A Dual-layer CRFs Based Method for Chinese Nested Named Entity Recognition

Chunyuan Fu
School of Computer Science and Technology
Heilongjiang University
Harbin 150080, China
fuchunyuan@yahoo.cn

Guohong Fu
School of Computer Science and Technology
Heilongjiang University
Harbin 150080, China
ghfu@hlju.edu.cn

***Abstract*—While substantial studies have been performed on named entity recognition to date, nested named entity recognition as a research issue has not been well studied, especially for Chinese. In this paper, we take Chinese nested named entity recognition as a cascaded chunking problem on a sequence of words. To approach this problem, we first make a corpus-based investigation of nested structures for Chinese entities and thus propose a dual-layer conditional random fields (CRFs) based solution. To exploit more informative clues for nested named entity recognition, we employ a hybrid chunking scheme to represent the nest structures in Chinese named entities. Moreover, we have also examined the performance of different dual-layer models. Experimental results on different data sets show that the dual-layer CRFs with a hybrid chunk scheme achieve the best performance.**

Keywords: Chinese named entity recognition; nested named entities; dual-layer CRFs; entity chunk

I. INTRODUCTION

Named entity recognition (NER) is a process of identifying phrases in sentences that indicate the names of persons, organizations, locations, times or quantities. As a fundamental task in information extraction, NER plays a key role in many natural language processing (NLP) applications such as text mining, automatic summarization and machine translation, and has drawn much attention within the NLP community over the past years. It has been shared tasks of a number of conferences, such as the Message Understanding Conferences (MUCs) [1], the Conferences on Natural Language Learning (CoNLL) [2], the International Conferences on Language Resources and Evaluation (LRECs) [3], and the ACL-SIGHAN Bakeoffs [4].

Current research on NER focuses on machine learning approaches, and a variety of methods have been attempted, such as Hidden Markov Models (HMMs) [5], Maximum Entropy (ME)[6], Support Vector Machines (SVMs) [7], and Conditional Random Fields (CRFs) [8]. While machine-learning methods prove to be robust in open applications, it is still a challenge for most of them to keep a balance between capacity and computational cost [5]. Although HMMs are very speedy in training and tagging, they can only take into account category tags in context, and ignores some informative cues like contextual word forms for NER. On the contrary, some other machine learning techniques like SVMs and CRFs offer a straightforward way for exploring much richer lexical features for NER, but they usually need much more time for training and tagging, which will become a serious problem in processing a large amount of data or some on-line applications such as text mining.

While substantial studies have been performed on NER to date, the recognition of nested named entities is still a challenging task. Alex et al. (2007) apply a CRF-layer model to recognize nested named entities in biomedical texts [9], which first identifies the simple named entities embedded in nested NEs, and then recognizes other NEs. Finkel and Manning (2009&2010) present a discriminative constituency parser for English nested NER [10][11]. While their methods work well for nested named entity recognition, it is much slower than common flat techniques.

In this paper, we present a dual-layer CRFs based method for Chinese nested named entity recognition. We first perform an investigation of the structural characteristics of Chinese nested named entities using the entity-tagged PKU corpus [5], showing that most Chinese nested named entities have two-level nested structures. As such, we reformulate Chinese nested entity recognition as a cascaded chunking problem on a sequence of words, and thus propose a dual-layer solution under the framework of machine learning. To better handle nested structures in Chinese named entities and to exploit more informative features, particularly entity-internal structural cues for Chinese nested entity recognition at the same time, we employ different chunking schemes in different layers. Furthermore, we also consider four dual-layer models in the present study. Our experimental results over different data sets show that the dual-layer CRFs with a hybrid chunk scheme achieve the best performance.

The rest of this paper is organized as follows: Section 2 provides a brief investigation of nested structures in Chinese named entities. Section 3 describes in detail the proposed dual-layer method for Chinese nested NER. We report our experimental results and conclude our work in Sections 4 and 5, respectively.

II. CHINESE NESTED NAMED ENTITIES

In general, named entities (NEs) can be classified into two groups, namely simple NEs and nested NEs, in terms of

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

their structures. Simple NEs have a single-level flat structure and thus do not involve any other NEs within them, while nested NEs have multiple levels of structures and usually contain other NEs inside them.

To investigate the structural characteristics of nested NEs in Chinese, we utilize an entity-tagged version of the PKU corpus [5]. This corpus consists of one month of news texts from the *People's Daily* in 1998, and has been annotated with 46 different part-of-speech tags [12] and 13 different NE tags [5], respectively. It contains a total of 106430 named entities. It is noteworthy that nested structures mostly emerge in location names and organization names. So, in present study we focus on these two types of named entities.

Table I shows the distribution of different nested location and organization names in Chinese. As can be seen from this table, more than 18% of location and organization names are nested with other NEs, which accounts for about 7.5% of all NEs under discussion, This demonstrates once again the importance of nested named entity resolution for NER systems. In addition, nested structures are much more common in organization names than in location names. This is probably the reason why the state-of-the-art performance in organization name recognition is still not satisfactory to date.

TABLE I. DITRIBUTIONS OF CHINESE NESTED NES WITH TYPES

Entity category	Total	Number / Percentage of nested NEs over all NEs of the same category	Percentage of nested NEs over all NEs
LOC	26031	113/4.35%	1.06%
ORG	17086	6862/40.16%	6.45%
Total	43117	7993/18.54%	7.51%

To further investigate the hierarchy structures of Chinese nested NEs, we have also performed a statistical analysis of named entities in terms of their number of nested levels. The results are presented in TABLE II.

TABLE II. DISTRIBUTION OF CHINESE NES WITH NESTED LEVELS

Nested levels	Example	Number	Percentage
1	[中国/ns]、[新华社/nt]	35124	81.5%
2	[[中共中央/nt] 国务院/nt] [[黑龙江/ns] 哈尔滨/ns]	6864	16.0%
3	[[[中国/ns] 中医药/n 学会/n] 急诊/n 医学/n 分会/n] [[[海南/ns] 汽车/n 工业/n 公司/n] 北海/ns 分公司/n]	1046	2.4%
4	[[[[长沙市/ns] 公安局/n] 交警/j 支队/n] 党委/n] [[[[天津市/ns] 和平区/ns] 碧云里/ns] 3/m 号/q 楼/n]	83	0.1%

As illustrated in TABLE II, among all NEs in the entity-tagged PKU corpus, some 18.5% are nested NEs. Moreover, most nested NEs have a two-level hierarchy structure, and the number of levels for the deepest nested NEs is four.

III. THE APPROACH

Since most Chinese nested NEs have a two-level structure, we can reformulate Chinese nested named entity recognition as a dual-layer cascaded chunking task on a sequence of words, which identifying all simple NEs in the first layer, and then resolve nested NEs in the second layer. This section details the proposed dual-layer method.

A. CRFs for NER

We choose CRFs based modeling for NER in that it proves to be one of the most effective techniques for sequence labeling tasks [13]. In comparison with other methods, CRFs allow us to exploit a number of observation features as well as state sequence based features or other features to NER.

Let $X = (x_1, x_2, \cdots, x_T)$ be an input sequence of Chinese characters or words, $Y = (y_1, y_2, \cdots, y_T)$ be a sequences of entity-level chunk tags as shown in TABLE VI. From a statistical point of view, the goal of NER is to find the most likely sequence of entity chunk tags $\hat{Y}$ for a given sequence of characters or words X that maximizes the conditional probability $p(Y \mid X)$. CRFs modeling uses Markov random fields to decompose the conditional probability $p(Y \mid X)$ of a tag sequence as a product of probabilities below

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{T} \sum_{j} \lambda_j f_j(y, x, i)\right) \quad (1)$$

Where $f_j(y, x, i)$ is the j^{th} feature function at position i, associated with a weight λ_j , and $Z(x)$ is a moralization factor that guarantees that the summation of the probability of all sequences of entity-level chunk tags is one, which can be further calculated by

$$Z(x) = \sum_{y} \exp\left(\sum_{i=1}^{T} \sum_{j} \lambda_j f_j(y, x, i)\right) \quad (2)$$

B. Hybrid Label Scheme

To explore more entity-internal informative clues for nested named entity recognition, and to consider the efficiency in training at the same time, we distinguish simple NEs from nested NEs and use a hybrid label scheme, which employs two separate tagsets to label simple NEs and nested NEs, respectively.

For simple NEs, we use the traditional BIO tagset for chunk representation, which defines three tags, namely *B*, *I* and *O*, to denote the respective position patterns of a token within NEs. Where *B* indicates that the token is at the beginning of a multi-token entity, *I* denotes that the token is at the middle or the end of multi-token named entities, and *O* denotes that the token is an independent NE by itself.

As shown in TABLE III, we utilize an extended version of the BIO tagset for nested NEs. For convenience, we refer

it to as BIO-E tagset. We believe that the BIO-E tags could better represent the relatively complicated structures of nested NEs and thus provide a convenient way for exploiting more potential clues, in particular the entity-internal structural features for nested named entity recognition.

TABLE III. THE BIO-E TAGSET FOR LABELING NESTED NES

Tag	Definition
B	The current token is at the beginning of a multi-token NE.
I	The current token is at the second position of a multi-token NE.
M	The current token is at the middle of a multi-token NE.
E	The current token is at the end of a multi-token NE.
O	The current token is an independent NE by itself.

Furthermore, in the present study we take lexicon words (viz in-vocabulary words) as the basic tokens that form NEs. So in addition to the above two levels of chunk representation for simple NEs and nested NEs, we also introduce a lexical chunk labeling scheme to represent Chinese out-of-vocabulary words. TABLE VI illustrates the word-level and entity-level chunk representations of the sentence "广州/ns 标志/nz 公司/n 与/p 北京/ns 地质部/nt" (*Guangzhou Peugeot Company and China Ministry of Geology*) under the hybrid scheme. Where "n", "ns", "nt", "nz", and "p" are the PKU POS tags [12] for common nouns, toponyms, organization nouns, other proper nouns, and prepositions, respectively.

TABLE IV. AN EXAMPLE: CHUNK REPRESENTATION OF NES

Lexicon words	POS tag	Chunk tags for words	Chunk tags for simple NEs	Chunk tags for nested NEs
广州	ns	O-ns	O-LOC	B-ORG
标	nz	B-nz	O-nz	I-ORG
志		I-nz	O-nz	M-ORG
公司	n	O-n	O-n	E-ORG
与	p	O-p	O-p	O-O
中国	ns	O-ns	O-LOC	B-ORG
地质部	nt	O-nt	O-ORG	E-ORG

C. Feature Templates

TABLE V. FEATURE TEMPLATE FOR SIMPLE NER

NO.	Feature	Definition
0	w_0	The current token
1	t_0	The tag for the current token
2	w_{-1}/w_0	The preceding token / the current token
3	w_0/w_1	The current token / the following token
4	t_{-1}/t_0	The tags for the preceding / current tokens
5	t_0/t_1	The tags for the current /following tokens
6	$t_{-1}/t_0/t_1$	The tags for the preceding / current / following tokens
7	$w_{-1}/w_0/w_1$	The preceding / current / following tokens

With a view to the structural difference between simple named entities and nested named entities, we use different feature templates during the dual-layer chunking, as shown in TABLE IV and TABLE V, respectively.

TABLE VI. FEATURE TEMPLATE FOR NESTED NER

NO.	Feature	Definition
0	w_0	The current token
1	t_0	The tag for the current token
2	t_0	The current simple NE
3	t_0/t_1	The current / following simple NEs
4	t_{-1}/t_0	The preceding / current simple NEs
5	w_{-1}/w_0	The preceding / current tokens
6	w_0/w_1	The current / following tokens
7	t_{-1}/t_0	Tags for the preceding / current tokens
8	t_0/t_1	Tags for the current / following tokens
9	$t_{-1}/t_0/t_1$	Tags for the preceding / current / following tokens
10	$w_{-1}/w_0/w_1$	The preceding / current / following tokens

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate our approach, we have conducted several experiments on different data sets. The two open toolkits, namely the CRF++ toolkit [13] and the MaxEnt toolkit [14] are used in our experiments. The section reports the related experimental results.

A. Data Sets

We performed experiments on three data sets, namely the entity-tagged PKU corpus [5], the IEER-99 newswire data, and the MET2 test data. TABLE VII and TABLE VIII present the basic statistics of these three data sets.

TABLE VII. BASIC STATITISTICS OF THE ENTITY-TAGGED PKU CORPUS

Entity type	Training data		Test data	
	Total	*Nested NEs*	*Total*	*Nested NEs*
PER	27913	-	1796	-
LOC	23770	935 / 3.93%	2261	196 / 8.67%
ORG	15483	5897 / 38.09%	1603	965 / 60.20%

TABLE VIII. BASIC STATITISTICS OF IEER-99 AND MET2 DATA SETS

Entity type	IEER-99		MET2	
	Total	*Nested NEs*	*Total*	*Nested NEs*
PER	489	—	174	—
LOC	1026	60 / 5.85%	750	84 / 11.20%
ORG	497	236 / 47.48%	377	224 / 59.42%

B. Experimental Results

Our first experiment is aiming at examining the performance of different dual-layer models in nested entity

recognition. They are ME+ME, ME+CRF, CRF+ME, and CRF+CRF. In this experiment, we use the IOB tagset through the two-layer chunking. In addition, two single-layer models, namely the single-layer ME and CRFs based systems, are also involved as baselines. This experiment is conducted over the entity-tagged PKU corpus. The results are summarized in TABLE IX and TABLE X.

TABLE IX. RESULTS OF SINGLE-LAYER MODELS FOR NER

NE	System	Type	P (%)	R (%)	F (%)
Simple NEs	ME	-	99.53	98.02	98.70
	CRF	-	99.78	99.89	99.83
Nested NEs	ME	LOC	76.92	43.47	55.55
		ORG	56.14	47.76	51.61
		Total	57.60	47.32	51.96
	CRF	LOC	85.71	49.08	62.41
		ORG	88.03	62.19	72.89
		Total	88.44	58.04	70.08
All NEs	ME	-	85.83	81.08	83.39
	CRF	-	92.25	80.14	85.77

TABLE X. COMPARISON OF DUAL-LAYER MODELS FOR NESTED ENTITY RECOGNITION

Dual-layer models	Nested entity type	P (%)	R (%)	F (%)
ME+ME	LOC	84.62	47.83	61.11
	ORG	59.64	49.25	53.95
	Total	61.45	49.10	54.59
CRF+ME	LOC	83.33	43.49	57.14
	ORG	59.65	52.23	55.70
	Total	61.17	51.34	55.82
ME+CRF	LOC	85.71	45.09	59.09
	ORG	86.39	63.18	72.99
	Total	86.36	59.38	70.37
CRF+CRF	LOC	88.88	46.78	61.29
	ORG	84.84	69.65	76.50
	Total	85.54	66.07	74.55

As shown in TABLE IX, the performance difference between the two single-layer models is not significant for simple NER. However, for nested NER, the single-layer CRFs based method is much better than the single-layer ME model.

By comparing the results in TABLE IX and TABLE X, we can observe that the dual-layer models consistently outperform the single-layer models in nested named entity recognition. Meanwhile, the dual-layer CRFs yield the best F-score among all dual-layer models under discussion.

The goal of our second experiment is to examine the effects of different chunk schemes on nested entity recognition. In particular, we introduce two schemes in this experiment, namely BIO+BIO and BIO+BIO-E for comparison under the framework of the dual-layer CRFs. The results are presented in TABLE XI.

From Table XI, we can see that the original dual-layer CRFs based system with a unified BIO scheme obtains an F-score of 74.55% on nested named entity recognition. This figure can be increased by 3.41% after using a hybrid scheme of BIO+BIO-E.

TABLE XI. EFFECTS OF DIFFERENT CHUNK SCHEMES ON NESTED ENTITY RECOGNITION

Chunk Schemes	Nested entity type	P (%)	R (%)	F (%)
BIO+BIO	LOC	88.88	46.78	61.29
	ORG	84.84	69.65	76.50
	Total	85.54	66.07	74.55
BIO+BIO-E	LOC	84.51	49.72	62.60
	ORG	88.12	74.68	80.84
	Total	88.09	69.91	77.96

To demonstrate the effectiveness of the proposed dual-layer CRFs based method with a hybrid label scheme, we have also conducted an open evaluation on the IEER-99 newswire test data and the MET2 test data, and compare it with other state-of-the-art systems for the two shared tasks. The experimental results are presented in TABLE XII and TABLE XIII, respectively.

TABLE XII. COMPARISON OF DIFFERENT SYSTEMS ON MET2 DATA

System	Entity type	P (%)	R (%)	F (%)
Our system	ORG	82.38	88.27	85.22
	LOC	90.24	89.63	89.94
The KRDL system [15]	ORG	88.00	89.00	88.50
	LOC	91.00	89.00	90.00
The NTU system [16]	ORG	78.00	85.00	81.30
	LOC	78.00	69.00	73.20

TABLE XIII. COMPARISON OF DIFFERENT SYSTEMS ON IEER-99 DATA

System	Entity type	P (%)	R (%)	F (%)
Our system	ORG	86.01	82.37	84.15
	LOC	84.37	91.19	87.64
The Microsoft system [17]	ORG	62.30	88.03	72.96
	LOC	80.18	79.09	79.63
The CAS-IA system [18]	ORG	71.08	86.09	84.61
	LOC	84.69	88.31	86.47

It can be seen in TABLE XII and TABLE XIII that our system can achieve performance that is better than or comparable to the state-of-the-art systems over both test data sets, illustrating in a sense the effectiveness of the proposed method.

V. CONCLUSIONS

In this paper, we have presented a dual-layer CRFs based method for Chinese nested named entity recognition.

We show that most nested named entities in Chinese have a two-level structure and the dual-layer CRFs with a hybrid label scheme can achieve state-of-the-art performance, illustrating that it is very suitable for Chinese nested entity recognition.

While our current method proves to be effective for most nested entity, it takes simple named entity decoding and nested entity decoding as two separate tasks. To further enhance our system, in future we intend to take into account the interaction of dual-layer decoding and exploit a unified decoding strategy for Chinese NER.

ACKNOWLEDGMENT

This study was supported by National Natural Science Foundation of China under Grant No.60973081 and No.61170148, the Returned Scholar Foundation of Educational Department of Heilongjiang Province under Grant No.1154hz26, and Harbin Innovative Foundation for Returnees under Grant No.2009RFLXG007, respectively.

REFERENCES

[1] Nancy A. Chinchor, "Overview of MUC-7 and MET2", Proceedings of 7th Message Understanding Conference, 1998.

[2] Erik F. Tjong Kim Sang, and Fien De Meulder, " Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition", Proceedings of CoNLL'03, 2003, pp. 142-147.

[3] Feorge Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel, "The automatic content extraction program: Tasks, data and evaluation", Proceedings LREC'04, 2004.

[4] Guangjin Jin, and Xiao Chen, "The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging", Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, 2008, pp.69-81.

[5] Guohong Fu, and Kang-Kwong Luke, "Chinese named entity Recognition using lexicalized HMMs" , ACM SIGKDD Explorations Newsletter, vol.7, no.1, 2005, pp.19-25.

[6] Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra, "A hybrid feature set based maximum entropy hindi named entity recognition", Proceedings of ACL-IJCNLP'08, 2008, pp.343-349.

[7] Asif Ekbal, and Sivaji Bandyopadhyay, "Bengali named entity recognition using support vector machine", Proceedings of ACL-IJCNLP'08 Workshop on NER for South and South East Asian Languages, 2008, pp.51-58.

[8] Jingchen Liu, Minlie Huang, and Xiaoyan Zhu, "Recognizing biomedical named entities using skip-chain conditional random fields," Proceedings of ACL'10, 2010, pp.10-18.

[9] Beatrice Alex, Barry Haddow, Claire Grover, "Recognising nested named entities in biomedical text", Proceedings of the BioNLP Workshop at ACL'07, 2007, pp. 65-72.

[10] Jenny Rose Finkel, and Christopher D.Manning, "Nested named entity recognition", Proceedings of EMNLP'09, 2009, pp.141-150.

[11] Jenny Rose Finkel, and Christopher D. Manning, "Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data", Proceedings of ACL'10, 2010, pp.720-728.

[12] Shiwen Yu, Huimin Duan, Xuefeng Zhu, Bin Swen, and Baobao Chang, "Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation" , Journal of Chinese Language and Computing, 2003, vol. 13, no.2, pp.121-158.

[13] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Proceedings of ICML'01, 2001, pp.282-289.

[14] Le Zhang. "Maximum entropy modeling toolkit for Python and C++", 2004, available at *http://homepages.inf.ed.ac.uk/lzhang10/software/maxen/manual.pdf*.

[15] Shihong Yu, Shuanhu Bai, and Paul Wu, "Description of the Kent Ridge Digital Labs system used for MUC-7", Proceedings of MUC-7, 1998.

[16] Hsin-Hsi Chen, Yung-Wei Ding, Shih-Chung Tsai, and Guo-Wei Bian, "Description of the NTU system used for MET2", Proceedings of MUC-7, 1998.

[17] J. Sun, J. Gao, L. Zhang, M. Zhou, and C. Huang, "Chinese named entity identification using class-based language model.", Proceedings of COLING'02, 2002, pp.967-973.

[18] Youzheng Wu, Jun Zhao, and Bo Xu, "Chinese named entity recognition combining a statistical model with human knowledge", Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 2003, 65-72.