ELSEVIER

Computers in Biology and Medicine 37 (2007) 1327–1333

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Rich features based Conditional Random Fields for biological named entities recognition

Chengjie Sun*, Yi Guan, Xiaolong Wang, Lei Lin

School of Computer Science, Harbin Institute of Technology, Mailbox 319, West Da-zhi Street 92, Harbin, Heilongjiang 150001, China

Received 3 June 2006; received in revised form 17 November 2006; accepted 4 December 2006

Abstract

Biological named entity recognition is a critical task for automatically mining knowledge from biological literature. In this paper, this task is cast as a sequential labeling problem and Conditional Random Fields model is introduced to solve it. Under the framework of Conditional Random Fields model, rich features including literal, context and semantics are involved. Among these features, shallow syntactic features are first introduced, which effectively improve the model's performance. Experiments show that our method can achieve an F-measure of 71.2% in an open evaluation data, which is better than most of state-of-the-art systems.
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Conditional Random Fields; Named entities recognition; Chunking; Sequential labeling problem; Text mining

1. Introduction

With the development of computational and biological technology, the amount of biological literature is increasing fleetingly. MEDLINE database has collected 11 million biological related records since 1965 and is increasing at the rate of 1500 abstracts a day [1]. The research literature is a major repository of knowledge. From them, researchers can find knowledge, such as connections between diseases and genes, the relationship between genes and specific biological functions and the interactions between different proteins and so on.

The explosion of literature in the biological field has provided an opportunity for natural language processing techniques to aid researchers and curators of databases in the biological field by providing text mining services. Yet typical natural language processing tasks such as named entity recognition (NER), information extraction, and word sense disambiguation are particularly challenging in the biological domain with its highly complex and idiosyncratic language.

Biological NER is a critical task for automatically mining knowledge from biological literature. Two special workshops for biological NER BioCreAtIvE [2] (Critical Assessment for Information Extraction in Biology) and JNLPBA [3] (Joint Workshop on Natural Language Processing in Biomedicine and its Applications) were held in 2004 and each of them contained an open evaluation of biological NER technology. The data and guidelines afforded by the two workshops greatly promote the biological NER technology. According to the evaluation results of JNLPBA2004, the best system can achieve an F-measure of 72.6%. This is somewhat lower than figures for similar tasks from the news wire domain. For example, extraction of organization names has been done at over 90% F-measure [2]. Therefore, biological NRE technology needs further study in order to make it applicable.

Current research methods for NER can be classified into three categories: dictionary-based methods [4], rule-based methods [5] and machine learning based methods. In biological domain, dictionary-based methods suffer from low recall due to new entities appearing continually with the advancing biology research. Biological named entities do not follow any nomenclature, which makes rule-based method hard to be perfect. Besides, rule-based method itself is hard to port to new applications. More and more machine learning methods are introduced to solve the biological NER problem, such as Hidden Markov

* Corresponding author. Tel.: +86 451 86413322 89; fax: +86 451 8641332293.
E-mail address: cjsun@insun.hit.edu.cn (C. Sun).

0010-4825/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2006.12.002

Model [6] (HMM), Support Vector Machine [7] (SVM), Maximum Entropy Markov Model [8] (MEMM) and Conditional Random Fields [1,9] (CRFs).

Biological NER problem can be cast as a sequential labeling problem. CRFs for sequences labeling offer advantages over both generative models like HMM and classifiers applied at each sequence position [10]. In this research, we utilize CRFs model involving rich features to extract biological named entities from biological literature. The feature set includes orthographical features, context features, word shape features, prefix and suffix features, Part of Speech (POS) features and shallow syntactic features. Among these features, shallow syntactic features are first introduced to CRFs model and do boundary detection and semantic labeling at the same time, which effectively improve the model's performance. Although some features have been used by some researchers, we show the effect of each kind of feature in detail, which can afford valuable reference to other researchers. Our method does not need any dictionary resources and post-processing, so it has strong adaptability. Experiments show that our method can achieve an F-measure of 71.2% in JNLPBA test data which is better than most of state-of-the-art systems.

The remainder of this paper is structured as follows. In Section 2, we define the problem of biological NER and introduce its unique characteristics compared to news wire domain. In Section 3, a brief introduction of linear-chain CRFs model is given. In Section 4 we explain the features involved in our method. Experiment results are shown in Section 5. Section 6 is a brief conclusion.

2. Biological NER

Biological NER can be addressed as a sequential labeling problem. It is defined as recognizing objects of a particular class in plain text. Depending on required application, NER can recognize objects ranging from protein/gene names to disease/virus names. In practice, we regard each word in a sentence as a token and each token is associated with a label. Each label with a form of B–C, I–C or O indicates not only the category of a named entity (NE) but also the location of the token within the NE. In this label denotation, C is the category label; B and I are location labels, standing for the beginning of an entity and inside of an entity, respectively. O indicates that a token is not part of an NE. Fig. 1 is an example of biological NER.

Biological NER is a challenging problem. There are many different aspects to deal with compared to news wire domain. In general, biological NEs do not follow any nomenclature [11] and can comprise long compound words and short abbreviations. Biological NEs are often English common nouns (as opposed to proper nouns, which, are the nouns normally associated with names) and are often descriptions [12]. For example, some Drosophila (fruit fly) gene names are *blistery*, *inflated*, *period*, *punt and midget*. Some NEs contain various symbols and other spelling variations. On average, any NE of interest has five synonyms. An NE may also belong to multiple categories intrinsically; an NE of one category may contain an NE of another category inside it [13].

Table 1
Biological named entities label list

Meaning	Label		
Beginning of protein	B-protein	Inside protein	I-protein
Beginning of DNA	B-DNA	Inside DNA	I-DNA
Beginning of RNA	B-RNA	Inside RNA	I-RNA
Beginning of cell_type	B-cell_type	Inside cell_type	I-cell_type
Beginning of cell_line	B-cell_line	Inside cell_line	I-cell_line
Others	O		

In natural language processing domain, Generative Models and Discriminative Models are often used to solve the sequential labeling problem, such as NER. Recently, Discriminative Models are preferred due to their unique characteristics and good performance [14]. Generative Models define a joint probability distribution $p(X, Y)$ where X and Y are random variables, respectively, ranging over observation sequences and their corresponding label sequences. In order to define a joint distribution of this nature, generative models must enumerate all possible observation sequences—a task which, for most domains, is intractable unless observation elements are represented as isolated units, independent from the other elements in an observation sequence. Discriminative Models directly solve the conditional probability $p(Y|X)$. The conditional nature of such models means that no effort is wasted on modeling the observations and one is free from having to make unwarranted independent assumptions about these sequences; arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related.

This paper utilizes a Discriminative Model—CRFs to solve biological NER problem. Using the definition in [2], we recognize five categories of entities. There are 11 labels in all using BIO notation mentioned above. All labels are shown in Table 1. Each token in the biological text will be assigned with one of the 11 labels in the recognition results.

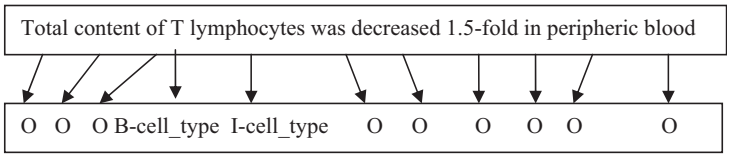

Fig. 1. An example of biological NER.

3. CRFs model

CRFs model is a kind of undirected graph model [14]. A graphical model is a family of probability distributions that factorize according to an underlying graph. The main idea is to represent a distribution over a large number of random variables by a product of local functions that each depend on only a small number of variables [15]. The power of graph model lies in that it can model multi-variables, while an ordinary classifier can only predicate one variable.

The result of NER is a label sequence, so linear-chain CRFs model is adopted in this research.

Let $\mathbf{y}$, $\mathbf{x}$ be random vectors, $\Lambda = \{\lambda_k\} \in \Re^K$ be a parameter vector, and $\{f_k(y, y', \mathbf{x}_t)\}_{k=1}^K$ be a set of real-valued feature functions. Then a linear-chain CRF is a distribution $p(\mathbf{y}|\mathbf{x})$that takes the form

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}, \quad (1)$$

in which $Z(\mathbf{x})$ is an instance-specific normalization function.

$$Z(\mathbf{x}) = \sum_{y} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1,\mathbf{x}_t})\right\}. \quad (2)$$

For the application of linear-chain CRFs model, the key problem is how to solve the parameter vector $\theta = \{\lambda_k\}$. This is done during the training process.

Suppose there are iid training data $D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N$, where each $\mathbf{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}\}$ is a sequence of inputs and each $\mathbf{y}^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \ldots, y_T^{(i)}\}$ is a sequence of corresponding predictions. Then parameter estimation is performed by penalized maximum conditional log likelihood$l(\theta)$,

$$l(\theta) = \sum_{i=1}^{N} \log\ p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}). \quad (3)$$

Putting formula (1) into formula (3), we get

$$l(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_t^i, y_{t-1}^i, \mathbf{x}_t^{(i)}) - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(i)}). \quad (4)$$

In order to avoid overfitting, a penalty term is involved, formula (4) becomes

$$l(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_t^i, y_{t-1}^i, \mathbf{x}_t^{(i)}) - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(i)}) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2}. \quad (5)$$

In formula (5), σ^2 determines the strength of the penalty. Finding the best σ^2 can require a computationally intensive parameter sweep. Fortunately, according to [15], the accuracy of the final model does not appear to be sensitive to changes in σ^2. In our experiment, σ^2 is set to 10. Given formula (5), we can use Improved Iterative Scaling (IIS) method or Numerical Optimization Techniques to find its maximum value and solve $\theta = \{\lambda_k\}$. We adopt L-BFGS [16] afforded by MALLET toolbox [17] to do that, which is a Numerical Optimization Techniques with high efficiency compared to IIS method. If $\theta = \{\lambda_k\}$ is available, we can use formula (1) to do NER.

Linear-chain CRFs model has efficient algorithms when using formula (1), such as forward backward or Viterbi. Similar to HMM, we can define the forward backward probability for linear-chain CRFs. The forward value $\alpha_i(y)$ is defined as the probability of being in state y at time i given the observation up to i. The recursive step is $\alpha_{i+1}(y) = \sum_{y'} \alpha_i(y') \exp(\sum_k \lambda_k f_k(y', y, \mathbf{x}, i+1))$.

Similarly, $\beta_i(y)$ is the probability of starting from state y at time i given the observation sequence after time i. The recursive step is $\beta_i(y') = \sum_y \exp(\sum_k \lambda_k f_k(y', y, \mathbf{x}, i+1))\beta_{i+1}(y)$.

The forward–backward and Viterbi algorithms can be derived accordingly [15].

For biological NER problem, the input sequence $\mathbf{x}$ is a sentence, the output sequences $\mathbf{y}$ are corresponding labels. The function set $\{f_k(y, y', \mathbf{x}_t)\}_{k=1}^K$ contains binary-value functions, which embody the features of the training data. For example $f_k(y, y', \mathbf{x}_t)$ may be defined as

$$f_k(y, y', \mathbf{x}_t) = \left\{\begin{array}{ll} 1 & \text{if WORD}_t = T, \text{WORD}_{t+1} = \text{cells}, \\ & y' = O, y = B\text{-cell\_type} \\ 0 & \text{others} \end{array}\right\}.$$

4. Features

In order to describe the complexity language phenomena in biological literatures, we involve orthographical features, context features, word shape features, prefix and suffix features, Part of Speech (POS) features and shallow syntactic features. Compared to other existing biological NER systems using CRFs, we first introduce shallow syntactic features in CRFs model. Shallow syntactic features are embodied using chunk labels (therefore, chunking features and shallow syntactic features have the same meaning in this paper). One of the most remarkable advantages of CRFs model is that it is convenient to involve rich features without considering the dependency of features. Also, when new features are added, the model does not need modification.

4.1. Shallow syntactic features

In order to get shallow syntactic features, we use GENIA Tagger [18] to do text chunking. Text chunking is the technique of recognizing relatively simple syntactic structures. It consists of dividing a text into phrases in such a way that syntactically related words become members of the same phrase. These phrases are non-overlapping which means that one word can only be a member of one chunk [19]. After chunking, each token will be assigned a chunk label.

Table 2
Orthographical features

Feature name	Regular expression
ALLCAPS	[A–Z]+
INITCAP	^[A–Z].*
CAPSMIX	.*[A–Z][a–z].*|.*[a–z][A–Z].*
SINGLE CHAR	[A–Za–z]
HAS DIGIT	.*[0–9].*
SINGLE DIGIT	[0–9]
DOUBLE DIGIT	[0–9][0–9]
NATURAL NUMBER	[0–9]+
REAL NUMBER	$[-0-9]+[.,]+[0-9.,]+$
HAS DASH	.*–.*
INIT DASH	–.*
END DASH	.*–
ALPHA NUMERIC	(.*[A–Za–z].*[0–9].*)|(.*[0–9].*[A–Za–z].*)
ROMAN	[IVXDLCM]+
PUNCTUATION	$[,.;:?!-+]$

The syntactic information contained in chunk labels can afford much more reliable clues for NER than literal information. For example, a noun chunk is more likely to form an entity. In our research, shallow syntactic features include chunk labels with a window of size 5. We use "c" which denotes a chunk label, −n denotes n position prior to target token, +n denotes n position after target token. The chunk features can be denoted as $c-2$, $c-1$, c0, c1, c2. Besides, some combined features are used in order to make full use of syntactic features. We employ three kinds of combined features: p-1c0, c0t0 and p0c0, where p denotes a POS tag and t denotes a token.

4.2. Other features

Orthographical features: Orthographical features describe how a token is structured. For example, whether it contains both upper and lower letters, whether it contains digits and whether it contains special character. Orthographical features are important to biological NER for its special structures. We use regular expressions to characterize orthographical features which are listed in Table 2. Some of them are also used in [1,9].

Word shape features: Tokens with similar word shape may belong to the same category [13]. We come up with a simple way to normalize all similar tokens. According to our method, upper-case characters are all substituted by "X", lower-case characters are all substituted by "x", digits are all substituted by "0" and other characters are substituted by "_". For example, "IL-3", "IL-4" and "IL-5" will be normalized as "XX_d". Thus, these tokens can share the weight of feature "XX_d". To further normalize these tokens, we substitute all consecutive strings of identical characters with one character. For example, "XX_d" is normalized to "X_d".

Prefix and suffix features: Some prefixes and suffixes can provide good clues for NER. For example, tokens ending in "ase" are usually proteins; tokens ending in "RNA" are usually RNAs. In our work, the length range of affix is 3–5. If the length is too short, the distinguishing ability of the affix will decrease. The frequency of the affix will be low if the length of affix is too long.

Context feature: Tokens near the target token may be indicators of its category. For example, "IL-3" may belong to "DNA" or "protein". If we know the next token is "gene", we can decide that it belongs to "DNA" category. According to [1,9], we choose five as the context window size, i.e. the target token, the two tokens right prior to target token and the two tokens right after target token.

POS features: The granule of POS features is larger than context features, which will help to increase the generalization of the model. GENIA Tagger is used to do POS tagging. GENIA Tagger is trained on biological literatures, whose accuracy is 98.20% as described in [18]. For POS features, we use the same window size as context features.

5. Experiment

In the experiments, JNLPBA data set[1] is adopted. The motivations of our experiments lie in two folders: (1) to indicate the effect of our method; (2) to show the function of each kind of feature in detail.

5.1. Experiment data sets

The basic statistics information of the data set is summarized in Tables 3 and 4. JNLPBA training set consists of 2000 MEDLINE abstracts retrieved using the search terms "human", "blood cell" and "transcription factor". These abstracts are then annotated manually based on the GENIA ontology. Each biological NE is annotated into one of five NE classes in Table 1 according to its chemical structure, which is usually independent of the biological context in which it appears.

For JNLPBA testing set, a new collection of MEDLINE abstracts are annotated. Four hundred and four abstracts are used that are annotated for the same classes of entities as in the training set: half of them were from the same domain as the training data and the other half of them were from the super-domain of "blood cells" and "transcription factors". That would provide an important test of generalization in the methods used [2].

Table 3
Composition of JNLPBA data set

Data set	# abs	# sen	# tokens
Training set	2000	18,546	472,006
Test set	404	3856	96,780

Table 4
Entity distribution in JNLPBA data set

Data set	# protein	# DNA	# RNA	# cell_type	# cell_line	All
Training set	30,269	9533	951	6718	3830	51,031
Test set	5067	1056	118	1921	500	8662

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html

5.2. Experiment results

We use JNLPBA training set to train our model. Evaluation is done at JNLPBA test set. Training our model with all feature sets in Section 4 took approximately 45 h (3.0 G CPU, 1.0 G Memory, 400 iterations). Once trained, the model can annotate the test data in less than a minute. The experiment results are shown in Table 5. In Table 5, P, denoting the precision, is the number of NEs a system correctly detected divided by the total number of NEs identified by the system. R, denoting the recall, is the number of NEs a system correctly detected divided by the total number of NEs contained in the input text. $F = 2PR/(P + R)$ stands for the synthetic performance of a system.

Our system achieves an F-measure of 71.20%, which is better than most of the state-of-the-art systems. Especially for protein, the most important entity category, our system's F-measure is 73.27%, which is much closer to the best system with F-measure 73.77% of protein in JNLPBA 2004.

5.2.1. The effect of each features set

Table 6 shows our system's performance with different feature sets. The baseline feature set includes orthographical features, context features, word shape features and affixes features. These features are literal features and easy to be collected. So they are often adopted by most biological NER systems, such as [1,9,13]. POS features contain larger granule knowledge than literal features. They can increase the model's generalization, so the F-measure increases to 70.33% from 69.52% when adding them into the model. Chunk features contain syntactic information which is more general linguistic knowledge than POS features, so we can see that involving shallow syntactic features can increase the performance from 70.33% to 71.20%. From Table 6, we can conclude that features containing large granule linguistic knowledge can prompt the CRFs model's generalization and get better results.

Following example shows the effect of each kind of feature vividly.

Example 1. *Input sentence*: Coordinate regulation of *HLA class II gene* expression during development and... . The correct label sequence of the italic part is "B-DNA I-DNA I-DNA I-DNA". The 2-best label sequences of the input sentence are outputted by our system ("2-best label sequences" means two label sequences with the top two highest probabilities). The 2-best label sequences corresponding to the italic part of the input sentence when using different feature sets in Table 6 are shown in Fig. 2. When baseline feature set is used, the correct result is not in the 2-best candidates. After adding POS features to the baseline feature set, the correct result is the second candidate in 2-best list. When chunk features are involved, the first best result of our system is the correct label sequence.

5.2.2. Comparison with other works

In order to compare our work with others, Table 7 lists the performance of other systems adopting CRFs model and the state-of-the-art system. All results are tested in the same data set, so they are comparable.

System 3 only involves orthographical features, context features, word shape features and prefix and suffix features. Its performance is near our baseline system. System 2 adds POS features and lexical features into system 1. Besides, system 2 adopts two post-processing methods including nested NE resolution and reclassification based on the rightmost word. But the F-measure of system 2 is still lower than our system with 1%. This also shows that syntactic features are effective in prompting the model's performance. System 4 is the state-of-the-art

Table 5
Experiment results

Entity category	P(%)	R(%)	F(%)
Protein	69.03	78.05	73.27
DNA	70.98	66.48	68.66
RNA	68.91	69.49	69.20
Cell_line	52.21	56.60	54.32
Cell_type	80.23	64.45	71.48
Overall	70.16	72.27	71.20

Table 6
The effect of different features set

Feature set	P(%)	R(%)	F(%)
Baseline	69.01	70.03	69.52
+POS features	69.17	71.53	70.33
+Chunk features	70.16	72.27	71.20

Table 7
Performance comparison

Number	System name	P(%)	R(%)	F(%)
1	Our system	70.2	72.3	71.2
2	Tzong-han Tsai (CRF) [1]	69.1	71.3	70.2
3	Settles et al. (2004) (CRF) [9]	69.3	70.3	69.8
4	Zhao [6]	69.4	76.0	72.6

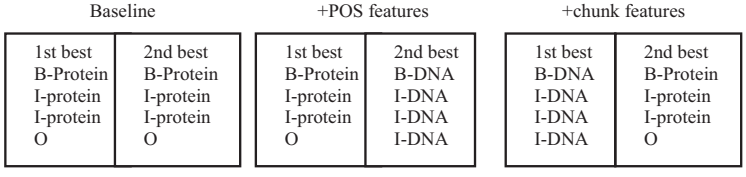

Fig. 2. Two-best results of each feature set.

system in JNLPBA2004. But according to [6], system 4 also needs lexical resource and post-processing. The F-measure of system 4 will be below 70% if post-processing is removed. Our system needs no lexical resource and post-processing. It achieves good performance with good adaptability.

6. Conclusion

Conditional Random Fields for sequences labeling offer advantages over both generative models like HMM and classifiers applied at each sequence position. In this paper, we cast biological NER as a sequential labeling problem and utilize Conditional Random Fields model involving rich features to solve it.

The main contributions of this research are:

- Firstly introduce shallow syntactic features to CRFs model and do boundary detection and semantic labeling at the same time. Experiment shows that shallow syntactic features greatly improve the model's performance.
- Show the effect of POS features and shallow syntactic features in detail; conclude that large granule linguistic knowledge can prompt the CRFs model's generalization, which can afford valuable reference to other researchers.
- Achieve a biological NER system with an F-measure of 71.2% in JNLPBA test data and which is better than most of state-of-the-art systems. The system has strong adaptability because it does not need any dictionary resources and post-processing.

7. Summary

Biological research literature is a major repository of knowledge. Unfortunately, the amount of literature has become so large that it is hard to find the information of interest on a particular topic quickly. Thus automatic literature mining is an urgent demand of biological researchers. Biological named entity recognition is a critical task for automatically mining knowledge from biological literature.

In this research, biological named entities recognition is cast as a sequential labeling problem, which means each token in a sentence will be assigned a label. The label will indicate whether the associated token is a part of an entity and the entity category if it is. Conditional Random Fields model is introduced to address this sequential labeling problem due to sound theoretical basis and outstanding practical performance in similar tasks compared to other stochastic models.

Under the framework of Conditional Random Fields model, rich features including literal, context and semantics are involved. In these features, shallow syntactic features are first introduced into Conditional Random Fields model and do boundary detection and semantic labeling at the same time, which effectively improve the model's performance. Experiments show that our method can achieve an F-measure of 71.2% in an open evaluation data, which is better than most of state-of-the-art systems.

The effect of each kind of feature for biological named entities recognition is also shown in detail through experiments. One can conclude that features containing large granule linguistic knowledge can prompt the CRFs model's generalization and get better recognition results according to the experiment results, which can afford valuable reference to other researchers. A comparison with others works shows that our method achieves the best result among the known systems adopting CRFs model. Also, our method needs no lexical resource and post-processing. It achieves good performance with good adaptability.

Acknowledgments

This work is partly supported by National Natural Science Foundation of China (60504021) and the 863 high Technology Research and Development Programme of China (2002AA117010-09).

References

[1] T.H. Tsai, W.C. Chou, S.H. Wu, et al., Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities, Expert Systems Appl. 30 (1) (2006) 117–128.
[2] L. Hirschman, A. Yeh, C. Blaschke, A. Valencia, Overview of BioCreAtIvE: critical assessment of information extraction for biology, BMC Bioinformatics 6 (Suppl. 1) (2005).
[3] J.D. Kim, T. Ohta, Y. tsuruoka, Y. Tateisi, Introduction to the bio-entity recognition task at JNLPBA, in: Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004, pp. 70–75.
[4] Z. Kou, W.W. Cohen, R.F. Murphy, High-recall protein entity recognition using a dictionary, Bioinformatics 21 (Suppl. 1) (2005) i266–i273.
[5] A.M. Cohen, W.R. Hersh, A survey of current work in biomedical text mining, Briefings Bioinformatics 6 (1) (2005) 57–71.
[6] G. Zhou, J. Su, Exploring deep knowledge resources in biomedical name recognition, in: Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004, pp. 96–99.
[7] J. Kazama, T. Makino, Y. Ohta, J. Tsujii, Tuning support vector machines for biomedical named entity recognition, in: Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain, 2002, pp. 1–8.
[8] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, G. Sinclair, Exploiting context for biomedical entity recognition: from syntax to the web, in: Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004, pp. 88–91.
[9] S. Burr, Biomedical named entity recognition using conditional random fields and novel feature sets, in: Joint Workshop on Natural Language Processing in Biomedicine and its Application, 2004, pp. 104–107.
[10] F. Sha, F. Pereira, Shallow parsing with conditional random fields, in: Proceedings of HLT-NAACL, 2003, pp. 213–220.
[11] H. Shatkay, R. Feldman, Mining the biomedical literature in the genomic era: an overview, J. Comput. Biol. 10 (6) (2003) 821–855.
[12] A.S. Yeh, A. Morgan, M. Colosimo, L. Hirschman, BioCreAtIvE task 1A: gene mention finding evaluation, BMC Bioinformatics 6 (Suppl. 1) (2005).
[13] T.H. Tsai, C.W. Wu, W.L. Hsu, Using maximum entropy to extract biomedical named entities without dictionaries, in: Proceedings of IJCNLP2005, 2005, pp. 270–275.
[14] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the International Conference on Machine Learning, 2001, pp. 282–289.
[15] C. Sutton, A. McCallum, An Introduction to Conditional Random Fields for relational Learning, ⟨http://www.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf⟩, 2005.

[16] H.M. Wallach, Efficient training of conditional random fields, Master's Thesis, University of Edinburgh, 2002.
[17] A. McCallum, MALLET: A Machine Learning for Language Toolkit, ⟨http://mallet.cs.umass.edu⟩, 2002.
[18] Y. Tsuruoka, Y. Tateishi, J.D. Kim, et al., Developing a robust part-of-speech tagger for biomedical text, in: Advances in Informatics—10th Panhellenic Conference on Informatics, 2005, pp. 382–392.
[19] F. Erik, T.K. Sang, S. Buchholz, Introduction to the CoNLL-2000 shared task: chunking, in: Proceedings of CoNLL-2000 and LLL-2000, 2000, pp. 127–132.

Chengjie Sun is a Ph. D. student of School of Computer Science, Harbin Institute of Technology, Harbin, China. He received his B.E. degree in computer science from Yantai University in 2002, and also completed M.E. in computer science in Harbin Institute of Technology in 2004. His research interests include machine learning, information extraction and text mining.

Dr. Xiaolong Wang is a professor of Computer Science in Harbin Institute of Technology currently. He received the B.E. degree in computer science from Harbin Institute of Electrical Technology, China, 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 joined Harbin Institute of Technology, China, as an assistant lecture in 1984 and was an associate professor in 1990. He was a senior research fellow at the polytechnic University from 1998 to 2000. His research interest includes artificial intelligence, machine learning, computational linguistics, bioinformatics and Chinese information processing.

Dr. Yi Guan is presently a professor of the School of Computer Science and Technology at Harbin Institute of Technology. He 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 a research assistant in Human Language Technology Center at Hong Kong University of Science and Technology, and he was a research scientist in Weniwen.com limited (Hong Kong) in 2001. In October 2001, he became an associate professor in School of Computer Science and Technology in Harbin Institute of Technology. His research interests include question answering, statistical language processing, parsing and text mining.

Dr. Lei Lin is presently an associate at the School of Computer Science and Technology at Harbin Institute of Technology. He received the Ph.D. degree in computer science and technology from Harbin Institute of Technology in 2004. His research interests include bioinformatics, pattern recognition, information fusion and information processing.