J. Intelligent Learning Systems & Applications, **2010, 2, 139-146**
doi:10.4236/jilsa.2010.23017 Published Online August 2010 (http://www.SciRP.org/journal/jilsa)

Chunk Parsing and Entity Relation Extracting to Chinese Text by Using Conditional Random Fields Model

Junhua Wu, Longxia Liu

College of Electronics and Information Engineering, Nanjing University of Technology, Nanjing, China.
Email: wujh@njut.edu.cn

Received March 19th, 2010; revised July 20th, 2010; accepted July 30th, 2010.

ABSTRACT

Currently, large amounts of information exist in Web sites and various digital media. Most of them are in natural language. They are easy to be browsed, but difficult to be understood by computer. Chunk parsing and entity relation extracting is important work to understanding information semantic in natural language processing. Chunk analysis is a shallow parsing method, and entity relation extraction is used in establishing relationship between entities. Because full syntax parsing is complexity in Chinese text understanding, many researchers is more interesting in chunk analysis and relation extraction. Conditional random fields (*CRFs*) *model is the valid probabilistic model to segment and label sequence data. This paper models chunk and entity relation problems in Chinese text. By transforming them into label solution we can use CRFs to realize the chunk analysis and entities relation extraction.*

***Keywords*:** *Information Extraction*, *Chunk Parsing*, *Entity Relation Extraction*

1. Introduction

At present, information is presented in various digital media. Many of them are organized in natural language, such as information in Web pages, text document in digital library etc. They are non structural or semi-structural and difficult to understand by computer. Further processing to the information is blocked. It makes large amounts of information wasted. So research on semantic Web, natural language understanding is developed in order to structure and retrieve information from Web pages or other natural language documents. And information extraction is important task in the work.

Information extraction is a process to retrieve information from large text set. It may be concerned with identifying named entity, extracting relationship and label properties of sentence etc. It is a subfield of natural language understanding. There are some methods for information extraction including methods based on rules [1,2] and statistical model [3-6].

Chunk analysis and relation extraction play the important roles in information extraction. It is a simplified syntax paring technology to define and label chunk based on syntax and semantics [7]. Comparing with full parsing this method only identifies the partial structure in a sentence, such as noun phrase or verb phrase. Through which, the simple syntax parsing can be implemented and information extraction may be more effective and simple.

The objective of entity relation extraction is identifying the relationship between entities in text. Miller et al. considered the problem of relation extraction in the context of natural language parsing and augmented syntactic parses with semantic relation-specific attributes [8]. It will be critical in events detecting and describing for research on information extraction. Entity relation may be explicit and implicit. Some encountered problems make studying on entities relation hard such as few dataset, difficult extraction of implicit relation and immature parsing to Chinese.

Conditional random fields model is a valid probabilistic model to segment and label sequence data [9]. In Chinese understanding, some research use CRFs in Chinese part-of-speech and word segmentation [10,11], but seldom in chunk parsing and entity relation extraction.

Compared with other statistical model CRFs can represent long-range dependences and multiple interacting features. Our innovation is that we analyze Chinese characteristics and then model chunk and entity relation problems as label problem. Moreover using CRFs real-

izes the chunk analysis and relation extraction.

2. Related Work

A number of approaches currently have been used for natural language tasks as part of speech tagging and entity extraction. They are usually based on rules or statistic models.

Text chunk divides a text in syntactically correlated parts of words. Steven introduced chunks [12] firstly. Many machine learning approaches, such as Memory-based Learning (MBL) [13], Transformation-based Learning (TBL) [14], and Hidden Markov Models (HMMs), have been applied to text chunking [15] for parsing.

Named entity is important linguistic unit. So there are many works such as named entity recognition, disambiguation, and relationship extraction on it [16-20]. The problem of relation extraction is starting to be addressed within the natural language processing and machine learning communities. Since it is proposed, many methods have been suggested. Methods based on knowledge base were used in decision of relation extraction firstly. But it is difficult to construct knowledge base. Therefore some methods based on machine learning were emerged, such as feature-based [16], kernel-based [17] method. Approach kernel-based is a valid one for relation extracting, but its training and testing time is long for large amounts of data.

2.1 Model for Information Extraction

A lot of research to information extraction is based on machine learning methods using statistic model because by the model sentence can be segmented and labeled. Statistical language model is a probability model which estimate probability of expected text sequence by computing probability. These models are concerned with Hidden Markov Model (HMM), Maximum Entropy Model (ME), Maximum Entropy Markov Model (MEMM) and conditional random fields model CRFs. Our method is also based on statistic model.

Hidden Markov models (HMMs) are a powerful probabilistic tool for modeling sequential data, and have been applied with success to many text-related tasks, such as part-of-speech tagging, text segmentation and information extraction [6]. HMM can be considered as a finite state machine that presents states and transition chains of an application. The model is built either by manual or training. Usually extracting text information is concerned with training and labeling. Maximum likelihood and Baum-Welch algorithm are used to learning sample data labeled or unlabeled. And then Viterbi algorithm is used to label state sequence with maximum probability in text needed processing.

HMM is easy to build. It needn't large dictionary or rule sets with well flexibilities. There are many improved HMM model and their application in information extraction. Freitag and McCallum's paper [3] uses stochastic optimization to search the fittest HMM. Souyma Ray and Mark Crave [4] choose HMM to represent sentence structure. Scheffer T, Decomain C and Wrobel S [5] proposes a method which uses active learning to minimize the label data for HMM training. But HMM is a generative model and independent hypothesis is needed, so it will ignore the context of information and lead to an unexpected result.

Maximum Entropy (ME) method [21] converts the sequence label into data classifying. Its principle can be stated as follows [22]:

1) Reformulate the different information sources as constraints to be satisfied by the target (combined) estimate.

2) Among all probability distributions that satisfy these constraints, choose the one that has the highest entropy.

The advantage of ME is [21]: It makes the least assumptions about the distribution being modeled other than those imposed by constraints and given by the prior information. The framework is completely general in that almost any consistent piece of probabilistic information can be formulated a constraint. Moreover, if the constraints are consistent, that is there exists a probability function which satisfies them, then amongst all probability functions which satisfy the constraints, there is a unique maximum entropy.

This ME method will lost sequence properties. So a model combining ME and MM (Markov Model) is emerged, that is MEMM [6].

In MEMM, the HMM transition and observation functions are replaced by a single function $P(s|s',o)$ that provides the probability of the current state s given the previous state s' and the current observation o. In this model, as in most applications of HMMs, the observations are given—reflecting the fact that we don't actually care about their probability, only the probability of the state sequence (and hence label sequence) they induce.

Conditional probability of transition between states is introduced in MEMM, which makes the arbitrary choice of properties possible. But MEMM is partial model which needs normalization for each node. Therefore only a localized optimization value is obtained. Also the problem named length bias and label bias [9] may be caused. It means the method will ignore those not in training dataset.

2.2 Label Bias

Classical discriminative Markov models, maximum entropy taggers (Ratnaparkhi, 1996), and MEMMs, as well as non-probabilistic sequence tagging and segmentation models with independently trained next-state classifiers are all potential victims of the label bias problem [9].

Consider a MEMM model shown in **Figure 1** which is a finite-state acceptor for shallow parsing of two sentences:

The robot wheels Fred round.

The robot wheels are round.

Here [B-NP] etc. are labels for sentence. NP, VP, ADJP and PP mean Noun Phrase, Verb Phrase, Adjective Phrase and Prep Phrase. B or I stand for word location, begin or inter of a phrase.

It is obvious that sum of transition probability is 1 from a state i to other adjacent states. Because there is only one transition in state 3 and 7, while current state and observed value *Fred* are specified, conditional probability of next state is:

$$p(4 \mid 3, Fred) = p(8 \mid 7, Fred) = 1$$

But this equation will face to some problems if there isn't existing a transition from state 7 to state 8 while observe value is *Fred* in training dataset. Generally a low probability is specified if an unknown event exists in training dataset. But for state with single output, the follow equation have to be given:

$$\sum_{\substack{allstates \\ fromstate7to}} p(s \mid 7, Fred) = 1$$

It means that the observed value *Fred* is ignored. This will result in that label sequence is not related to observed sequence. That is label bias.

Proper solutions require models that account for whole state sequences at once by letting some transitions "vote" more strongly than others depending on the corresponding observations [9].

Lafferty suggests a global model CRFs that can solve the problems discussed before. Instead of local normalizing CRFs can realize global processing, so a global optimization value will be produced. CRFs is a new graph model of probability which can represent the long-range dependences and multiple interacting features. Domain knowledge is represented conveniently by the model. McCallum use this model to process named entity recognition [23]. His experiments shows F value is 84.04% while processing English, F value is 68.11% while processing German. Hong mingcai uses CRFs to label Chinese part-of speech [11]. But information extraction of Chinese is still a difficult task presented in many subfields such as chunk analysis and entity relation extraction. So this paper explores the methods about chunk analysis and entity relation extraction to Chinese text based on CRFs.

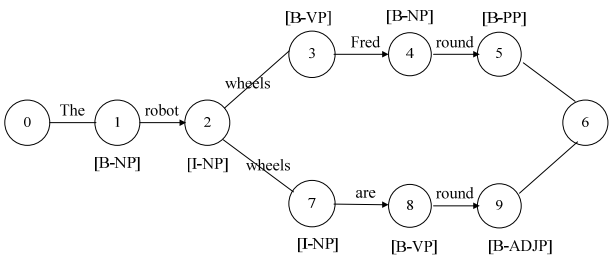

Figure 1. Finite-state acceptor for shallow parsing of two sentences

3. Conditional Random Fields (CRFs) Model

Conditional random fields model is a probabilistic model to segment and label sequence data based on statistic. It is a non-directional graph model that can compute conditional probability of output sequence when conditioned on input sequence of model.

Definition 1 [9]. Let $G = (V,E)$ be a graph such that $Y = (Y_v)$ $v \in 2V$, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

CRFs is a random field globally conditioned on the observation X. if $X = \{x_1, x_2, \ldots x_n\}$ is specified as data sequence needed label then $Y = \{y_1, y_2, \ldots y_n\}$ is the result data which have been segmented or labeled by the model. The model computes the joint distribution over the label sequence Y given X instead of only defining next state in terms of current state.

The conditional probability of label sequence Y depends on the global interactional features with different weight.

Assume $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ is a vector of features, conditional probability, for a given X, $P_\Lambda(Y|X)$ is defined as follow:

$$p \wedge (Y \mid X) = \frac{1}{Z_X} \exp\left\{\sum_{t=1}^{T}\sum_{k} \lambda_k f_k\left(y_{t-1}, y_t, X, t\right)\right\} \quad (1)$$

$$Z_X = \sum_{Y} \exp\left\{\sum_{t=1}^{T}\sum_{k} \lambda_k f_k\left(y_{t-1}, y_t, X, t\right)\right\} \quad (2)$$

Z_x is a normalized value that makes the total probability of all state sequence is 1 for given X. $f_k(y_{t-1}, y_t, X, t)$ is a feature function to mark the feature at position t and t-1 for observed X. Its value is between 0 and 1. $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ is corresponding to the context of data sequence and is a weight set of $f_k(y_{t-1}, y_t, X, t)$.

If we want to use the CRFs model to obtain expected result the critical task is training model. A model trained can produce optimization $P(Y|X)$, that is $Y^* = \arg\max_Y p(Y \mid X)$. It also means $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ will be deter-

mined. Training may use log-likehood algorithm that is independent of applications.

In this paper chunk analysis and entity relation extraction will be converted into the label solution. Data sequence X is made up of some words. For each word W_0 there are some words ahead or back of it. It is represented $W = \{W_{-n}, \ldots W_{-1}, W_0, W_{+1}, \ldots W_{+n}\}$. W_{-n}, stands for nth word previous to W_0 and W_{+n} is the nth one following W_0. $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$, in model, can be thought as feature weights related to $W = \{W_{-k}, \ldots W_{-1}, W_0, W_{+1}, \ldots W_{+k}\}$. Each λ is specified in model after training, and label sequence can be produced by Viterbi algorithm through running the model .

4. Chunk Analysis Based on CRFs

Chunk is firstly proposed by Abney [12]. He thinks chunk is the syntax element between word and sentence and with non-recursive properties.

Chunk analysis is partial parsing, also named shallow parsing, relative to complete parsing with simplified policy [15]. It is a new technology of natural language processing. Full parsing can produce a complete parse tree finally by series analysis process to sentence, which needs large cost. But chunk analysis only needs to identify some structures of the sentences such as non-recursive noun phrase, verb phrases etc, called chunk. By dividing sentence into different chunks in syntax or semantics and labeling chunks we can improve the efficiency of information extraction. It is a policy between lexical analysis and syntax analysis. Chunk partitioning and identifying are completed by chunk parsing in natural language processing.

4.1 The Definition and Label of Chunks

Definition 2. Chunk is a structure that is non-recursive phrase meet syntax. Each chunk has a head word and begins or ends at this word.

Non-recursive phrase means nested structure not exist. That is, all chunks are the same level.

Conference on Computational Natural Language Learning (CoNLL-2000) developed a dataset of English chunk which provided a platform to evaluating and test chunk analysis algorithms. There are 11 types chunks defined. They are NP, VP, ADVP, ADJP, PP, SBAR, CONJP, PRT, INTJ, LST, UCP [24].

Most of Chinese chunks present the same properties compared with English. But there is some difference. By analyzing the properties of Chinese we defined some chunk types: noun chunk(NP), verb chunk(VP), adjective chunk(AP), adverb chunk(DP), preposition chunk(SP), time chunk(TP), quantifier chunk(MP), conjunction chunk (CONJP) and other chunk(UCP).

In fact chunk analysis based on CRFs has become a process of labeling chunk like tagging part-of speech. Generally there are two kinds of standard method to label: Inside/Outside and Start/End methods. Inside/Outside policy, named IOB1, uses tag set {I,O,B} [25] to label internal, outside and first word of a chunk. Combining it with chunk type we will have chunk labeled. Such as B-VP, it shows that is a first word of a verb chunk. O means the word doesn't belong to any chunk. Start/End method, named IOBES, uses tag set {I,O,B,E,S}. When chunk only includes one word, S tag is used. E labels the last word of a chunk. Other tags are the same as Inside/Outside. For example S-NP means a chunk is constructed by one word. **Table 1** presents the label chunks of a sentence. The first column of table is Chinese words and the second column is corresponding to English for reader understanding. Next two columns are notations used IOB1 and IOBES.

In the table, row 4-6 represent a verb chunk which consist of three Chinese words labeled B-VP, I-VP and E-VP if use IOBES method. By these label chunk analysis is considered as chunk label which can be implemented by training CRFs model.

4.2 Model Training

CRFs model must be trained using labeled dataset to determine the model parameters. That trained model can be used to realize processing text which expects to be segmented and labeled. If X is sentences that have been labeled and Y is corresponding label sequence of chunk CRFs model training will make label sequence $Y^* = \arg\max_Y p(Y \mid X)$ optimal.

Here we use CRF++0.50 as training and testing tool. CRF++0.50 is a string learning tool based on CRFs principle. The training sample file and feature template file are needed in training process. Training will result in a CRfs model which will be used in labeling chunk to Chinese text.

Table 1. An example of label chunks

Chinese	English	IOB1	IOBES
因而	So	I-CONJP	S-CONJP
我们	we	B-NP	S-NP
可能	may	B-VP	B-VP
会	be	I-VP	I-VP
面临	Face to	I-VP	E-VP
一个	a	B-MP	S-MP
不	un	B-NP	B-NP
稳定	stable	I-NP	I-NP
时期	period	I-NP	E-NP
。	.	O	O

The training sample file is made up of some blocks and each block represents a sentence. The block form of training sample file is presented in **Table 2**.

There is a blank row between blocks. Each block includes some tokens and each token is a label word in one row. First column is the Chinese word and next column is the English word to help understanding. Third column lists the properties of the word (may be more than one column). Last column is the tag notations.

In section 3 we know some feature weights used in representing context of a word W_0. So it is important to select feature set. Generally context of a word and their properties are very useful for decision of feature. That means we can use some words which are previous or succeed to word W_0 as features. Features may be N-gram. Features $\{\ldots W_{-2}, W_{-1}, W_0, W_1, W_2 \ldots\}$ is named Uni-gram basic features and $\{\ldots W_{-2}, W_{-1}, W_{-1}, W_0, W_0, W_1, W_1, W_2 \ldots\}$ is named Bi-gram basic features. Here W_n stands for a word. In addition, advanced features $\{\ldots WP_{-2}, WP_{-1}, WP_0, WP_1, WP_2 \ldots\}$ which combine the word and its property together are also used to improve result of analysis. P means property of a word. **Table 3** shows various features of word "赤字" (deficit) in **Table 2**. The last column of **Table 3** is English word corresponding to Chinese word.

So an observing window of token W_0 need to be given for training. The window includes W_0 and some words before and after it, that is $W = \{W_{-n}, W_{-(n-1)}, \ldots, W_0, \ldots, W_{n-1}, W_n\}$. **Table 4** is an example about observing window. It is used as feature source for training vector $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$. Larger window provides more context feature, but it will increase the cost of processing. Too small window may lose important features. So we define windows size as 5, that is $W = \{W_{-2}, W_{-1}, W_0, W_1, W_2\}$.

The template file of features defines the feature item for training. After training $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ is produced, that is CRFs model has been available.

Table 2. Form of experiment dataset sample

Chinese token	English token	Property	Notation
他	He	PRP	B-NP
认为	reckons	VBZ	B-VP
当前的	current	JJ	I-NP
赤字	deficit	NN	I-NP
将	will	MD	B-VP
缩小	narrow	VB	I-VP
到	to	TO	B-PP
仅	only	RB	B-NP
1800	18000	CD	I-NP
万	thousands	CD	I-NP
9 月	september	NNP	B-NP
.		.	O

Table 3. Feature instance

Feature	Feature item	Feature value	Value in English
Uni-gram basic features	W_{-2}	认为	reckons
	W_{-1}	当前的	current
	W_0	赤字	deficit
	W_1	将	will
	W_2	缩小	narrow
Bi-gram basic features	$W_{-1}W_0$	当前的/赤字	current/ deficit
	W_0W_1	赤字/将	deficit/will
Uni-gram advanced features	WP_{-1},	当前的/JJ	current/JJ
	WP_1	将/MD	Will/MD

Table 4. Observing window of features

Feature position	Description	Chinese example
$W = W_0$	Token	当前的
$W = W_{-1}$	Last word of token	认为
$W = W_{+1}$	Next word of token	赤字
$W = W_0W_{+1}$	Token and next word	当前的 赤字
$W = W_{-1}W_0W_{+1}$	Last word, token and next word	认为 当前的 赤字

5. Entity Relation Extraction

Entity is the basic element in natural text, such as place, role, organization, thing etc. Entities play important roles in natural language text. Generally there are some relationships between them. Such as locating, belong to, adjacent and so on. These relationships may be explicit or implicit. Implicit relationship needs reasoning by knowledge. Entity relation extraction is the process of identifying the relationship between entities in text and labeling them. It is not only an important work in information extraction but also useful in automatic answer or semantic network.

Testing from MUC shows that many systems are able to process named entity to large of English document [26]. But entity relation extraction to Chinese may be difficult. As we known machine learning is the valid method for extracting, but it needs dataset labeled. Currently, "People's Daily" labeled by Beijing university is perhaps a better choice. This dataset has been labeled in

part-of speech, properties of word, named entity. But extracting implicit relationship needs more external knowledge.

5.1 The Definition and Label of Entity Relation

ACE 2006 defines six classes and 18 subclasses relationships between two entities. They are shown in **Table 5**. The dataset provided by ACE covers English, Chinese and Arabie.

This paper focuses on Chinese entity relation. Here we only illustrate two kinds of relationship of physical class, and only implement extraction based on the definition of these two relationships.

Definition 3: Relation 1 (M, N) is defined as located relationship. With Entity M, N $\in$ geographical entity and N $\in$ M.

Definition 4: Relation 2(M, N) is defined as near relationship. With M, N $\in$ geographical entity and 2(M, N) = 2 (N, M).

In this paper the same principle as chunk analysis is used to realize extraction which training CRFs model through label sample dataset. Then using CRfs model realizes the extraction. Here we suggest nine kinds of notation to label the position relationship in dataset. **Table 6** presents these nine notations.

In the table each row presents one or two entities and their relationship. For a notation 1-B, 1 stands for Relation 1 (M, N) and B stands for the first entity M. The third column lists the sentence including entities. The end column shows the instance corresponding to notation.

5.2 Experiment of Entity Relation Extraction

Experiment is designed based on the definition of Relation 1 (M,N) and Relation 2 (M,N). The paper uses label dataset of "people's daily" on January 1998 as sample dataset named M, size 3.2 MB. It is divided into 10 subsets from M1 to M10. M1 includes text on Jan 1st and M2 ranges from 1st to 2nd. By the way M10 ranges from 1st to 10th. That is Mn means newspaper contents of *n* days.

M1-M9 is considered as training dataset and M10 as test set. **Table 7** is the dataset and result of experiment.

In the table Recall R, Precision P and F-Score are metrics to information extraction. Let r_1 be numbers of relationships extracted correctly, r_2 as numbers of relationships extracted actually, r_3 as numbers of original relationships in text. Then:

Table 5. Named entity relation types and instances

Class	Subclass
Physical	Located, Near
Part-Whole	Artifact, Geographical, Subsidiary
Personal-Social	Business, Family, Lasting-Personal
ORG-Affiliation	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
Agent-Artifact	User-Owner-Investor-Manufacturer
General-Affiliation	Citizen-Resident-Religion-Ethnicity, Org-Location

Table 6. Nine kinds of labeling and instances

Notation	Description	Sentences for example	Instance
1-B	Entity M in Relation1(M,N)	中国首都北京 (Beijing of china capital)	中国 (China)
1-E	Entity N in Relation1(M,N)	中国首都北京 (Beijing of china capital)	北京 (Beijing)
2	Entity M,N in Relation 2(M,N)	城市北京和天津 (City Beijing and Tianjin)	北京 ,天津 (Beijing,Tianjin)
1-E-1-B	Entity N in Relation 1(M,N), 1(N,S)	位于中国首都北京的西单 (Xidan in Beijing of China capital)	北京 (Beijing)
1-E-1-E	Entity S in Relation 1(M,N), 1(N,S)	位于中国首都北京的西单 (Xidan in Beijing of China capital)	西单 (Xidan)
1-B-2	Entity M, S in Relation 1(M,N), 2（M,S)	美国总统访问中国 (The president of America visits China)	美国，中国 (America, China)
1-E-2	Entity N,S in 1(M,N), 2（N,S)	中国城市北京和天津 (City Beijing and Tianjin of China)	北京，天津 (Beijing, Tianjin)
2-1-B	Entity N in 2(M,N), 1（N,S)	美国总统访问中国北京 (The president of America visits China)	中国 (China)
2-1-E	Entity S in 2(M,N), 1（N,S)	美国总统访问中国北京 (The president of America visits China)	北京 (Beijing)

Table 7. Experiment result of entity relation extraction

Dataset	CRFs trained	Time (s)	Precision (%)	Recall R(%)	F-Score (%)
M1	Model 1	48	39.8	45.3	42.4
M2	Model 2	112	38.2	57.2	45.8
M3	Model 3	203	45.3	59.6	51.5
M4	Model 4	293	50.2	65.5	56.8
M5	Model 5	547	63.1	69.8	66.3
M6	Model 6	923	73.1	74.9	74.0
M7	Model 7	1982	76.9	84.7	80.8
M8	Model 8	2439	87.9	89.4	88.6
M9	Model 9	3010	90.5	95.8	93.1

$$P = \frac{r_1}{r_2} \times 100\%$$

$$R = \frac{r_1}{r_3} \times 100\%$$

$$F = \frac{2 \times P \times R}{P + R}$$

The data in table is the result of using M10 to test each model trained through M1-M9. Experiment shows that *P, R* and *F* increase with the more dataset used in training model. When we use M1 training the model F-Score is 42.4%. This is a disappointed value. But it only uses few sample dataset (newspaper of one day) for training. When using data of nine days we have obtained *P*,*R*,*F* values as 90.5%, 95.8% and 93.1% by use Model10. It illustrates it's a valid method. If we provide enough sample dataset we may win better result.

6. Conclusions

This paper discusses the information extraction of Chinese text based on CRFs which aims at the chunk parsing and relation extraction. Processing Chinese text is a complex system. This is an exploration because we haven't enough sample dataset for training CRFs model by now. But we think it's an effective method by experiments since CRFs model possesses working with global features.

At present we are developing a prototype of information extraction so a lot of work will be continued. Absence of training database is the common problem for many kinds of language. But manual label will be large cost. Therefore there are some researches on automatic or semi-automatic constructing dataset. In addition, how to select suitable feature set and improve precision is also future works.

REFERENCES

[1] E. C. Mary and J. M. Raymond, "Relational Learning of Pattern-match Rules for Information Extraction," Ph.D. Thesis, University of Texas, Austin, 1998.

[2] S. Stephen, "Learning Information Extraction Rules for Semi-Structured and Free Text," *Machine Learning*, Vol. 34, No. 13, 1999, pp. 233-272.

[3] D. Freitag and A. McCallum, "Information Extraction with HMM Structures Learned by Stochastic Optimization," *Proceedings of* 18*th Conference on Artificial Intelligence*, AAAI Press, Edmonton, 2002, pp. 584-589.

[4] R. Souyma and C. Mark, "Representing Sentence Structure in Hidden Markov Models for Information Extraction," *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Washington, 2001, pp. 1273-1279.

[5] T. Scheffer, C. Decomain and S. Wrobel, "Active Hidden Markov Models for Information Extraction," *Proceedings of the Fourth International Symposium on Intelligent Data Analysis*, Springer, Lisbon, 2001, pp. 301-109.

[6] D. Freitag, A. McCallum and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, 2000, pp. 591-598.

[7] H. L. Sun and S. W. Yu, "Shallow Parsing: An Overview," Contemporary Linguistics, 2000.

[8] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone and R. Weischedel, "Algorithms that Learn to Extract Information-BBN: Description Of The SIFT System as Used for MUC-7, *Proceedings of MUC*-7, Fairfax, 1998.

[9] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proceedings of the International Conference on Machine Learning* (*ICML*), 2001, pp. 282-289.

[10] Y. Y. Luo and D. G. Huang, "Chinese Word Segmentation Based on the Marginal Probabilities Generated by CRFs," *Journal of Chinese Information Processing*, Vol. 23, No. 5, 2009, pp. 3-8.

[11] M.-C. Hong, K. Zhang, J. Tang and J.-Z. Li "A Chinese Part-of-Speech Tagging Approach Using Conditional Random Fields," *Computer Science*, Vol. 33, No. 10, 2006, pp. 148-152.

[12] S. P. Abney and C. Tenny, "Parsing by Chunks. Principle based Parsing: Computation and Psycholinguistics," Kluwer Academic Publishers, Dordrecht, 1991, pp. 257-278.

[13] F. Erik, "Tjong Kim Sang and Sabine Buch holz. Introduction to the Conll-2000 Shared Task: Chunking," *Proceedings of CoNLL*-2000 *and LLL*2000, Lisbin, 2000, pp. 127-132.

[14] L. Ramshaw and M. Marcus, "Text Chunking Using Transformation-Based Learning," In: D. Yarovsky and K. Church, Eds., *Proceedings of the Third Workshop on Very Large Corpora*, Association for Computational Linguistics, Somerset, 1995, pp. 82-94.

[15] J. Hammerton, M. Osborne, S. Armstrong and W. Daelemans, "Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing," *Journal of Machine Learning Research*, Vol. 2, No. 3, 2002, pp. 551-558.

[16] K. Nanda, "Combining Lexical, Syntactic and Semantic Features with Maximum Entropy Models for Extracting Relations," *Proceedings of the ACL* 2004 *on Interactive poster and demonstration sessions*, Barcelona, 2004, pp. 22-25.

[17] D. Zelenko, C. Aone and A. Richardella, "Kernel Methods for Relation Extraction," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1083-1106.

[18] C. Whitelaw, A. Kehlenbeck, N. Petrovic, *et al*., "Web-Scale Named Entity Recognition," *Proceeding of ACM* 17*th Conference on Information and Knowledge Management*, Napa Valley, 2008, pp. 123-132.

[19] Z. Q. Chen, D. V. Kalashnikov and S. Mehrotra, "Adaptive Graphical Approach to Entity Resolution," *Proceedings of ACM IEEE Joint Conference on Digital Libraries*, Vancouver, 2007, pp. 204-213.

[20] X. P. Han and J. Zhao, "Person Name Disambiguation Based on Web-Based Person Mining and Categorization," 2*nd Web People Search Evaluation Workshop in conjunction with WWW*2009, Madrid, 2009.

[21] S. D. Pietra, R. L. Mercer and S. Roukos, "Adaptive Language Modeling Using Minimum Discriminate Estimation," *Proceedings of the Speech and Natural Language DARPA Workshop*, San Francisco, 1992, pp. 103-106.

[22] R. Rosenfeld, "Adaptive Statistical Language Modeling: A Maximum Entropy Approach," Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1994.

[23] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields Feature Induction and Web-Enhanced Lexicons," *Proceedings of CoNLL*-2003 *Association for Computational Linguistics*, Daelemans, 2003, pp. 188-191.

[24] K. Tjong, E. F. Sang and S. Buchholz, "Introduction to the CoNLL-2000 Shared Task: Chunking," *Proceedings of CoNLL*-2000 *and LLL*-2000 *Association for Computational Linguistics*, Lisbon, 2000, pp. 127-132.

[25] K. Tjong, E. F. Sang and J. Veenstra, "Representing Text Chunks," *Proceedings of EACL*'99, Association for Computational Linguistics, Bergen, 1995, pp. 173-179.

[26] J. Zhao, "A Survey on Named Entity Recognition, Disambiguation and Cross 2 Lingual Conference Resolution," *Journal of Chinese Information Processing*, Vol. 23, No. 2 March 2009, pp. 3-17.

