Semi-joint Labeling for Chinese Named Entity Recognition

Chia-Wei Wu[1], Richard Tzong-Han Tsai[2,*], and Wen-Lian Hsu[1,3]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2] Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, Taiwan
[3] Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan
cwwu@iis.sinica.edu.tw, thtsai@saturn.yzu.edu.tw,
hsu@iis.sinica.edu.tw

Abstract. Named entity recognition (NER) is an essential component of text mining applications. In Chinese sentences, words do not have delimiters; thus, incorporating word segmentation information into an NER model can improve its performance. Based on the framework of dynamic conditional random fields, we propose a novel labeling format, called *semi-joint labeling* which partially integrates word segmentation information and named entity tags for NER. The model enhances the interaction of segmentation tags and NER achieved by traditional approaches. Moreover, it allows us to consider interactions between multiple chains in a linear-chain model. We use data from the SIGHAN 2006 NER bakeoff to evaluate the proposed model. The experimental results demonstrate that our approach outperforms state-of-the-art systems.

Keywords: Named entity recognition, Chinese Word Segmentation.

1 Introduction

Named entity recognition (NER) is widely used in text mining applications. English NER achieves a high performance, but Chinese NER needs to be improved substantially. A named entity (NE) is a phrase that contains predefined names, such as person names, location names, and organization names. Named Entity Recognition (NER) is the process used to extract named entities in many applications, such as question answering systems, relation extraction, and social network analysis. Several conferences have been held to evaluate NER systems, for example, CONLL2002, CONLL2003, ACE (automatic context understanding), and SIGHAN 2006 NER Bakeoff. In many works, the NER task is formulated as a sequence labeling problem. Such problems have been discussed extensively in the past decade and several practical machine learning models have proposed, for example, the maximum entropy (ME) model[1], the hidden Markov model (HMM) [8], memory-based learning[5], support vector machines (SVM)[6] and conditional random fields (CRFs)[13].

Chinese NER is particularly difficult because of the word segmentation problem. Unlike English, Chinese sentences do not have spaces to separate words. Therefore,

[*] Corresponding author.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 107–116, 2008.

word segmentation information is important in many Chinese natural language applications. Depending on the way such information is incorporated, NER approaches can be classified as either character-based or word-based. In character-based approaches, segmentation information is used as features, whereas word-based methods use the output of the word segmentation tagger as the basic tagging unit. However, irrespective of the method used, the interactions between NER and word segmentation tags can not be considered jointly and dynamically.

One solution for handling multiple related sub-tasks like word segmentation and named entity recognition is to use joint learning methods, for example, jointly tagging parts-of-speech and noun phrase chucking using dynamic CRFs [13], incorporating features into different semantic levels using a log-linear joint model [3], and using a re-ranking model to jointly consider parsing and semantic role labeling [12]. These joint learning methods yield richer interactions between sub-tasks, which they consider dynamically.

In this paper, based on the concept of joint learning, we propose a novel Chinese NER tagging presentation, called the *semi-joint labeling* which partially integrates segmentation labels and named entity labels. The format can represent the interactions between the named entity and word segmentation states. It also facilitates dynamic consideration of NER and word segmentation states in a linear chain to alleviate the problem of potentially higher computation costs incurred by multiple layer tagging. Because it uses semi-joint tagging, the proposed system outperforms state-of-the-art systems.

The remainder of this paper is organized as follows. In Section 2, we introduce Chinese NER and word segmentation. In Section 3, we describe the proposed method. Then, in Section 4, we discuss the features of our system. Section 5 details the experiment results, and Section 6 contains our conclusions.

2 Chinese Word Segmentation and Named Entity Recognition

In this section, we introduce the Chinese word segmentation and named entity recognition task, and consider existing approaches that incorporate word segmentation information in NER models. In Table 1, the first row shows a series of Chinese characters with word segmentation and named entity labels. We list two segmentation tagging formats, BI and IE, in the next two rows. In the BI format, B denotes that *a character is at beginning of a word* and I denotes that *a character is in a word*. In the EI format, E denotes that *a character is at the end of a word* and I denotes *the inside character of a word.* In the named entity tagging format, a label is defined as a named entity type extended with a boundary tag. For example, B-LOC denotes that a character is at the beginning of a location name, while O denotes that the character is not part of a named entity. Word segmentation can provide valuable information for NER. For example, the boundaries between a word and a named entity can not cross or overlap. Previous works, such as Guo et. al.[4] and Chen et. al.[2], have shown that word segmentation information can improve NER performance.

There are two ways to incorporate word segmentation information into an NER model, namely, character-based approaches and word-based approaches. Unlike

Table 1. Examples of word-based and character-based tagging representation and their corresponding English phrases with NER tags

Character-based	**Characters**	俄	罗	斯	总	统	普	京	说
	BI-format Word Segmentation	B	Ｉ	I	B	Ｉ	B	I	B
	IE-format Word Segmentation	I	Ｉ	E	I	E	I	E	E
	Named entity labels	B-LOC	I-LOC	I-LOC	O	O	B-PER	I-PER	O
Word-based	**Words**	俄罗斯		总统		普京	说		
	Named entity labels	Ｂ-LOC		O		B-PER	O		
English	**Words**	Russian		president		Putin	says		
	Named entity labels	LOC		O		PER	O		

English NER, Chinese character-based NER uses characters as the basic tokens in the tagging process. Chen et. al.[2] and Wu et. al.[14] use a character-based approach in their NER models. The advantage of this approach is that it avoids the propagation of potential errors by the segmentation tagger. However, this approach does not consider the word segmentation information. One common approach employs a cascaded training and testing method that uses the output of the segmentation tagger as a feature in the NER model. For example, Guo et. al.[4] use word segmentation information as a feature in a character-based model.

Word-based NER uses words as the basic tokens. A number of systems, like those of Ji and Grishman [7] and Sun et. al. [11] use the word-based approach. In Figure 1, the first row of the word-based section shows an example of a phrase with word-based NER tags. Comparison with the first row of the English section shows that the NER tags of Chinese word-based and English word-based segmentation are the same. However, since word-based segmentation needs the output of a word segmentation tagger as the basic tagging token, propagated errors will be passed on to the NER model.

No matter whether the cascaded approach uses word segmentation information in character-based tagging or uses word-based tagging directly, the interactions between word segmentation and NER can be represented in is limited and can not be considered dynamically. Next, we introduce dynamic CRFs and the semi-joint labeling format used to represent more complex interactions.

3 Methods

3.1 Dynamic Conditional Random Fields

Dynamic conditional random fields (DCRF) [13] are generalizations of linear-chain conditional random fields (CRF) in which each time slot contains a set of state variables and edges. The form of a dynamic CRF can be written as follows:

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{t=1} \prod_{c \in C} \exp\left(\sum_{k} \lambda_k f_k \left(\boldsymbol{y}_{t,c}, \boldsymbol{x}, t \right) \right) \tag{1}$$

where $\boldsymbol{y}$ is a label sequence over observation sequence $\boldsymbol{x}$**;** c denotes a clique in a graph; λ_k and f_k are, respectively, the weights and feature function associated with the clique index k; t denotes a time slot; and Z is a normalization constant. By using different definitions of c, DCRFs can represent various interactions within a time slot. For example, if we define c as a combination of labels in multiple tagging layers, then $\boldsymbol{y}_{\mathrm{t,c}}$ denotes a joint label of multiple layers in time slot t. Therefore, we can identify rich interactions between word segmentation information and named entity recognition.

We use DCRFs to present a graphical model that considers the interactions of named entities and word segmentation tags in a multiple chain structure. In Figure 2a, the two chains correspond to the state sequences of named entities and word segmentation tags. Using DCRFs, we can represent this structure by the following equation:

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{t=1}^{T-1} \Omega\left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t+1,n}, \boldsymbol{x}, t \right) \prod_{t=1}^{T-1} \omega\left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t,s}, \boldsymbol{x}, t \right) \tag{2}$$

where $\boldsymbol{y}_{\mathrm{n}}$ denotes the named entity label sequence and $\boldsymbol{y}_{\mathrm{s}}$ denotes the segmentation label sequence; Ω denotes the function of the interactions between $\boldsymbol{y}_{n,t}$ and $\boldsymbol{y}_{n,t+1,}$ the labels of named entities in the adjacent time slot; and ω denotes the function of the interactions between the named entities and word segmentation labels in the same time slot. Based on the feature f_k and the parameter λ_k, Ω and ω can also be presented as:

$$\begin{aligned} \Omega\left(\boldsymbol{y}_{t,n}, \boldsymbol{x}, t \right) &= \exp\left(\sum_{k} \lambda_k f_k \left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t+1,n}, \boldsymbol{x}, t \right) \right) \\ \omega\left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t,s}, \boldsymbol{x}, t \right) &= \exp\left(\sum_{k} \lambda_k f_k \left(\boldsymbol{y}_{t,s,n}, \boldsymbol{x}, t \right) \right), \end{aligned} \tag{3}$$

Figure 1(b) is a three-chain structure in which the chains corresponding to the tagging sequence from the top to the bottom represent named entity segmentation, BI-format word segmentation, and EI-format segmentation, respectively. Using DCRFs, we can represent this structure by the following equation,

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{t=1}^{T-1} \Omega\left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t+1,n}, \boldsymbol{x}, t \right) \prod_{t=1}^{T-1} \omega\left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t,s}, \boldsymbol{x}, t \right) \prod_{t=1}^{T-1} \Psi\left(\boldsymbol{y}_{t,n}, \boldsymbol{y}_{t,s}, \boldsymbol{x}, t \right) \tag{4}$$

where Ψ denotes the interactions between the state sequence of named entity labels and the EI-format word segmentation labels.

3.2 Semi-joint Labels in Linear Conditional Random Fields

If DCRFs are used to represent complicated structures, such as multiple layers of tags, the number of cross-products of states between the layers will cause the inference space increases. For example, the cross-product space of the segmentation labels and named entity labels is twice as large as the original named entity label space. We propose a semi-joint model to reduce the inference spaces.

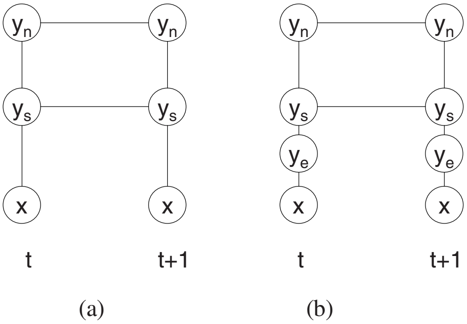

Fig. 1. The two tagging structures

A semi-joint label set partially integrates the original labels in different layers. Semi-joint labeled CRFs are linear-chain CRFs transformed from multiple chain CRFs by applying the semi-joint label set. Next, we define a semi-joint label set.
Let a semi-joint label set $\boldsymbol{q}$ be { $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, … ,$\boldsymbol{q}_m$ } where $\boldsymbol{q}_k$ is a vector of the label set selected from the Cartesian product of the original label set. The selection rule can be decided manually or systematically For example, Table 1 shows a Chinese phrase with word segmentation tags and named entity tags that are integrated by semi-joint labeling tags. The second column shows the phrase's corresponding English translation. Each character has segmentation tags in two formats, a named entity tag and two semi-joint labeling tags, as shown in the last two columns Note that semi-joint labeling only integrates a segmentation tag with a named entity tag if the named entity is "O". Other named entity tags will be reserved. The number of distinct tags in the semi-joint labeling format is only one more than in the original named entity format. Even in semi-joint labeling II, which integrates two kinds of segmentation format, there are only three more distinct tags than in the original format.

In this example, we also find that integrating the word segmentation tag with the named entity tag "O" can provide boundary information, which can not be derived from the original tag "O". For example, the named entity tag sequence of the first three characters is (B-LOC, I-LOC, I-LOC); hence, the next tag can not be I-O it would be against the constraint that the word boundary can not cross or overlap the named entity boundary. This constraint helps us rule out impossible inference paths and thereby improve the precision of named entity diction boundaries.

Next, we replace $\boldsymbol{y}$ with the semi-joint labels q_{kj} in linear-CRFs, as shown in Equation 5.

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_k \lambda_k f_k \left(q_t, q_{t+1}, \boldsymbol{x}, t \right) \right) \quad (5)$$

By combining the BI and IE formats, we can identify more significant interactions, such as constraints, when considering the transition of labels. For example, in the last row of Table 2, the tag before B-PER can only be I-E-O and B-E-O; otherwise, it would be against the cross-overlap constraint on words and named entities.

Table 1. An example of a Chinese phrase with different tagging representation

Token	Meaning	Segmentation Tag		NER	Semi-joint Label	Semi-joint Label II
		BI format	IE format			
俄	Russian	B	I	B-LOC	B-LOC	B-LOC
罗		I	I	I-LOC	I-LOC	I-LOC
斯		I	E	I-LOC	I-LOC	I-LOC
总	president	B	I	O	B-O	B-I-O
统		I	E	O	I-O	I-E-O
普	Putin	B	I	B-PER	B-PER	B-PER
京		I	E	I-PER	I-PER	I-PER
说	says	B	E	O	B-O	B-E-O

4 Features

4.1 Basic Features

The basic features of our NER model are:

State features

- C_n (n = -2, -1, 0, 1, 2)
- C_nC_{n+1} (n = -2, -1, 0, 1)

Transition features

- C_n (n = -1, 0, 1) ,

where C denotes a character and n denotes its position. For example, C_0 denotes the current character and C_nC_{n+1} denotes its bi-gram feature, which is a combination of the previous character and the current character. A state feature is a feature that only corresponds to the current label, whereas transition features relate the previous and current labels.

4.2 Knowledge Features

Knowledge features are semantic hints that help an NER model identify more named entities. Several Chinese NER models use knowledge features; for example, Youzheng Wu [14] collects named entity lists to identify more named entities and thereby resolves the data sparseness problem in Chinese NE.

To compare our system with other approaches, we observe the closed task rules, which do not allow the use of external resources. Therefore, we only generate knowledge lists from the training corpus. For example, we compile the surname list from the tagged person names in the corpus. The knowledge feature types are listed in the Table 3. Since the features are generated automatically, we filter out those that occur less than twice [2], [14]. The table also shows the number of each distinct feature that we obtain from the training corpus.

Next, we consider how we represent knowledge in feature functions. If a character is included in a list of knowledge features, the feature's value is set at 1; otherwise, it is set at 0.

Table 2. The list of knowledge feature types

Feature type	**Description**	**#**
Person surname	The first character of a person name	678
Person name	The characters of a person name, except the first character.	1374
Previous characters of a person name	The previous single character of a person name and the previous two characters of the name	1847
The characters after a person name	The first character after a person name and the first bigram characters after the name	2467
Location name	The characters of a location name	778
Organization name	The characters of an organization name	823
Suffix characters of an organization name	The last two characters of an organization name	417

5 Experiment

In this section, we describe the experimental data, introduce the parameters used in the CRF model, and detail the experiment results.

5.1 Data Source

To evaluate our methods, we use the City University of Hong Kong (CityU) Chinese corpus from SIGHAN 2006 [10], the Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics. We choose the CityU corpus because it provides both segmentation tags and named entity tags. The corpus contains 1.6M words and 76K named entities in the training part, and 220K words and 23K named entities in the test part. It also contains three named entity types: person names, organization names, and location names.

5.2 Settings

We use CRF++[1] to implement our CRF models. The parameters we use in CRF++ are f, the cut-off threshold, which is set to 1; and c, the C value that prevents over fitting, which is set to 3. The maximum number of training iterations is 1000, and the training environment is Windows Server 2003 with an AMD 2.39GHz CPU and a 10 Gigabyte RAM.

[1] Information about CRF++ can be found at http://chasen.org/~taku/software/CRF++/

5.3 Results and Discussion

Table 4 shows the results achieved by the three NER models. Each row shows the performance of an NER model for three types of NE with specific tagging formats, as well as the model's overall performance. The models are evaluated on the full test set (220K words and 23K NEs) of the CityU corpus. BIO uses the traditional format, i.e., a named entity type extended with a boundary, while the Semi-Joint labeling and Semi-Joint II labeling formats use the methods proposed in Section 3.2. Basic and knowledge features are included in all three models. The only difference is that the models using semi-joint formats do not include word segmentation features. By contrast, in the model that uses the BIO format, the output of a segmentation tagger includes word segmentation features. The results show that the Semi-joint format outperforms the BIO format for all three NE types with an overall F-score of approximately 1.41%. Meanwhile, the Semi-joint II format outperforms the Semi-joint format with an overall F-score of approximately 0.24%.

Table 3. The results of the BIO, semi-joint, and semi-joint II formats

		precision	**recall**	**F-score**
Baseline	**PER**	91.42	85.35	88.28
System	**ORG**	90.31	77.19	83.24
	LOC	92.09	91.85	91.97
	Overall	**91.49**	**86.29**	**88.82**
Semi-Joint	**PER**	93.50	88.25	90.80
Labeling A	**ORG**	90.43	78.14	83.89
	LOC	92.35	92.74	92.55
	Overall	**92.27**	**87.83**	**89.99**
Semi-Joint	**PER**	93.51	89.32	91.37
Labeling B	**ORG**	90.05	77.81	83.48
	LOC	92.43	93.32	92.87
	Overall	**92.23**	**88.31**	**90.23**

Since the proposed semi-joint labeling method integrates word segmentation with an NER model, and word segmentation can help detect the boundaries of named entities, it is worth discussing changes in the error rates due to named entity boundary detection. We define a boundary error as a named entity is identified and their lengths are different with the correct ones. Each row in Table 5 shows the reduced boundary error rate achieved by using semi-joint labeling. The error rate is computed by dividing the number of named entities with boundary errors in the semi-joint labeling method by those in the baseline system. We observe that semi-joint labeling reduces boundary errors, especially the semi-joint labeling II, which integrates two word segmentation formats.

Next we consider different types of boundary error. Suppose the boundary of a named entity in a sentence is $<i, j>$ where i is start position and j is end position. We define boundary detection error type I as $i_{\text{guessed entity}} = i_{\text{correct entity}}$ and $j_{\text{guessed entity}} \neq j_{\text{correct entity}}$, and boundary error type II as $i_{\text{guessed entity}} \neq i_{\text{correct entity}}$ and $j_{\text{guessed entity}} = j_{\text{correct entity}}$. Semi-joint B in boundary error type II is more significant than semi-joint A. We infer that, with the IE-format word segmentation information, the beginning character of a named entity can be identified more easily by the "E" label, which refers to the end of a word.

Table 4. Reduced boundary error rates achieved by the two semi-joint methods

	Reduced Boundary Error Rates
Semi-Joint A	3.90 %
Semi-Joint B	8.62 %

Table 5. Reduced error rates of boundary errors achieved by the two semi-joint methods

	Reduced Error Rates of Boundary Error Type I	**Reduced Error Rates of Boundary Error Type II**
Semi-Joint A	8.33 %	2.33 %
Semi-Joint B	15.87 %	18.14 %

We list the performance of the top five teams at the SIGAHN NER bakeoff for the CityU corpus. The performance of the proposed model is better than the top one in 1.2% F-scores. The major difference between our results and those of NII is that the latter approach uses word segmentation information as features, while we partially join word segmentation tags with named entity tags.

Table 6. The performance of the top five teams for the CityU corpus at the SIGHAN 2006 NER bakeoff

Team Name	**Precision**	**Recall**	**F-score**
Our Results	92.23	88.31	90.23
NII	91.43	86.76	89.03
Yahoo!	92.66	84.75	88.53
Chinese Academy of Sciences	92.76	81.81	86.94
Alias-i, Inc.	86.90	84.17	85.51

6 Conclusion

We propose a semi-joint tagging format that partially combines word segmentation and named entity recognition labels. The format allows us to consider the interactions between multiple labeling layers in a linear-chain CRF model. To evaluate our model, we use the CityU corpus of SIGHAN 2006 NER bakeoff. The model based on semi-joint labeling outperforms the model that uses word segmentation tags as features, with an overall F-score of approximately 1.41%. Because of the novel labeling format, the proposed model outperforms the top one system by about 1.2% in terms of the F-score.

In our future work, we will explore other possible interactions between word segmentation information and NER. We also plan to apply our method to other applications that would be improved by incorporating word segmentation information.

Acknowledgements

We are grateful for the support of thematic program of Academia Sinica under Grant AS 95ASIA02 and the Research Center for Humanities and Social Sciences, Academia Sinica.

Reference

1. Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition. New York University, New York (1999)
2. Chen, W., Zhang, Y., Isahara, H.: Chinese Named Entity Recognition with Conditional Random Fields. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 118–121 (2006)
3. Duh, K.: A Joint Model for Semantic Role Labeling. In: Proceedings of the 9th Conference on Computational Natural Language Learning, pp. 173–176 (2005)
4. Guo, H., Jiang, J., Hu, G., Zhang, T.: Chinese Named Entity Recognition Based on Multilevel Linguistic Features. In: International Joint Conference on Natural Language Processing, pp. 90–99 (2004)
5. Hendrickx, I., Bosch, A.v.d.: Memory-based One-step Named-entity Recognition: Effects of Seed List Features, Classifier Stacking, and Unannotated Data. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL, pp. 176–179 (2003)
6. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th international conference on Computational linguistics, pp. 1–7 (2002)
7. Ji, H., Grishman, R.: Improving Name Tagging by Reference Resolution and Relation Detection. In: Proceedings of the 43rd Annual Meeting of the ACL, pp. 411–418 (2005)
8. Klein, D., Smarr, J., Nguyen, H., Manning, C.D.: Named Entity Recognition with Character-Level Models. In: Conference on Computational Natural Language Learning, pp. 180–183 (2003)
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning, pp. 282–289 (2001)
10. Levow, G.-A.: The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 108–117 (2006)
11. Sun, J., Gao, J., Zhang, L., Zhou, M., Huang, C.: Chinese named entity identification using class-based language model. In: Proceedings of the 19th international conference on Computational linguistics, pp. 1–7 (2002)
12. Sutton, C., McCallum, A.: Composition of Conditional Random Fields for Transfer Learning. In: Proceedings of Human Language Technologies / Empirical Methods in Natural Language Processing, pp. 748–754 (2005)
13. Sutton, C., Rohanimanesh, K., McCallum, A.: Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. In: Proceedings of the Twenty-First International Conference on Machine Learning, pp. 99–107 (2004)
14. Wu, Y., Zhao, J., Xu, B.: Chinese Named Entity Recognition Combining Statistical Model wih Human Knowledge. In: Dignum, F.P.M. (ed.) ACL 2003. LNCS (LNAI), vol. 2922, Springer, Heidelberg (2004)