A Novel Hybrid Approach Incorporating Entity Characteristics for Vietnamese Chunking

Liu Yanchao1, Guo Jianyi1, 2, Xian Yantuan 1,2 ,Yu Zhengtao 1,2 ,Li Jia1

1 *School of Information Engineering and Automation*, Kunming University of Science and Technology, Kunming 650500, China
2 Intelligent Information Processing Key Laboratory, Kunming University of Science and Technology, Kunming 650500, China
{898559856@qq.com, gjade86@hotmail.com, yantuan.xian@gmail.com, ztyu@hotmail.com, 270981402@qq.com}
Corresponding author: E-mail gjade86@hotmail.com

***Abstract* -- Natural language processing is the key to realize artificial intelligent, text chunking plays an important role in natural language processing. Aiming at the shortage of Vietnamese chunking corpus as well as the low accuracy, this paper proposed a novel hybrid approach of Vietnamese chunking based on the Conditional Random Fields (CRFs) and Transformation-Based Error-Driven Learning incorporated with the Entity features. First of all, many features including the context information, words, part of speech and entity are all incorporated in CRFs statistical analysis, which is labeled the initial chunking recognition results. Secondly, On the basis of these results and in terms of the template of the rules, the paper analyzed the statistical information of the chunking tags by Transformation-Based Error-Driven Learning (TBL), the candidate transformation rules set can be obtained. Thirdly, according to filtering the candidate set using the function of evaluation, the final set of transformation rules were obtained; Finally, the final rules set were applied to correct the marked result of CRFs model for improving the recognition rate of chunking based on the TBL. The experiments were based on Vietnamese chunking corpus which was built by this paper and the results showed that the average accuracy reached to 89.7%; compared with Vietnamese Language Processing (VLSP), the Vietnamese chunking precision increases obviously.**

Keywords -- Vietnamese Chunking; Shallow Parsing; Conditional Random Fields; Transformation-Based Learning; Entity Characteristics

I. INTRODUCTION

Chunking is a fragment of a sentence which is decomposed into relatively independent fragments, these ordered fragments are adjacent, no nested, and the inner of fragments don't contain other types of chunking, but have the same kinds of chunks, meanwhile, and they have some meanings and syntactic function; chunking is the important part in the natural language processing. Therefore, text chunking was proposed as a shared task in the International Conference CONLL-2000 [1]. For example, the complexity of the complete syntactic parsing and phrase structure could be greatly reduced by chunking which could reduce ambiguity by increasing the granularity of words. Text chunking can provide some auxiliary information for the construction of phrase trees and dependency tree. Text chunking has played an key role in many fields such as Information Retrieval [2], Morphology Analysis [3], Semantic Analysis [4], [5], Opinion Mining [6], Information Extraction [7], Teaching [8], Syntactic Analysis [9], Algorithm Analysis [10] and machine translation.

The chunking research has now gained good achievements in English and Chinese. There are three types of methods mainly used in the chunk:1) Rules-Based [11][12] [13] [14], as in literature [11] the method of the finite state automaton was proposed by ABNEYS et al.(1991); Grace Ngai et al.(2001) [12], who combined English characteristics, had analyzed chunk with the method of Transformation-Based Learning and achieved good effects; the rule matching method of the error-driven-based learning was proposed to analyze chunking by RAMSHAW LA et al.(2002) [13]; afterwards, TBL was applied for improving Chinese text Chunking precision by Liu Y et al.(2006) [14]. For the diversity of language, rule-based method has many limitations of wasting time and energy, and difficult to count completely etc. 2) Statistical-Based [15] [16] [17], for instance in the literature, combining with English characteristics, the shallow syntactic was analyzed with the CRFs by Fei Sha et al.(2003)[15], gaining good result ; the method of the Chinese Chunking Using ESVM-KNN was proposed by Gao H et al.(2006)[16]; Chinese chunking were identified with conditional random fields algorithm by Sun G L et al.(2008)[17]; 3) Hybrid-Based [18][19] [20] [21] [22], for example, Li Sujian (2002) [18] used a combination method of statistical and rules to identify Chinese chunking, making up for their shortcomings each other; besides, Li Heng et al.(2005) [19] proposed stacking algorithm, every layer contained four classifiers(fnTBL,

SNoW, SVM, MBL), it used the contextual information as inputting feature vectors; Ying Liu et al. (2006) [20] identified Chinese chunking by the combining based on HMM and Transformation-Based Error-Driven Learning, it had achieved a certain results in English and Chinese; in the sequel, Wei Y et al. (2010) [21] identified the Chinese chunking with the method of combining Support Vector Machines, Border revised Rules and Transformation-Based Error-Driven Learning, and some effects were got by combining Chinese characteristics; Huang D G etc.(2009)[22] proposes a distributed strategy for Chinese text chunking based on Conditional Random Fields(CRFs) and Error-driven technique. The experimental results show that this kind of method is effective, outperforming the single CRFs-based approach, distributed method and other hybrid approaches.

In Vietnamese, there is only a little research work about chunking: tagging 1) the Vietnamese Noun phrases identification, which was identified with discriminant model (Le Minh Nguyen et al.)(2009)[23]; and the Vietnamese Noun Phrase Chunking were identified with CRFs model (Thao N T H et al., 2009) [24]. 2) Le Minh Nguyen et al.(2008)[25] identified chunking tags with the CRFs model in the limited Treebank resources, which included only 10,000 sentences syntactic trees. Owing to the limited scale of Vietnamese corpus as well as the low accuracy, this paper puts forwards a novel approach incorporating entity features to identify Vietnamese text chunking.

II. RELATED WORKS

A. CRFs-based Chunking

The CRFs are usually used to label and divide the probability model of the data sequence structure, it was proposed by J.Lafferty et al. (2001) in early period, and widely used in natural language processing and image processing. It constructs a discriminative framework which has been emerged recently in the field of statistical machine learning. This framework directly models the conditional probability distribution rather than the joint probability assumed. Moreover, the strong independence assumption of generative models is relaxed by CRFs. It has also addressed the problem of other discriminative models such as maximum entropy models which suffers from unexpected bias toward states. And it are mainly used in the sequence labeled research such as word segmentation[26], part of speech tagging[27], named entity recognition[28] and recognition chunking[25].

The principle of CRFs is as follows. Suppose, X indicates the observed sequence to be mark, C indicates the respectively joint probability distribution of the observed sequence, given an effective observation sequence X length of which is m, $X = x_1, x_2, \ldots\ldots, x_m$, obtained marked sequence C, $C = c_1, c_2, c_3, \ldots\ldots c_m$, defined as follows :

$$P(C / X) = \frac{1}{Z(x)} \exp\left\{\sum_{k=1}^{k} \lambda_k f_k(c_t, c_{t-1}, x_t)\right\} \quad (1)$$

In Equation (1), $Z(x)$ indicates a normalization factor constant, it makes all probability value in [0,1] of the result state sequence . $Z(x)$ expressed as follows:

$$Z(x) = \sum_{c} \exp\left\{\sum_{k=1}^{k} \lambda_k f_k(c_t, c_{t-1}, x_t)\right\} \quad (2)$$

In Equation (1), (2), is a transfer function , and represents the transition probability of marking sequence in the observing sequence X when the position in t and t-1, λ_k indicates the corresponding feature weight vector which need to be calculated in the training sample.

Minh Nguyen et al. (2008) [25] identified chunking by CRFs model with Treebank training corpus. Features such as words, part of speech tagging and the context information were fully used in his method. But due to the limited corpus scale, the method performance is limited. In general, the performance of statistical model depends on the corpus scale and selecting features. Therefore the training corpus scale is a bottleneck for all statistical method in some extent; Furthermore, the training features template has certain limitations for the language diversity, thus the features are difficult to find.

B. TBL-based Chunking

In 1992, Transformation-Based Error-Driven Learning (TBL) was proposed by Eric Brill which is a symbolic machine learning method. It can obtains rules automatically in the learning process and has also been applied in many fields and has achieved good results, such as part of speech tagging[29], word segmentation[30], Information Extraction[31], Phrases Extraction[32], word sense disambiguation and Chunking[33] etc..

The principle of TBL is as follows: one first applies the baseline heuristic to produce initial hypotheses of the training corpus, but the baseline prediction isn't correct, then the templates are used to form the candidate rules. This process identifies all the candidate rules generated by template set that would have a positive effect on the current tag assignments in corpus. Those candidate rules are tested against the rest of corpus, to identify the negative changes. Last, the entire process is repeated on the transformed corpus deriving candidate rules, scoring them, and selecting one with the maximal positive effect.

Ramshaw L A etc.(2002)[13] presented that the text chunking could be as a tagging problem, so transformation-based learning was applied in the text chunking, the approach could automatically induce a chunking model from supervised training. But the performance relies on rules set, and the template set was

somewhat hard to obtain and limited, it needs a simple or complex taggers or manual marking.

C. Hybrid Approach

Both CRFs and TBL have the limitations when they were used respectively, therefore, many researchers used them together to draw the strong points of others to offset their own weakness, which would achieved better results and was widely used in the field of natural language processing such as part of speech tagging, chunking tagging etc. [34]. The hybrid method uses the rules obtained by TBL to correct the chunking results produced by the CRFs, it has effectively overcome the shortage of the statistical method. In general, the chunk length is larger than the window size of CRFs template and the training corpus size is limited; once the training corpus has noise, it will seriously affect the performance of statistical models. Therefore, it is feasible to use TBL to improve the annotation results.

Sun Guanglu[35] et al. put forward the hierarchical clustering algorithm based on information entropy combined with Chinese words characteristics, and the word clusters which was generated by the algorithm are applied to the Chinese chunking model. Words and chunking tag were used as the basic features in Chinese chunking corpus. The syntactic function word clustering was produced by two hierarchical clustering. Results show that the proposed algorithm could improves the clustering efficiency, and the performance of the Chinese chunking system improved effectively. It means that the features selection is an important and indispensable part for chunking.

For chunking analysis, the entity features are perhaps a main factor, which concern the recognition of chunking boundary and types. Referring to the existing methods and ideas, Based on a certain scale corpus which were constructed by us, this paper proposes a novel hybrid method based on CRFs and TBL incorporating entity features. This approach makes fully use of the advantages of serialization marking, and overcomes the shortage of the corpus scale and the template windows size. Because of incorporating the entity features, this approach identifies the block boundaries and types more exactly. Experimental results show that our method has achieved better results and is feasibly and effectively.

III. VIETNAMESE CHUNKING DEFINITION

A. Vietnamese Characteristics

Each language has its own characteristics. If the language structure or semantics was different, the expression of word order would be different; in addition, if the complexity degree of phrase structure was different, then the sentence structure would be different. For instance, once noun phrase was divided into the simple noun phrase structure and the complex noun phrases, the complex noun phrases would be mainly reflected in the complexity of modifier component and other issues, it will brings difference to the chunk identification work.

Vietnamese is the official language in Vietnam. Influenced by multi-culture, it shows complex forms in the formation of words. The Vietnamese is an isolating language, whose word is constituted by one or more syllable (morpheme), which results in the complexity of word formation. After detailed analysis, it can be known that the proportion of noun chunking is about 86% in Vietnamese text chunking. Because of the uncertainty, complexity and inconsistency of the block length, the structure of the group is complicated; e.g., the noun phrase usually contains a core noun as well as the former modifying word or the post modifying word, the sequence of the modification is so complicated that it leads to the different word order and would affect the complexity of the noun phrase types. These problems as well as the lack of resources of the Vietnamese block etc. will bring severe difficulties and challenges for Vietnamese chunking.

B. Chunk Definition

Chunking definition has a great impact on the chunking results. Generally, the block definition is based on the description on CONLL-2000. But it should be combined with different languages characteristics. This paper made chunking definition with Vietnamese characteristics. Chunking is a fragment of a sentence which is decomposed into relatively independent fragments, these ordered fragments are adjacent, no nested, and the inner of fragments don't contain other chunking types, but they could contain the same chunking types, and had also some meanings and syntactic function; Vietnamese chunking is non-recursive, but could be nested, and the word order doesn't cross, it is unnecessary to cover the entire sentence for the Vietnamese chunking. For instance, the sentence punctuation doesn't belong to the inner chunk element.

TABLE 1. CHUNKING TYPES IN THE PAPER

No.	Tag	Type	For example
1	NP	Noun Chunk	quầy thể thao hàng hiệu...
2	VP	Verb Chunk	nhanh chóng...
3	AP	Adjective Chunk	đẹp...
4	PP	Preposition Chunk	từ...
5	QP	Quantitative Chunk	2.95 triệu đồng...
6	TP	Time Chunk	năm 2001...
7	WH	Doubt Chunk	vì sao...
8	ADV	Adverb Chunk	có khả năng...
9	O	others	。, “”...

The IOB2 sets were used to mark the sentence for Vietnamese chunking tags, and the B-X indicates the

beginning of X chunking type, I-X indicates the inside of X chunking type, O doesn't belong to any chunking types.

According to the statistics of Vietnamese chunking corpus, it was found that the chunking types such as Noun, Verbs, Adjectives, Preposition, Number, Time, Doubt and Adverb totally are about 98.3% and they are often used. Other chunking types as list chunking and exclamation chunking only have a little proportion. Therefore, a larger proportion chunking types was considered in the paper. 8 kinds of chunking types were defined as shown in Table 1.

IV. PROPOSED FRAMEWORK

A novel hybrid Vietnamese chunking method was proposed in the paper, which analyzes Vietnamese chunking and incorporates the entity features. The framework is as shown in Figure 1.

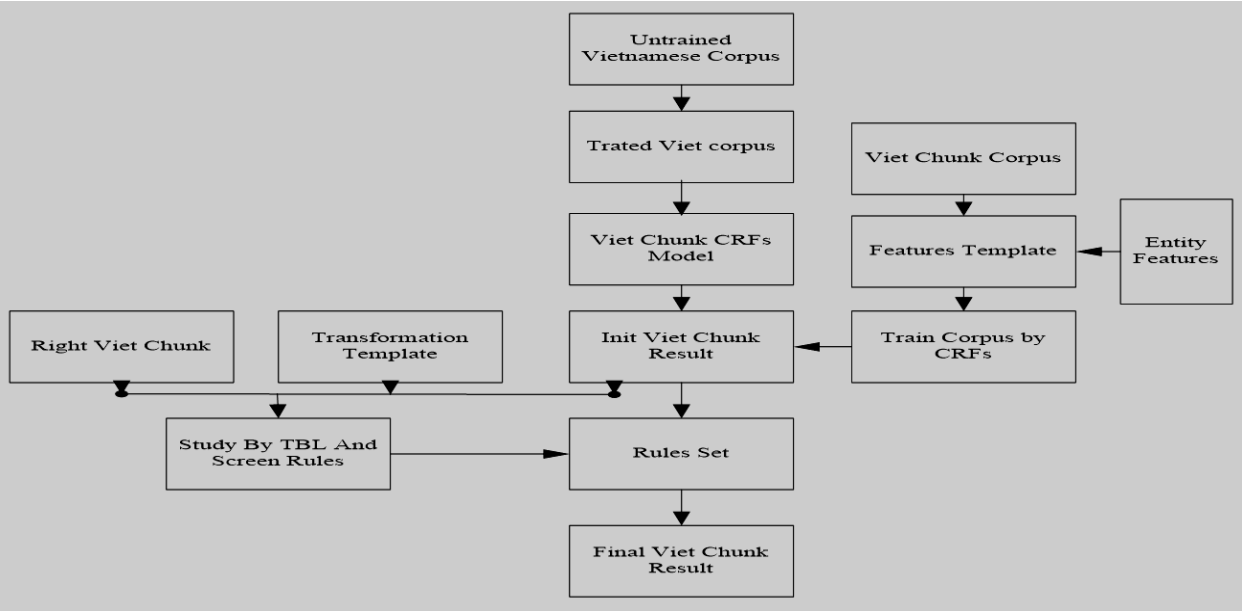

Figure 1 Framework of Vietnamese Chunk Analysis

Firstly, collecting and organizing the Vietnamese chunking corpus, applying the CRFs model for Vietnamese chunking and the preliminary recognition results of Vietnamese chunking were obtained; Secondly, after making the conversion rules template, combined the basic features with the entity features, the candidates rules set can be obtained by TBL method; further, the candidate rules set was filtered by evaluation function so as to obtain the conversion rules set. Finally, the untreated Vietnamese corpus were made for word segmentation, part of speech tags and manual proofreading, the CRFs model were used to label the chunking with the processed corpus; and the chunking was corrected by the conversion rules set, which would improve the accuracy of the chunking tagging.

Take a vietnamese sentence for instance: "Liệu Fenqing/Np có/V mãi_mãi/R là/V Fenqing/Np không/R ? (Will Fenqing remain to be Fenqing forever?) ", Firstly, the sentence are labeled by CRFs model, obtaining preliminary marks (Liệu/B-Np Fenqing/I-Np có/B-Vp mãi_mãi/I-Vp là/I-Vp Fenqing/B-Np không/B-Vp ?/CH), but "Liệu/B-Np Fenqing/I-Np" are labeled wrongly.

TABLE 2-A SAMPLE

Word	Pos	Entity tags	CRFs Label	TBL Revising Label
Liệu	Np	E-O	B-Np	B-Np
Fenqing	Np	E-O	I-Np	B-Np
có	V	E-O	B-Vp	B-Vp
mãi_mãi	R	E-O	I-Vp	I-Vp
là	V	E-O	I-Vp	I-Vp
Fenqing	Np	E-O	B-Np	B-Np
không	R	E-O	B-Vp	B-Vp
?	CH	E-O	O	O

It need to be revised with the generating conversion rules set. After Revising, the sentence are labeled properly as "Liệu/B-Np Fenqing/B-Np có/B-Vp mãi_mãi/I-Vp là/I-Vp Fenqing/B-Np không/B-Vp ?/CH ".

A. The Feature Selection

The selecting features is a key part for the probability CRFs model, it would directly affect the effectiveness and

chunking performance. Therefore, it is very important to select useful features.

B. The Basic Features

By analyzing the corpus which has been marked with chunking types, it can be found that features such as Words, Parts Of Speech, the Entity play a important role in identifying chunking types, as can be shown in Figure 2. The following features are selected: (1) the current word; (2) the current word pos; (3) the current word Entity tags; (4) the current word chunking tags, as is shown in Table 2.

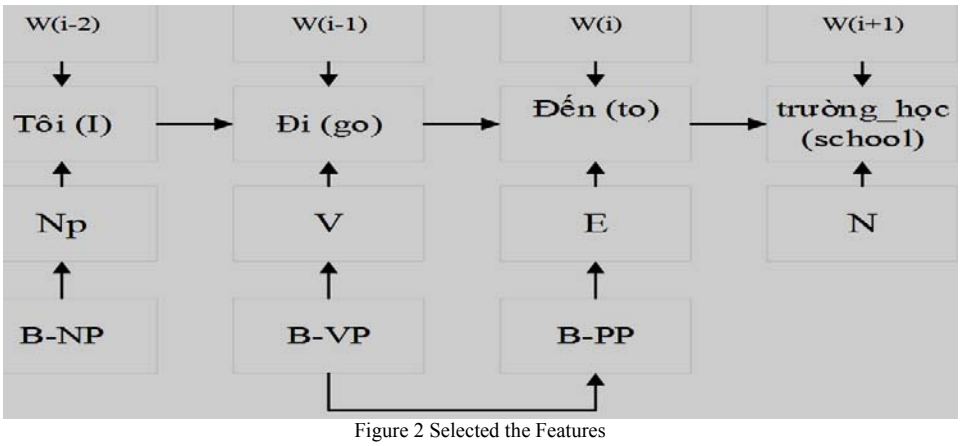

Figure 2 Selected the Features

TABLE 2-B BASIC FEATURES SET IN VIETNAMESE

No.	Feature	Means	No.	Feature	Means
1	w_i	Current word	3	c_i	Current Chunk tag
2	p_i	Current pos	4	e_i	Current Entity tag

C. Entity Features

There are many different entities in Vietnamese News corpus, including person name, place name, organization name and time etc. These entities would impact the chunking effect. Therefore, incorporating the entity features is beneficial to the chunking boundary and types identification. For instance, a sentence " The company manager Nguyen Thai " was represented in Vietnamese as:"Nguyễn_Thôi/B-NP quản_lý/I-NP công_ty/ I-NP.....", here,"Nguyễn_Thôi"is person name entity; another sentence " Import and export commercial manufacturing company " was represented in Vietnamese as "..... Công_ty/B-NP Sản_xuất/I-NP Kinh_doanh/I-NP hàng_xuất/I-NP nhập_khẩu/I-NP ....." here, "Công_ty Sản_xuất Kinh_doanh hàng_xuất nhập_khẩu" is an organization name, and the "Company" (Công_ty) is an the organization name's indication, it can be as a valid judgment condition in the chunking boundary identification. Time entity are also profit to the boundary identification and types. In addition, other entities as place name and the currency have the same effects. Therefore, it can be seen that the entity features were closely linked with chunking identification, the person name, place names, organization names, the currency and time entity were selected as entity features in this paper, as is shown in Table 3.

TABLE 3 NAMED ENTITIES CHARACTERISTICS

NO.	word	pos	Entity tags	Chunk tags
0	...	...	...	...
1	Tổng giám đốc	N	E-O	B-NP
2	CN	N	E-ORG-B	I-NP
3	Công_ty	N	E-ORG-I	I-NP
4	Du_lịch	N	E-ORG-I	I-NP
5	Hà_Tây	N	E-ORG-I	I-NP
6	tại	N	E-ORG-I	I-NP
7	Hải_Phòng	N	E-ORG-I	I-NP
8	...	...	...	...

D. Complex Features

Apart from the basic feature, the context information was also selected for the chunking identification in this paper. For instance, surrounding the current word, surrounding the current pos, surrounding the current entity as well as surrounding the chunking tags are selected as the effective features. As shown in Table 4.

TABLE 4 COMPLEX FEATURES SET

No.	Feature	Means	No.	Feature	Means
1	$p_{i-1}p_i$	Current pos and before one pos	12	$w_iw_{i+1}w_{i+2}$	Current word and after two words
2	$w_{i+1}w_{i+2}$	Current word after two word	13	$p_{i-2}p_{i-1}p_i$	Current pos and before two pos
3	$w_{i-2}w_{i-1}$	Before two word of The Current word	14	$p_{i-1}p_ip_{i+1}$	Current pos and after/before one pos
4	w_iw_{i+1}	Current word and after one word	15	$p_ip_{i+1}p_{i+2}$	Current pos and after two pos
5	$p_{i+1}p_{i+2}$	Current word after two word pos	16	e_ie_{i+1}	Current Entity tag and after one tag
6	$w_{i-1}w_i$	Current word before one word	17	$e_{i+1}e_{i+2}$	After two tag of Current Entity tag
7	$p_{i-2}p_{i-1}$	Before two word pos of the Current word	18	$e_{i-1}e_i$	Current Entity tag and before one tag
8	p_ip_{i+1}	Current word pos and after one pos	19	$e_{i-2}e_{i-1}$	Before two tags of Current Entity tag
9	$c_{i-2}c_{i-1}$	Before two chunk tag of the Current word	20	$e_ie_{i+1}e_{i+2}$	Current Entity tag and after two tags
10	$w_{i-2}w_{i-1}w_i$	Current word and before two words	21	$e_{i-1}e_ie_{i+1}$	Current Entity tag and after/before one tag
11	$w_{i-1}w_iw_{i+1}$	Current word and after/before one word	22	$e_{i-2}e_{i-1}e_i$	Current Entity tag and before two tags

In table 4, p represents the Vietnamese pos tags; c indicates the Vietnamese chunking tags, w means the Vietnamese words information, - / + is the surrounding current word features. i varies with the selecting window sizes as 2, it means that the two words before the current word and the two words after the current word as a range of templates considered, and c is only a negative value.

E. Obtaining the Rules Set

Each rule has two parts: (1) the correcting rule set; (2) the trigger environment. The correcting rule means that the marked words sequence of current chunk should be converted into the marked results of the target blocks. The trigger environment contains the information of words, part of speech tags, entity and chunking tagging of words sequence. Thereby, the key of TBL algorithm is to find the rules and triggering environment, and to correct the initial marked errors based on these rules as best as we can.

The processing flow of TBL is as follows:

(1) The preliminary Vietnamese Chunking marked results can be obtained by Using the CRFs model to label the corpus which had been tagged by the part of speech tagging tools and word-cut by the Word segmentation tools;

(2) The results of step(1) were used as the training data, and the TBL method was used to obtain the candidate rules set based on the template of the conversion rules;

(3) For each candidate rule, calculating the identification correct rate of chunking with using the rule r in the training corpus as ;

(4) Calculates the identification correct rate of chunking without using the rule r in the training corpus as ;

(5) The rule r will be reserved which meet the threshold conditions by screenings and will be put into the final correcting rules set;

(6) Until the candidate rules set had not included a candidate rule.

The rules of the evaluation function is used in this paper, here, represents the identification correct rate of chunking with using the rule r in the training corpus, represents the identification correct rate of chunking without using the rule r in the training corpus; represents the difference value between and . This paper selects zero as the critical point, that is: , when , it represents that the numbers of the right chunking are larger than that of the wrong chunking, and the rule r is added to the rules set; otherwise, the rule r is no sense. If the threshold chosen is too high, it would result in a smaller scale set, and the optimization result of chunking labeling is not obvious again.

V. EXPERIMENT RESULTS AND ANALYSIS

A. Prepare Corpus

In the paper, the corpus stems mainly from two aspects, The one is from the Vietnamese website in which Vietnamese sentences were crawled and labeled with CRFs model identification, and then the corpus were proofread by Vietnamese experts; The other is from Vietnamese Language Processing (VLSP) [36] website, namely http://vlsp.vietlp.org:8080/demo/, which was also proofread, marked, duplication removed by artificial. The total corpus have 203,902 words, all corpus is saved with UTF-8 and as shown in Table 5.

TABLE 5 SOURCE OF THE CORPUS

Source	Numbers	Proportion
VLSP	183512	90%
Manual processing	20390	10%

B. Experiment Assessment Standard

To evaluate the performance of the proposed method, we use well-accepted performance measures: precision (P), recall (R), and F-measure (F).experimental evaluation criteria are defined as follows:

$$\text{Precision} = \frac{\text{CorrectlyIdentifying Number}}{\text{Identifying Number}} * 100\% \quad (3)$$

$$\text{Recall} = \frac{\text{Correctly Identifying Number}}{\text{Chunk Number In testing set}} * 100\% \quad (4)$$

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Recall} + \text{Precision}} * 100\% \quad (5)$$

C. Experimental Design

The novel hybrid model which combined CRFs with TBL is employed to identify Vietnamese chunking in the paper. To evaluate the proposed method, we conduct a series of chunking experiments and compare the performance with other techniques.

Experiment 1: In order to verify the influence of the training corpus scale to hybrid model and hybrid model incorporating entity features, the experiments were conducted in different scale:100k, 120k, 140k, 160k, 180k and all; as can be shown in Figure 3,4.

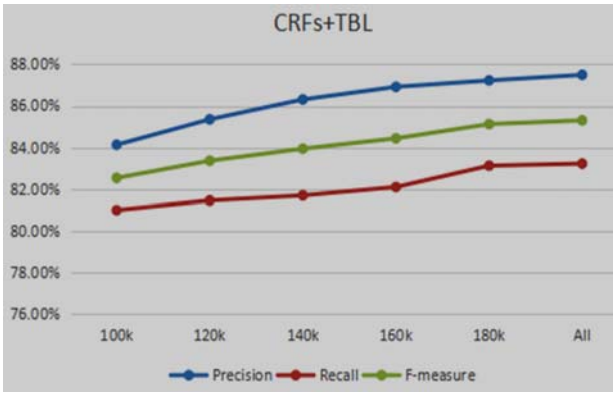

Figure 3 Influence of the Corpus Scale for Hybrid Model

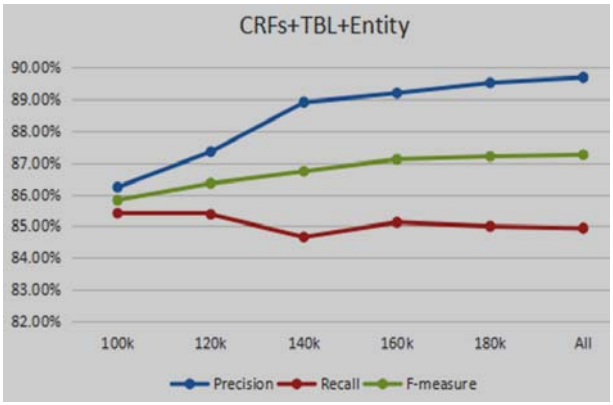

Figure 4 Influence of the Corpus Scale for Hybrid Model Incorporating Entity Features

From figure 3 and figure 4, with increasing corpus scale, the accuracy of Vietnamese chunking and F-measure will increase. Results shows that the corpus scale has a direct impact on chunking; after incorporating entity features, the accuracy of the hybrid model incorporated entity features is higher 2.2% than the accuracy of hybrid model. So incorporating entity features is beneficial to chunking identification.

Experiment 2: In order to evaluating the model performance more accurately, the corpus are divided into five equal parts, then made five-folds cross-validation experiments with CRFs, TBL and entity features, using 80% of the corpus as training data to obtaining the CRFs, TBL and entity features model and the transform learning rules set, using 20% of the corpus as testing data, we evaluate the model performance with average accuracy, as can shown in Figure 5.

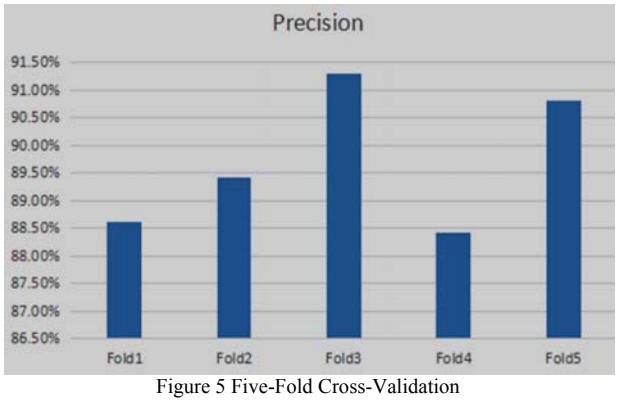

Figure 5 Five-Fold Cross-Validation

From figure 5, the fold 3 precision reaches 91.3% as a local maximum. The average accuracy is about 89.7% with CRFs, TBL and entity features model.

Experiment 3: To verify the accuracy of proposed methods is higher than that of the state of art methods-VLSP which based on CRFs, 10,000 syntactic trees from Vietnamese treebank are used to train model in VLSP, 3

groups experiments were conduct in the paper. As shown in Table 6.

TABLE 6 COMPARISON EXPERIMENTS

Method	P	R	F
VLSP(CRFs)	80.77%	79.85%	80.31%
CRFs+TBL	87.5%	83.21%	85.3%
CRFs+TBL+ Entity Features	89.7%	84.93%	87.25%

It can be seen from table 6 that the results of hybrid model is higher 6.73% than that of the VLSP, the hybrid model incorporated entity features is higher 2.2% than hybrid model. Therefore, in Vietnamese chunking recognition, the identification rate of the hybrid model incorporated entity features is higher than that of the single model of VLSP or the hybrid model, so the methods proposed in the paper is useful for chunking identification.

VI. CONCLUSION AND THE FUTURE WORK

There are some difficulties and challenges in Vietnamese chunking for the complex structure of blocks and the shortage of the Vietnamese chunking corpus. Based on the chunking corpus constructed by this paper, a novel hybrid approach is presented in this paper of Vietnamese chunking combining the statistical learning method with the rule-based one incorporated with many key features. This approach makes fully use of the advantages of serialization marking, and overcomes the shortage of the corpus scale and the template windows size. Because of joining the entity features, this approach identifies more precisely the block boundaries and types. Experimental results show that our method has achieved better results and is feasible and effective. Our future work will incorporated more the Vietnamese language characteristics such as the structural features of noun phrase and the phrase tree structure information etc. to analyze Vietnamese Chunking.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61262041, 61363044 and 61472168) and the key project of National Natural Science Foundation of Yunnan province (Grant No. 2013FA030).

REFERENCES

[1] Conll-2000,http://www.cnts.ua.ac.be/conll2000/chunking/

[2] Hongguang Suo,Shuying Cao.Chinese Automatic Abstracting System Based On Chunking[J]. Computer System Application,2007,03:97-100.

[3] Jing Shi,Guozhong Dai.Automatic Pos Based On Chunking and Memory[J].Journal of jilin University Engineering and Technology Edition,2006,4:560-563.

[4] Liang Yang,Fengming Pan,Hongfei Lin.Evaluate Object Recognition and Application Based On Chunking[J].Journal of Guangxi Normal University(Natural Science Edition),2011,01:151-156.

[5] Xiulong Zhang.Xinde Li,Xianzhong Dai.Path Natural Language Semantic Role Labeling Based On Chunking[J]. Journal of Southeast University(Natural Science Edition) ,2012,s1:127-131.

[6] Dozza M, Bärgman J, Lee J D. Chunking: A procedure to improve naturalistic data analysis[J]. Accident Analysis & Prevention, 2013, 58(5):309-317.

[7] Jihao Yin,Xiaozhong Fan,Panchao Zhao etc.Chinese Organization Name Recognition Based On Chunking[C]//CNNC2006 and CAISC06.2006.

[8] Jianzhong P. Colligation, collocation, and chunk in ESL vocabulary teaching and learning [J]. Foreign Language Teaching and Research, 2003, 6: 438-45.

[9] Arman A A, Arif B P N, Purwarianti A, et al. Syntactic Phrase Chunking for Indonesian Language[J]. Procedia Technology, 2013, 11(1):635–640.

[10] OL Mangasarian, ME Thompson.Chunking for massive nonlinear kernel classification,Optimization Methods & Software,2008,23(3):365-374

[11] ABNEYS, ABNEY S P.parsing by Chunks: principle-Based Paring[M].Dordrecht : Kluwer Academic Publishers ,1991:257-278.

[12] Grace Ngai and Radu Florian. Transformation Based Learning in the Fast Lane. In: "Proceedings of NAACL 2001", Pittsburgh, PA, USA, 2001.

[13] Ramshaw L A, Marcus M P. Text Chunking Using Transformation-Based Learning[J]. Text Speech & Language Technology, 2002, 11:82--94.

[14] Liu Y, Liao P. Improving Chinese text Chunkings precision using Transformation-based Learning[C]// Industrial Technology, 2006. ICIT 2006. IEEE International Conference on IEEE, 2006:2480-2485.

[15] Fei Sha, Shallow parsing with conditional random fields, Proceeding NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1,Pages 134-141

[16] Gao H, Huang D, Yang Y, et al. Chinese Chunking Using ESVM-KNN[C]// Computational Intelligence and Security, 2006 International Conference on. IEEE, 2006:731-734.

[17] Sun G L, Liu B Q, Wang X L, et al. Chinese chunking algorithm based on conditional random fields[C]// Machine Learning and Cybernetics, 2008 International Conference on. IEEE, 2008:2509-2513.

[18] Sujian Li,Qun Liu,Shuo Bai.Chinese Chunking Based On Combination Of Statistical And Rules Analysis[J].Computer Research and Development | Comput Res Dev,2002,04:385-391.

[19] Heng Li,Jinbo Zhu,Tianshun Yao.Combined Classifiers And Application in Chinese Chunking Based On Stacking algorithm[J].Computer Research and Development | Comput Res Dev,2005,05:844-848.

[20] Ying Liu ; Tsinghua Univ., Beijing ; Panpan Liao.Improving Chinese text Chunkings precision using Transformation-based Learning.Industrial Technology, ,pp.2196-2201,2006

[21] Wei Y, Zhang L Y, Zhang Y X, et al. Combining Support Vector Machines, Border Revised Rules and Transformation-based Error-driven Learning for Chinese Chunking[C]// Artificial Intelligence and Computational Intelligence, International Conference on. IEEE, 2010:383-387.

[22] Huang D G, Jing Y U. A Distributed Strategy for CRFs Based Chinese Text Chunking [J]. Journal of Chinese Information Processing, 2009.

[23] Le Minh Nguyen, Huong Thao Nguyen,Phuong Thai Nguyen, Tu Bao Ho and Akira Shimaz. 2009. An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models. In Proceedings of the 7th Workshop on Asian Language Resources (In Conjunction with ACLIJCNLP), pp. 9-16.

[24] Thao N T H, Thai N P, Minh N L, et al. Vietnamese Noun Phrase Chunking Based on Conditional Random Fields[C]// Knowledge and Systems Engineering, International Conference on. IEEE, 2009:172-178.
[25] Le Minh Nguyen, Hoang Tru Cao. 2008. Constructing a Vietnamese Chunking System. In Proceedings of the 4th National Symposium on Research, Development and Application of Information and Communication Technology, Science and Technics Publishing House, pp. 249-257.
[26] Gui K Z, Ren Y, Peng Z M. CRFs Based Chinese Word Segmentation[J]. Applied Mechanics & Materials, 2014, 556-562:4376-4379.
[27] Wu Z Q, Yu H Z, Wan S H. Research on Automatic Tagging of Parts of Speech for Tibetan Texts Based on the Condition of Random Fields[J]. Applied Mechanics & Materials, 2014, 519-520:782-785.
[28] Wang H, Zhao T, Li S, et al. A conditional random fields approach to biomedical named entity recognition[J]. Journal of Electronics, 2007, 24(6):838-844.
[29] Alex M, Zakaria L Q. Kadazan Part of Speech Tagging Using Transformation-based Approach[J]. Procedia Technology, 2013, 11(1):621-627.
[30] XIA Xin Song XIAO Jian Guo. A New Error-driven Learning Approach for Chinese Word Segmentation[J]. Computer Science, 2006, 33(3):160-164.
[31] Zhang C, Cao C, Niu Z, et al. A Transformation-Based Error-Driven Learning Approach for Chinese Temporal Information Extraction[M]// Information Retrieval Technology. Springer Berlin Heidelberg, 2008:663-669.
[32] KAM-FAI WONG, TIMOTHY KUN-CHUNG CHAN, CHUN-HUNG CHENG. An investigation on transformation-based error-driven learning algorithm for Chinese noun phrase extraction.[J]. International Journal of Computer Processing of Languages, 2001, 14:47-69.
[33] Liu Y, Liao P. Improving Chinese text Chunkings precision using Transformation-based Learning[C]// Industrial Technology, 2006. ICIT 2006. IEEE International Conference on. IEEE, 2006:2480-2485.
[34] PVS A, Karthik G. Part-of-speech tagging and chunking using conditional random fields and transformation based learning[J]. Shallow Parsing for South Asian Languages, 2007, 21.
[35] Sun Guanglu, Wang Xiaolong, Liu Bingquan, et al. Chinese chunking based on word clustering Statistical[J].Acta Electronica Sinica | Acta Electr Sin,8,36(12):2450-2453
[36] VLSP Project,http://vlsp.vietlp.org:8080/demo/