistic Model And Its Application in Chinese Named Entity Recognition. [phD. dissertation]. Harbin Institute of Technlogy, 2006.

[11] Wei Jiang, Jian Zhao, Yi Guan et al. Chinese Word Segmentation based on Mixing Model. The 4th SIGHAN Workshop. 2005: p. 180-182.

[12] C. Kemp and K. Ramamohanarao. Long-term learning for web search engines. In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002). 2002: p.263-274.

[13] L. N. de Castro and J. Timmis. Artificial immune systems: a novel paradigm for pattern recognition. In: Artificial Neural Networks in Pattern Recognition. 2002: p.67-84.

[14] De Castro, L. N. & Von Zuben, F. J. The Clonal Selection Algorithm with Engineering Applications. GECCO'00 Workshop Proceedings. 2000: p.36-37.

[15] Wei Jiang, Yi Guan, Xiaolong Wang et al. Acquiring Feedback information for Named Entity Recognition based on Reinforcement learning. Computer Engineering and Application (checking). 2006.

[16] Rosenfeld. R, A Maximum Entropy Approach to Adaptive Statistical Language Modeling, Ph.D. thesis. Carnegie Mellon University. 1994.

[17] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. Journal of Artificial Intelligence Research. Vol. 4, 1996: p.237-285.

[18] Qiang Wang, Yi Guan, XiaoLong Wang et al. A Novel Feature Selection Method Based on Category Information Analysis for Class Prejudging in Text Classification. International Journal of Computer Science and Network Security (IJCSNS). Vol.6 No.1A. 2006: p.113-119.

[19] Jian Zhao, XiaoLong Wang, Yi Guan et al. Chinese Named Entity Recognition: an Approach Base on Conditional Random Fields Using Word Triggers Information. Chinese High Technology Letters (to be published). 2006.

[20] Yan Zhao, XiaoLong Wang, Yi Guan et al. Solution Strategies for Word Sense Problems Based On Vector Space Model and Maximum Entropy Model. Chinese High Technology Letters. Vol.1. 2005: p.1-6.

**Wei Jiang** reived the B.S.degrees in School of Computer Science and Technology from ChangChun University of Science and Technology in 2000 and now is a PH.D candidate in Harbin Institute of Technology. In 2005, he participated in The Second International Chinese Word Segmentation Bakeoff (SIGHAN2005) and ranked the top in open test in MSR corpus and got second in open test in PKU corpus. He is interested in the Theories and Methods for Question answering, Machine learning and Information Extraction.

**Yi Guan** holds a B.Sc. degree in Computer Science and Technology from Tianjin University in 1992, and a Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology in 1999. In 1996, Dr. GUAN was an invited visiting scholar in Canotec Co.,Japan. In 2000, Dr. GUAN was research associate in Human Language Technology Center at Hong Kong University of Science and Technology. In October 2001, he became an associate professor in School of Computer Science and Technology in Harbin Institute of Technology. Dr. GUAN's research interests include: question answering, statistical language processing, parsing, text mining.

**XiaoLong Wang** received the B.E. degree in computer science from Harbin Institute of Electrical Technology, China, and the M.E. degree in Computer Architecture from Tianjin University, China, in 1982, and 1984, respectively, and the Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology, China, in 1989. He was a senior research fellow at the polytechnic University from 1998 to 2000. Currently, he is a Professor of computer Science at Harbin Institute of Technology, China. His research interest includes artificial intelligence, machine learning, computational linguistics, and Chinese information processing.