Morpheme-Based 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—**Named entity recognition plays an important role in many natural language processing applications. While considerable attention has been pain in the past to research issues related to named entity recognition, few studies have been reported on the recognition of nested named entities. This paper presents a morpheme-based method to Chinese nested named entity recognition. To approach this task, we first employ the logistic regression model to extract multi-level entity morphemes from an entity-tagged corpus, and thus explore a variety of lexical features under the framework of conditional random fields to perform Chinese nested named entity recognition. Our experimental results on different data set show that our system is effective for most nested named entities under evaluation.**

Keywords: nested named entity; entity morphemes; conditional random fields

I. INTRODUCTION

As an important sub-task of information extraction, named entity recognition (NER) has received considerable attention from natural language processing community. It has been a shared task of a number of conferences, including the Message Understanding Conferences (MUCs), Automatic Content Extraction (ACE), the Conferences on Natural Language Learning (CoNLL) and the Special Interest Group of the Association for Computational Linguistics on Chinese Language Processing (SIGHAN).

These shared tasks have greatly promoted the development of NER technology, especially English NER technology. However, to develop a practical NER system for Chinese is still a challenge. On the one hand, Chinese NER is usually formalized as a sequence labeling task with characters or words as the basic tokens. However, it is difficult to handle entity-internal structural features for Chinese NER based on either characters or words. On the other hand, while the recognition of simple named entities (NEs) is well studied over the past years, few works have been reported on the identification of nested NEs, which might be the major source of errors for existing systems for Chinese NER.

To address the above issues, this paper introduces the concept of entity morphemes into Chinese NER and presents a morpheme-based two-layer method to Chinese nested NER. To approach this, we first employ the logistic regression model to extract multi-level entity prefixes and suffixes from an entity-tagged corpus, and thus explore a variety of lexical features under the framework of conditional random fields (CRFs) to perform Chinese nested NER. Our experimental results show that the introduction of entity morphemes is beneficial to the improvement of Chinese NER performance.

The rest of this paper is organized as follows: Section 2 provides a brief overview of the related work on NER. In section 3, we analyze the structures of Chinese nested named entities. Section 4 gives a morpheme-based representation for entity chunks. Section 5 describes the features used for nested NER and a logistic transformation-based technique for entity morpheme extraction. We report our experimental results in section 6 and finally conclude our work in section 7.

II. RELATED WORK

Over the past years, a number of machine learning approaches have been attempted for NER, such as the Hidden Markov Model(HMM)[1][2], the Maximum Entropy Model (MEM)[3], the Support Vector Machines (SVMs) [4], the Decision Tree Machine (DTM) and the Conditional Random Fields (CRF)[5]. Zhang et al. (2008) combined multiple features, including both local and global constraint information for Chinese NER under the ME framework. They also introduced heuristic knowledge to reduce the searchable space. They showed that their system can achieve an F-score of 86.31% over the SIGHAN Bakeoff 2008 data set. Tsai et al. (2005) employed the ME model to exploited multiple shallow linguistic information such as spelling, parts of speech, word forms, context and other features for biological NER [7]. However, the recognition of named entities with complex structures is not satisfactory. CRF proved to be more effective for NER [8][9]. Wang (2009) proposed a two-stage method that incorporates rules with CRFs to perform biological NER [8]. More recently, She and Zhang (2010) incorporated CRFs with MNE rules for musical NER [9]. Finkel and Manning attempted a complicated approach that combine three models for NER, namely a syntax model, an entity model and a joint model and tested it on a united marked corpus [10]. However, the precision and the recall are not satisfactory.

Nest NER is an important but difficult issue in the field of NER. Finkel and Manning (2010) proposed a discriminatory selection algorithm to train a structural model for English nested NER. However, the method did not work well in news

978-1-61284-181-6/11/$26.00 ©2011 IEEE

reports and biomedical field. Compared with other methods, it is also time-consuming. Alex et al. (2007) applied a CRF-layer model to English NER [12], which first identifies the simple NEs embedded in nested NEs, and then recognizes other NEs.

III. Nested Named Entities In Chinese

To investigate the structural characteristics of Chinese nested named entities, we use an entity-tagged version of the PKU corpus [2]. This corpus consists of one month of news text from the People's Daily, which has been annotated with 46 different part-of-speech tags and thirteen different named entity tags, respectively. It contains a total of 106430 named entities. In the present study, we focus on the three main named entities, namely person names, location names and organization names.

We analyze the form of nested named entities and educe three types.

(1) A same type of NEs paratactically embedded in one named entity: In this case, a named entity contains a same type of entities, and there is no hierarchical relationship between these embedded NEs.

(2) Different types of NEs paratactically embedded in one named entity: In this case, a named entity involves different types of entities, and there is no hierarchical relationship between these embedded NEs.

(3) Various types of named entities nested in one named entity: In this case, a named entity consists of multiple entities and there is a hierarchical relationship between these embedded NEs.

We further analyze the structure of nested named entities, and summarize four forms of nested named entities.

TABLE I. The structure of nested named entities

NO.	Type	Example
1	prefix + others + suffix	索非亚/nr 大/a 教堂/n (*Sofia Cathedral*)
2	prefix + PERs + others + suffix	湛江/B-ns 市/E-ns 惠珍/nr 联合 /v 医院/n (United Hospital of Huizhen in Zhanjiang City)
3	prefix + LOCs + others + suffix	中国/ns 驻/v 南非/ns 大使/B-n 馆/E-n (Chinese Embassy in South Africa)
4	prefix + ORGs + others + suffix	纽约/ns 联合/B-nt 国/E-nt 总部 /n (United Nations Headquarters in New York)

As illustrated in TABLE I, most nested NEs begin with an entity prefix and end with an entity suffix. Their middle part is made up of one or more person names, location names, organization names or other words. The entity prefix can be a person name, location name or organization name, and plays a very important role in nested NER.

TABLE II. Distribution Of Differeent Nested Named Entities Interms of Nested Levels

NE type	Nesting level	Number	Percentage
Simple NEs	One-level	35124	81.5%
Nested NEs	Two-level	6864	16.0%
	Three-level	1046	2.4%
	Four-level	83	0.1%

Furthermore, nested NEs usually have a complex hierarchy structure. Thereby, we can further distinguish different nested NEs in terms of their number of nested levels. As illustrated in TABLE II, there are a large number of nested named entities in our corpus, and the deepest nested NEs have a four-level nested structure.

IV. Morpheme-based chunk representation

Morpheme is a grammatical unit that is smaller than words. It will be effective to extract the important information of named entities. Accordingly it can improve the accuracy of identification. At the same time, we use a deeper sequence tag method to exploit the features of nested named entities. TABLE III shows the tag set for representing entity chunks.

TABLE III. Tagset of named entity chunking

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

In this study, we transform our corpus into the morpheme-based corpus, and then merge the lexical chunks and entity chunks in Chinese NER.

TABLE IV. An example: Lexical chunk representation vs. Entity chunk representation

Word	POS tag	Morpheme	Lexical chunk tag	Entity chunk tag
参观	v	参观	Q-v	Q
了	u	了	Q-u	Q
北大	j	北	B-j	B-ORG
		大	E-j	E-ORG
、	w	、	Q-w	Q
清华	j	清	B-j	B-ORG
		华	E-j	E-ORG
和	c	和	Q-c	Q
抗日战争	nz	抗日	B-nz	B-LOC
		战争	E-nz	I-LOC
纪念馆	n	纪念	B-n	M-LOC
		管	E-n	E-LOC

As illustrated in TABLE IV, lexical chunk tags and entity chunk tags follow a similar format like S_1-S_2, where S_1 refers to the position patterns of a token in chunk formation, and S_2 denotes the respective categories of words or entities that contain the token.

V. Features

The exploration of informative features is essential to nested NER. Based on the above morpheme-based representation of Chinese NEs, in this section we continue to exploit a variety of entity-internal and entity-external features for Chinese NER.

A. *Lexical features*

Lexical information plays an important role in NER. In the present study we consider multiple lexical features, such as morpheme forms, part of speech, and the position of entity morphemes within entities. Furthermore, we also take into account the contextual lexical information outside a lexical unit For example, given an n-morpheme sentence. The observation is not limited to the current morphme *i*, lexical features within a window of five morphemes, namely (*i*-2, *i*-1, *i*, *i*+1, *i*+2), are explored for NER.

B. Multi-level prefix and suffix features

In fact, most nested named entities in Chinese begin with a prefix and end with a suffix. Such morphological information is obviously an important source of indicators for nested NER. However, there are some words or morphemes that usually occur before or after named entities but hardly appear as part of nested named entities. In other words, we need to distinguish useless entity morphemes from informative entity morphemes while exploring morphological features for nested NER.

To approach this, we employ the logistic transform method in the logistic regression model to calculate the importance of some potential entity morphemes, and thus extract a set of useful entity prefixes and suffixes for nested NER. The principle of logistic transform is showed as follows:

First of all, a two-dimension contingency table is constructed from the training data.

TABLE V. N × 2-DIMENSIONAL CONTINGENCY TABLE TO DESCRIBE THE NUMBER OF PREFIXES AND SUFFIXES, CONSTITUTING NESTED NAMED ENTITIES

Entity morpheme	**Number of entity morphemes forming nested NEs**	**Total number of morphemes in training data**	**Probability**
w_1	m_1	n_1	p_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
w_n	m_n	n_n	p_n
$\sum_i$	$m_\bullet$	$n_\bullet$	$P_\bullet$
	Frequency		Probability

As illustrated in TABLE V, let $W=w_1\ w_2 \cdots w_n$ be a set of , entity morphemes extracted from the training data, m_i is the number of the entity morpheme w_i that constitutes nested named entities, n_i denotes the total number of morphemes in the training data, and p_i is the probability of the entity morpheme w_i constituting nested named entities.

From the mathematical point of view, directly calculating the value of the probability *p* may be difficult, for the range of *p* is from 0 to 1. So the relationship between the independent variable and *p* is hard to describe by linear model. Furthermore, it is also hard to find and deal with the tiny change when the value of *p* is close to 0 or 1. Under this situation, we do not deal with *p* directly. Instead, we count a strictly monotonic function of *p*. The function is represented by $Q=Q(p)$, where $Q(p)$ is sensitive enough to small changes when *p* trends to 0 or 1. That is, dQ/dp should be directly proportional to $p/(1-p)$. The function Q can be defined as:

$$Q = Logit(P) = \ln(\frac{P}{1-P}) \qquad (1)$$

When *p* changes from 0 to 1, the range of *Q* is $(-\infty,+\infty)$. For each entity morpheme w_i, we use P^* to represent *Q*, and the above formula can be further rewritten as

$$\mathrm{P_i}^* = Logit(P) = \ln(\frac{P_i}{1-P_i}) \qquad (2)$$

The formula requires $P_i^* \neq 0$ or $P_i^* \neq 1$. That is $m_i \neq 0$ and $m_i \neq n_i$. So, if $m_i = 0$ or $m_i = n_i$, we must amend P_i* using formula (3).

$$\mathrm{P_i} = \frac{m_i + 0.5}{n_i + 1} \qquad (3)$$

In this way, we can calculate the importance of entity prefixes and entity suffixes in forming nested NEs, and thus remove some useless morphemes. Meanwhile, we can figure out the weight of each entity prefix and suffix, as shown in TABLE VI and TABLE VII, respectively.

TABLE VI. THE WEIGHTS OF SOME ENTITY PREFIXES

Useful prefixes	**Weights**	**Useless prefixes**	**Weights**
中东 'Middle East'	5.5255	世界 'world'	-2.9435
美国 'America'	5.5636	华夏 'China'	-2.5649
淮 'Abbreviation of the Huaihe River'	4.4427	国家 'nation'	-1.5063
空军 'air force'	3.9120	中 'Abbreviation of China'	-1.1570
南亚 'South Asia'	2.5649	首都 'Capital'	-2.3573

TABLE VII. THE WEIGHTS OF SOME ENTITY SUFFIXES

Useful suffixes	**Weights**	**Useless suffixes**	**Weights**
议会 'parliament'	3.7613	学校 'school'	-1.6831
委 'committee'	3.9195	所 'institute'	-1.5077
公司 'company'	0.3809	部门 'department'	-4.2022
院 'institute'	2.8140	大会 'conference'	-0.6497
支队 'detachment'	2.1972	企业 'enterprise'	-4.6914

As illustrated in TABLE VI and TABLE VII, if the weights of entity prefixes or suffixes are greater than 0, they can be identified as important prefixes or suffixes for nested NER. Based on this standard, we extracted 1878 informative entity prefixes and 881 entity suffixes for nested NER from a total of 2804 candidate entity prefixes and 1402 candidate entity suffixes, respectively. After a manual checking, we finally obtain 950 entity prefixes and 316 entity suffixes..

TABLE VIII. RANKS OF ENTITY SUFFIXES

Level	**Suffix**	**Weight**
Level 1	议会 'parliament'	3.761273
	委 'committee'	3.919495
Level 2	公司 'company'	0.380874
	支队 'detachment'	2.197225

To further handle the particularity and complexity of a special useful entity suffix, we re-rank the extracted entity suffixes according to their weights, and thus divide them into two levels based on a given threshold. In the present study, the threshold for ranking entity suffixes is set to 3. TABLE 8 illustrates the division of some entity suffixes.

VI. Experimental Results And Discussions

To test the validity of our method, we used the CRF++ toolkit [], and conducted experiments on several data sets. This section reports the experimental results.

A. Experimental data

In our experiments we employ the entity-tagged corpus [2]. To achieve a morpheme-based system for Chinese NER, we transform the original word-based corpus to a morpheme-based format using the maximum forward matching method. This corpus is further divided into two parts: 90% is used as the training data, and the rest 10% is for test. TABLE IX shows the distribution of different named entities in the experimental corpora.

TABLE IX. Distribution of named entities In The Training Data And The Test Data

Type	Training data		Test data	
	Number of all NEs	Number of nested NEs	Number of all NEs	Number of nested NEs
PER	27913	—	1796	—
LOC	23770	935	2261	196
ORG	15483	5897	1603	965

B. Experimental metrics

In our experiments, we evaluate our system in terms of recall (R), precision (p) and F-score (F). The recall is defined as the number of correctly-recognized entities divided by the total number of entities in the test data, while the precision can be interpreted as the number of correctly-recognized entities is divided by the total number of entities yielded automatically by the system. The F-score is the balanced value of recall and precision, namely F-score=2*P*R/(P+R).

C. Experimental results

To evaluate our approach, we have conducted three experiments.

Our first experiment is designed to test the effects of different chunking tokens, namely morphemes and words on NER. We conduct this experiment by applying the CRF-based chunker to a morpheme-based data set (referred to as CRF_M) and a word-based data set (referred to as CRF_W), respectively. The experimental results are presented in TABLE X.

TABLE X. NER Performance: Morpheme-Based Chunking Vs. Word-Based Chunking

System	Nested NEs			NEs		
	P%	R%	F%	P%	R%	F%
CRF_W	90.5	55.8	68.0	94.3	60.7	73.1
CRF_M	81.9	61.3	70.1	91.9	63.3	75.0

As illustrated in TABLE X, the morpheme-based chunker overall outperforms the word-based chunker. The may be due to the fact that it is more straightforward to explore morphological features for NER under a morpheme-based framework than under a word-based framework.

Our second experiment is aiming at investigating the effects of lexical features and multi-level entity morphological features on nested NER. In this experiment, we take CRF_M in the first experiment as a baseline, and additionally introduce the lexical features (referred as CRF_ML) and the multi-level prefix and suffix morpheme features (referred to as CRF_MM), and then evaluate the related outputs, respectively. The experimental results are presented in TABLE XI and TABLE XII.

TABLE XI. Performance For ORG And LOC In The Second Experiment

System	Type	Nested NEs			NEs		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
CRF_M	ORG	85.0	65.1	73.7	89.0	67.2	76.6
	LOC	64.6	42.9	51.5	91.7	65.5	76.4
CRF_ML	ORG	86.2	74.5	79.9	90.6	75.2	82.2
	LOC	81.1	59.5	68.6	92.1	93.1	92.7
CRF_MM	ORG	84.9	84.3	84.6	89.1	84.3	86.7
	LOC	86.5	68.4	76.4	94.7	94.5	94.6

As can be observed from TABLE XI, the performance for nested NER increases with more features introduced. Take nested organization name recognition for example. the F-score is 73.7% for the baseline system. The number increase to 79.9% after using lexical features, and further to 84.4% after the introduction of multi-level entity morphological information. Similar trends can be observed with regards to nested location name recognition. This shows in a sense that both lexical features and entity morphological information are equally important cues for nested NER.

TABLE XII. Overall NER Performance For The Second Experiement

System	Nested NEs			NEs		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
CRF_M	81.9	61.3	70.1	91.9	63.3	75.0
CRF_ML	85.5	72.0	78.1	94.1	89.8	91.9
CRF_MM	85.1	81.7	83.3	94.4	92.9	93.8

TABLE XII presents the overall NER performance for the second experiments. As can be seen in this table, the overall F-score for all types of NEs can be improved from 75.0% to 91.9 after using lexical information, and further to 93.8%, illustrating not only the important roles of lexical information and entity-internal morphological features in nested NER, but also the significance of nested named entities in NER.

TABLE XIII. Distribution of NEs in IEER-99 and MET2 Test Data

NE Type	IEER-99		MET2	
	Number of NEs	Number of nested NEs	Number of NEs	Number of nested NEs
PER	504	—	170	—
LOC	962	60	585	84
ORG	483	236	318	224

To further demonstrate the effectiveness of our approach, we also conducted an open test on the IEER-99 and MET2 test data. TABLE XIII shows the distribution of named entities in the two corpora.

TABLE XIV. Results of nested named entities and total named entities on realistic corpus of IEEE and HKU-MET2

Corpus	Nested NEs			NEs		
	P%	R%	F%	P%	R%	F%
IEER-99	67.8	75.3	71.4	92.3	90.5	91.4
MET2	75.9	78.9	77.4	91.0	90.8	90.9

The results are listed in TABLE XIV. Our system performs better than the systems based on the lexicalized HMMs [1][2].

Although the proposed approach is effective for most nested NEs, it fails to yield correct results for some nested NEs that are embedded by other kinds of NEs, such as 中国人民银行江西金溪县支行 ‘*Jinxi branch of China People’s Bank in Jiangxi*’, or for some NEs that contain multiple prefixes and suffixes, such as 山东省石油公司加油站 ‘*Gas Station of Oil Company in Shandong Province*’.

VII. Conclusions

In this paper, we have presented a morpheme-based method to Chinese nested named entity recognition. In comparison with previous studies, our approach offers a straightforward framework for exploring more features, including entity-internal features and contextual features for nested NER. Our experimental results on different test set show that both lexical information and entity-internal morphological features are equally important for Chinese nested NER. To further enhance our system, in future we intend to exploit more complicated features such as deep entity formation patterns and structural features for recognizing complex nested NEs in real Chinese text.

Acknowledgment

This study was supported by National Natural Science Foundation of China under Grant No.60973081, 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] Guohong Fu, “Improving Chinese named entity recognition with lexical information. Proceedings of ICMLC’09, 2009, pp.12-15.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Yuejie Zhang, Zhiting Xu, and Xiangyang Xue, “Fusion of multiple features for Chinese named entity recognition based on maximum entropy model”. Computer Research and Development, 2008, pp.1004-1010.

[7] TzongHan Tsai, Chia-Wei Wu, and Wen-Lian Hsu, “Using maximum entropy to extract biomedical named entities without dictionaries”, Proceedings of ACL-IJCNLP’05, 2005, pp.268-273.

[8] Yefeng Wang, “Annotating and recognising named entities in clinical notes”. Proceedings of the ACL-IJCNLP’09 Student Research Workshop, 2009, pp.18-26.

[9] Jun She, and Xue-qing Zhang, “Musical named entity recognition method”, Proceedings of Association for Computer Applications, 2010, pp.2928-2931.

[10] 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.

[11] Jenny Rose Finkel, and Christopher D.Manning, “Nested named entity recognition”, Proceedings of EMNLP’09, 2009, pp.141-150.

[12] Beatrice Alex, Barry Haddow, and Claire Grover, “Recognising Nested Named Entities in Biomedical Text”, Proceedings of the Biological, Translational, and Clinical Language Processing, 2007, pp.65-72.

[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] Dingcheng Li, Karin Kipper-Schuler, and Guergana Savova, “Conditional random fields and support vector machines for disorder named entity recognition in clinical texts”, Proceedings of ACL’08, 2008, pp.94-95.