Biomedical Named Entities Recognition Using Conditional Random Fields Model*

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

School of Computer Science, Harbin Institute of Technology,
Heilongjiang Province, China
{cjsun, guanyi, wangxl, linl}@insun.hit.edu.cn
http://www.insun.hit.edu.cn

Abstract. Biomedical named entity recognition is a critical task for automatically mining knowledge from biomedical literature. In this paper, we introduce Conditional Random Fields model to recognize biomedical named entities from biomedical literature. Rich features including literal, context and semantics are involved in Conditional Random Fields model. Shallow syntactic features are first introduced to 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 JNLPBA test data and which is better than most of state-of-the-art system.

1 Introduction

With the development of computational and biological technology, the amount of biomedical literature is increasing unprecedentedly. MEDLINE database has colleted 11 million biomedical 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 the knowledge, such as connections between diseases and genes, the relationship between genes and specific biological functions and the interactions of different proteins and so on.

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

Biomedical Named Entities Recognition (NER) is a critical task for automatically mining knowledge from biomedical literature. Two special workshops for biomedical named entities recognition BioCreAtIvE [2] (Critical Assessment for Information Extraction in Biology) and JNLPBA [3] (Joint Workshop on Natural Language

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

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1279 – 1288, 2006.

Processing in Biomedicine and its Applications) were held in 2004 respectively and each of them contained an open evaluation of biomedical named entities recognition technology. The data and guidelines afforded by the two workshops greatly promote the biomedical NER technology. According to the evolution 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 0.90 F-measure [2]. Therefore, biomedical NER technology need further study in order to make it applied.

Current research methods for NER can be classified into 3 categories: dictionary-based methods [4], rule-based methods [5] and machine learning based methods. In biomedical domain, dictionary-based methods suffer from low recall due to new entities appear continually with the biology research advancing. Biomedical NEs do not follow any nomenclature, which makes rule-based methods to be helpless. Besides, rule-based method itself is hard to port to new applications. More and more machine learning methods are introduced to solve the biomedical NER problem, such as Hidden Markov Model [6] (HMM), Support Vector Machine [7] (SVM), Maximum Entropy Markov Model [8] (MEMM) and Conditional Random Fields [1, 9] (CRFs). Biomedical NER problem can be cast as a sequential labeling problem. Conditional random fields for sequences labeling offer advantages over both generative models like HMM and classifiers applied at each sequence position[10].

In this research, we utilize Conditional Random Fields model involving rich features to extract biomedical named entities from biomedical 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. Shallow syntactic features are first introduced to Conditional Random Fields 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 features 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 and which is better than most of state-of-the-art system.

The remainder of this paper is structured as follows. In section 2, we define the problem of biomedical named entities recognition and its unique characteristics. In section 3, a brief introduction of linear-chain conditional random fields model are given. In section 4 we explain the features involved in our system. Experiment results are shown in section 5. Section 6 is a brief conclusion.

2 Biomedical Named Entity Recognition

Biomedical 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 extract 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 the Named Entity (NE) but also the location of the token within the an NE. In this label denotation, C is the category label; B and I are location labels, standing

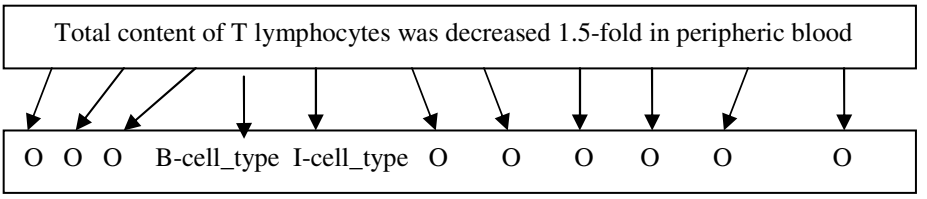

Fig. 1. An example of biomedical NER

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 biomedical NER.

Biomedical NER is a challenging problem. There are many different aspects to deal with. In general, biomedical NEs do not follow any nomenclature [11] and can comprise long compound words and short abbreviations. Biomedical 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].

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 characteristic 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 $p(Y \mid X)$ .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 independence 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.

Table 1. Biomedical Named Entities label list

Meaning	Label	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		

This paper utilizes a Discriminative Model – Conditional Random Fields to solve biomedical NER problem. Using the definition in [3], we recognize 5 categories entities. There are 11 labels in all using BIO notation mentioned before. All labels are show in Table 1. Each token in the biomedical text will be assigned one of the 11 labels in the recognition results.

3 Conditional Random Fields Model

Conditional Random Fields (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 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 CRFs is a distribution $p(\mathbf{y}|\mathbf{x})$ that takes the form

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

where $Z(\mathrm{x})$ is an instance-specific normalization function.

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

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 $\mathbf{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \cdots, x_T^{(i)}\}$ is a sequence of inputs and each $\mathbf{y}^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_T^{(i)}\}$ is a sequence of corresponding predictions. Then parameter estimation is performed by penalized maximum conditional log likelihood $l(\theta)$,

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

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

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

In order to avoiding overfitting, a penalty term is involved, the formula (4) becomes into

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

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, the 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 an Numerical Optimization Techniques with high efficiency compared to IIS method. If $\theta = \{\lambda_k\}$ is available, we can use formula (1) to do NER.

For biomedical NER problem, the input sequence $\mathbf{x}$ is a sentence, the output sequences $\mathbf{y}$ is 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 $f_k(y, y', \mathbf{x}_t) = \begin{cases} 1 \ \textit{if } WORD_t{=}T, WORD_{t+1}{=}\textit{cells}, y'{=}O, y{=}\textit{B-cell\_type} \\ 0 \ \textit{others} \end{cases}$.

4 Features

In order to describe the complexity language phenomena in biomedical literatures, we involve orthographical features, context features, word shape features, prefix and suffix features, Part of Speech (POS) features and shallow syntactic features. Compare to others exist biomedical NER system using CRFs, we first introduce shallow syntactic features in CRFs model. Shallow syntactic features are embodied using chunk labels (Therefore, chunk features and shallow syntactic features are 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 doesn't 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 techniques of recognizing relatively simple syntactic structures. It consists of dividing a text into phrases in such a way that syntactically related words become member 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.

The syntactic information contains 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. If we use "c" denote a chunk label, -n denote n position prior to target token, +n denote 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 3 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 biomedical 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].

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	[,.;:?!-+]

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, there 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 suffix 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 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 belong to "DNA" category. According to [1, 9], we choose 5 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 increasing the generalization of the model. GENIA Tagger is used to do POS tagging. GENIA Tagger is a trained on biology literatures, whose accuracy is 98.20% as described in [18]. For POS features, we use the same window size as context features.

5 Experiment

5.1 Experiment Dataset

In the experiment, JNLPBA 2004 dataset is adopted. Its basic statistics is summarized in Table 3 and Table 4. Only 106 abstracts' publish year among 404 in test dataset are same as training dataset [3]. The difference in publish year between training data and test data demands the model should have a good generalization.

Table 3. Dataset of JNLPBA

dataset	#abs	#sen	#tokens
Training set	2,000	18,546	472,006
Test set	404	3,856	96,780

Table 4. Entity distribution in JNLPBA dataset

dataset	protein	DNA	RNA	cell_type	cell_line	All
Training set	30,269	9,533	951	6,718	3,830	51,031
Test set	5,067	1,056	118	1,921	500	8,662

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 hours (3.0G CPU, 1.0G 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.

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

Our system achieves 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 JNLPBA2004.

Table 6 shows our system's performance with different feature sets. The baseline feature set includes orthographical features, context features, word shape features and prefix and suffix features. These features are literal features and easy to collection. So they are often adopted by most biomedical NER system, such as [1, 9, 13]. POS features contain larger granule knowledge than literal feature. 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. Involving shallow syntactic features can increase the performance from 70.33% to 71.20%. From Table 6, we can conclude that features contain large granule linguistic knowledge can prompt the CRFs model's generalization and get better results.

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

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 dataset, so they are comparable.

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, 2004[6]	69.4	76.0	72.6

System 3 only involves orthographical features, context features, word shape features and prefix and suffix features. Its performance is near to 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 percent. This also shows that syntactic features are effective in prompting the model's performance. System 4 is the state-of-the-art system in JNLPBA2004. But according to [6] system 4 also need lexical resource and post processing. The F-measure of system 4 will below 70% if post processing is removed. Our system need not any 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 biomedical 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:

- First 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 biomedical NRE system with an F-measure of 71.2% in JNLPBA test data and which is better than most of state-of-the-art system. The system has strong adaptability because it does not need any dictionary resources and post-processing.

References

1. Tsai, T.H., Chou, W.C., Wu, S.H., Sung, T.Y., Hsiang, J., Hsu, W.L.: Integrating Linguistic Knowledge into a Conditional Random Field Framework to Identify Biomedical Named Entities. Expert Systems with Applications. 30(1) (2006) 117-128.
2. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics. 6(Suppl 1) (2005)
3. KIM, J.D., OHTA, T., TSURUOKA, Y., TATEISI, Y.: Introduction to the Bio-Entity Recognition Task at JNLPBA. In Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004) 70-75
4. Kou, Z., Cohen, W.W., Murphy, R.F.: High-recall protein entity recognition using a dictionary. bioinformatics. 21(Suppl. 1) (2005) i266-i273
5. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. BRIEFINGS IN BIOINFORMATICS. 6(1) (2005) 57-71

6. Zhou, G.D., Su, J.: Exploring Deep Knowledge Resources in Biomedical Name Recognition. In Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004) 96-99
7. Kazama, J., Makino, T., Ohta, Y., Tsujii, J.: Tuning Support Vector Machines for Biomedical Named Entity Recognition. In Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain. (2002) 1-8
8. Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., Sinclair, G.: Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. In Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004) 88-91
9. Burr, S.: 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) 104-107
10. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In Proceedings of HLT-NAACL. (2003) 213-220
11. Shatkay, H., Feldman, R.: Mining the Biomedical Literature in the Genomic Era: An Overview. JOURNAL OF COMPUTATIONAL BIOLOGY. 10(6) (2003) 821-855
12. Yeh, A.S., Morgan, A., Colosimo, M., Hirschman L.: BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics 6(Suppl 1) (2005)
13. Tsai, T.H., Wu, C.W., Hsu, W.L.: Using Maximum Entropy to Extract Biomedical Named Entities without Dictionaries. In Proceedings of IJCNLP2005. (2005) 270-275.
14. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the International Conference on Machine Learning. (2001) 282–289
15. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. http://www.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf. (2005)
16. Wallach, H.M.: Efficient training of conditional random fields. Master's thesis. University of Edinburgh (2002)
17. McCallum, A.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. (2002)
18. Tsuruoka, Y., Tateishi, Y., Kim, J.D.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In Advances in Informatics - 10th Panhellenic Conference on Informatics. (2005) 382-392
19. Erik, F., Sang, T.K., Buchholz, S.: Introduction to the CoNLL-2000 Shared Task: Chunking. In Proceedings of CoNLL-2000 and LLL-2000. (2000) 127-132