# Using Non-Local Features to Improve Named Entity Recognition Recall*

Xinnian Mao[1], Wei Xu[1], Yuan Dong[1, 2], Saike He[2], and Haila Wang[1]
[1]France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
{xinnian.mao, wxu.ext, yuan.dong,haila.wang}@orange-ftgroup.com
[2]University of Posts and Telecommunications, Beijing, 100876, P.R.China
yuandong@bupt.edu.cn; hsk000@gmail.com

**Abstract.** Named Entity Recognition (NER) is always limited by its lower recall resulting from the asymmetric data distribution where the *NONE* class dominates the entity classes. This paper presents an approach that exploits non-local information to improve the NER recall. Several kinds of non-local features encoding entity token occurrence, entity boundary and entity class are explored under Conditional Random Fields (CRFs) framework. Experiments on SIGHAN 2006 MSRA (CityU) corpus indicate that non-local features can effectively enhance the recall of the state-of-the-art NER systems. Incorporating the non-local features into the NER systems using local features alone, our best system achieves a 23.56% (25.26%) relative error reduction on the recall and 17.10% (11.36%) relative error reduction on the F1 score; the improved F1 score 89.38% (90.09%) is significantly superior to the best NER system with F1 of 86.51% (89.03%) participated in the closed track.

**Keywords:** Named Entity Recognition, Non-local Feature, Conditional Random Field

## 1. Introduction

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify predefined entities, such as names of persons, locations, organizations, etc. in unstructured texts. It is the fundamental step to many natural language processing applications, like Information Extraction (IE), Information Retrieval (IR) and Question Answering (QA). Most empirical approaches currently employed in NER task make decision only on local context for extract inference, which is based on the data independent assumption (Krishnan and Manning, 2006). But often this assumption does not hold because non-local dependencies are prevalent in natural language (including the NER task). How to utilize the non-local dependencies effectively is a key issue in NER task. Unfortunately, few researches have been devoted to this issue, existing works mainly focus on using the non-local information for further improving NER label consistency.

There are two methods to use non-local information. One is to add additional edges to graphical model structure to represent the distant dependencies and the other is to encode the non-locality with non-local features. However, in the first approach, heuristic rules are used to find the dependencies (Bunescu and Mooney, 2004; Sutton and McCallum, 2004) or penalties for label inconsistency are required to handset ad-hoc (Finkel et al., 2005). Furthermore, high computational cost is spent for approximate inference. In order to establish the long dependencies easily and overcome the disadvantage of the approximate inference, Krishnan and Manning (2006) propose a two-stage approach using Conditional Random Fields (CRFs) with extract inference. They represent the non-locality with non-local features, and extract the non-local features from the output of the first stage CRF using local context alone; then they incorporate the non-local features into the second CRF. But the features in this approach are

only used to improve label consistency.

To our best knowledge, up to now, non-local information has not been explored to improve NER recall in previous researches; on the other hand, NER is always impaired by its lower recall due to the imbalanced distribution where the *NONE* class dominates the entity classes. Classifiers built on such data typically have a higher precision and a lower recall and tend to overproduce the *NONE* class (Kambhatla, 2006). In this paper, we employ non-local information to recall the missed entities. Similar to Krishnan and Manning (2006), we also encode non-local information with features and apply the simple two-stage architecture. Different from their work for improve label consistency, their features are activated on the recognized entities coming from the first CRF, the non-local features we design are used to recall more missed entities which are seen in the training data or unseen entities but some of their occurrences being recognized correctly in the first stage, our features are fired on the raw token sequence directly with forward maximum match. Compared to their non-local information extracted from training data with 10-fold cross-validation, our non-local information is extracted from the training date directly; our approach obtaining the non-local features is simpler. Moreover, we design different non-local features encoding different useful information for NER two subtasks: entity boundary detection and entity semantic classification. Our features are also inspired by Wong and Ng (2007). They extract entity majority type features from unlabelled data with an initial maximum entropy classifier. Our approach is validated on the third International Chinese language processing bakeoff (SIGHAN 2006) MSRA and CityU NER closed track, the experimental results show that non-local features can significantly improve the recall of the state-of-the-art NER system using local context alone.

The remainder of the paper is structured as follows. In Section 2, we introduce the first stage CRF with local features alone; then we describe the second stage CRF using non-local features we design in Section 3. We demonstrate the experiments in Section 4 and we conclude the paper in Section 5.

## 2. Our Baseline NER System

To validate the effectiveness of our approach of exploiting non-local features, we need to establish a baseline with state-of-the-art performance using local context alone. Similar to (Krishnan and Manning, 2006), we employ two-stage architecture under conditional random fields (CRFs) framework. In the first stage, we build the baseline with local features only, and then we build the second NER system with non-local features. We will introduce them step by step.

### 2.1. Conditional random fields

We regard the NER task as a sequence labeling problem and apply Conditional Random Fields (Lafferty et al., 2001; Sha and Pereira, 2003) since it represents the state of the art in sequence modeling and has also been very effective at NER task. It is undirected graph established on G = (V, E), where V is the set of random variables $Y = \{Y_i | 1 \le i \le n\}$ for each the n tokens in an input sequence and $E = \{(Y_{i-1}, Y_i) | 1 \le i \le n\}$ is the set of (n − 1) edges forming a linear chain. Following Lafferty et al. (2001), the conditional probability of the state sequence (s1, s2…sn) given the input sequence (o1, o2…on) is computed as follows:

$$P_\Lambda(s \mid o) = \frac{1}{Z_o} \prod_{c \in C(s,o)} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (1)$$

Where $f_k$ is an arbitrary feature function; and $\lambda_k$ is the weight for the feature function; it can be optimized through iterative algorithms like GIS (Darroch and Ratcliff, 1972) and IIS (Della Pietra et al., 1997). However recent research y been shown that quasi-Newton methods, such as L-BFGS, are significantly more efficient (Byrd et al., 1994; Malouf, 2002; Sha and Pereira, 2003).

### 2.2. Local features

The fist stage CRF labels for token directly depends on the labels corresponding the previous and next token, namely $C_{-2}$, $C_{-1}$, $C_0$, $C_{-1}$, $C_2$, $C_{-2}C_{-1}$, $C_{-1}C_0$, $C_0C_{-1}$, $C_1C_2$, and $C_{-1}C_1$, where $C_0$ is the current character, $C_1$ the next character, $C_2$ the second character after $C_0$, $C_{-1}$ the character preceding $C_0$, and $C_{-2}$ the second character before $C_0$. In addition, the first CRF used the tag bi-gram feature. Although these local features are simple, they give us state-of-the-art baseline using local information alone as described in Section 4.

### 2.3. Low recall in NER task

As Kambhatla (2006) points out that NER system typically have a higher precision and a lower recall and tends to overproduce the NONE class because the *NONE* class dominates all other classes in the task. In natural language, different sentences contain different useful contextual information; the missed entities are happened when their context surroundings are not indicative enough for the statistical-based approaches (including the CRFs) to make a correct decision. When we analyze these missed occurrences of the missed entities further, we can put them into three groups. The first is the seen entities in the training data; the second is the unseen occurrences, but some other occurrences of the entities have been correctly recognized in certain indicative context surroundings. The third is the unseen occurrences with no any occurrences recognized correctly. In NER task, considering influences between extractions can be very useful, if the context surrounding one occurrence of a token sequence is very indicative of it being an entity, then this should also influence the tagging of another occurrence of the same token sequence in a different context that is not indicative of entity (Bunescu and Mooney, 2004). So if we consider the non-local dependencies between the same entities, some of these missed occurrences will be recognized correctly. We will describe how to capture the non-locality to recall more missed entities in Section 3.

## 3. Recalling Missed Entities with Non-local Features

In natural language, difference sentences contain different useful context information; the missed entities happen when their context surroundings are not indicative enough for the first stage CRF to make correct decisions. If the context surrounding one occurrence of a token sequence is very indicative of it being an entity, then this should also influence the labeling of another occurrence of the same token sequence in a different context that is not indicative of entity (Bunescu and Mooney, 2004). So considering the non-local dependencies between the same entities can be very useful, if these non-local dependencies are incorporated into the CRF model, some of the missed entities will be recalled correctly.

### 3.1. Flow chart using non-local features

Figure 1 shows the flow using non-local features in two-stage architecture under CRFs framework. The first CRF is trained with local features alone as baseline (described in Section 2), and then we test the testing data with the first CRF and get the entities plus their type from the output. The second CRF utilizes the non-local features derived from the entity list which is merged by the output of the first CRF from the testing data and the entities extracted directly from the training data. To provide flexible and general conclusion, we only use non-local information found in labeled training data and test data rather than external knowledge sources, such as post-of-speech, gazetteers, external lexica and etc.

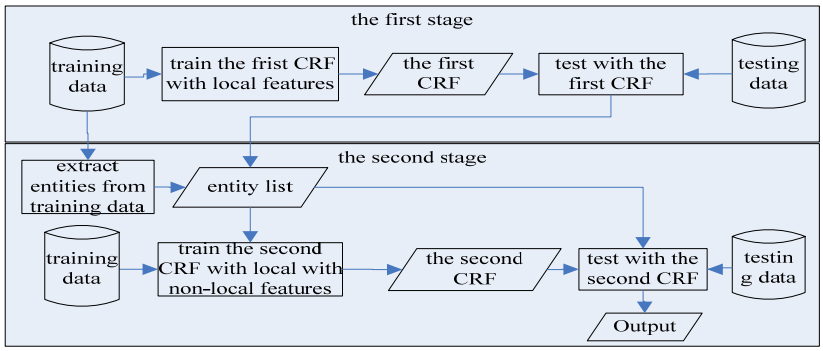

Figure 1. The flow using non-local features in two-stage architecture

### 3.2. Four kinds of non-local features

We design four kinds of non-local features which encode different useful information for the two NER subtasks, i.e. entity boundary detection and entity semantic classification, the non-local features are fired on the token sequences if they are matched with certain entity in the entity list in forward maximum matching (FMM) way. I will describe them one by one as follows.

**Entity-occurrence features (F1)**: These refer to the occurrence information assigned to the token sequence which is matched with the entity list exactly. These features capture the dependencies between the identical candidate entities; which results in the same candidate entities of different occurrences can be recalled favorably.

**Token-position features (F2)**: These refer to the position information (start, middle and last) assigned to the token sequence which is matched with the entity list exactly. These features enable us to capture the dependencies between the identical candidate entities and their boundaries.

**Entity-majority features (F3)**: These refer to the majority label assigned to the token sequence which is matched with the entity list exactly. These features enable us to capture the dependencies between the identical entities and their classes, so that the same candidate entities of different occurrences can be recalled favorably, and their label consistencies can be considered too.

**Token-position & entity-majority features (F4)**: These features capture non-local information from F2 and F3 simultaneously. They take into account the entity boundary and semantic class information at the same time.

These non-local features are applied in English NER in one-step approach (Krishnan and Manning, 2006; Wong and Ng, 2007), they employ these features to improve entity consistence among their different occurrences. These features are assigned to token sequences that are matched exactly with the (entity, majority-type) list in forward maximum matching (FMM) way. During training or testing, when the CRFs tagger encounters a token sequence $C_1...C_n$ such that $(C_k...C_s)$ $(k \geq 1, s \leq n)$ is the longest token sequence existing in the entity list; the correspondent features will be turned on to each token in $C_k....C_s$. For example, considering the following sentence: *我(wo)爱(ai)北(bei)京(jing)天(tian)安(an)门(men)(I love Beijing Tiananmen)*. If (*北京*, *Maj-LOC*), (*京*, *Maj-LOC*), and (*天安门*, *Maj-LOC*) are presented in the *(entity, majority type)* list, the features below will be turned on as table 1 shows.

Notice that the feature turned on for *京*is *E-Maj-LOC*, not *B-Maj-LOC*, because the longest matching sequence is *北京*. Different from (Krishnan and Manning, 2006; Wong and Ng, 2007), they only assign the majority type information, like *Maj-Loc,* to each token in matched candidates, boundary information like *B, I* and *E* is ignored, it is acceptable because they utilize these features only for English corpora, and the boundary information can be captured by the capitalization characteristics. But in Chinese NER, NED is more difficult than NEC, so we assign the boundary information, representing with *B*, *I* and *E*, to each token in the matched candidates. Please note that not all matching token sequences are true candidates. The false

candidates come from two aspects: the first is the boundaries are correct, but the occurrences are common words[1]; the second errors come from FMM, so the features are soft constraints.

Table 1. Example for Token-Majority-Type features

| *Token* | *Entity-Majority-Type Feature* |
|---|---|
| *我* | - |
| *爱* | - |
| *北* | *B-Maj-LOC* |
| *京* | *E-Maj-LOC* |
| *天* | *B-Maj-LOC* |
| *安* | *I-Maj-LOC* |
| *门* | *E-Maj-LOC* |

## 4. Experiments

### 4.1. Corpus analysis

Our investigation is based on the MSRA and CityU datasets from the NER closed track of the third International Chinese language processing bakeoff (SIGHAN 2006) (Levow, 2006); its goal is to perform NER on three entity classes: PERSON, LOCATION and ORGANIZATION. We give up the LDC corpus because it is initially designed for ACE Evaluation and the definition of named entity is different from traditional definition. The named entities in SIGHAN training data sets are labeled in IOB-2 format, we covert the corpus to OBIE as a pre-processing, because some existing work and our experiments show that OBIE scheme outperforms other formats when applying machine learning to NER. In OBIE format, tokens outside of entities are tagged with *O* (*NONE* class), while the first token in an entity is tagged with B-k to begin class k, the token inside the entity is tagged with I-k and the end token in the entity is tagged with E-k; single-token entity is labeled as B-k.

General information for each dataset appears in Table 2. It also summarizes the statistic information of seen and unseen entities in the test sets. A seen named entity in test set means that it exists in its correspondent training data set. From the table, we can find that the proportion of seen entities is very high. 71.86% of named entities in MSRA test data can be found in MSRA training data, while 73.53% for CityU corpus. In fact, most of named entities may appear frequently in our generally lives. To make use of existing named entities in training data is crucial to improving capability to capture seen entities and thereby unseen entities, since many models consider the possibilities of labels in context. We also see an interesting phenomenon in MSRA corpus that many named entities are consecutive without punctuations, especially the person names. Particularly, in MSRA testing data, nearly 20% named entities appear consecutively. It brings great difficulties for NER system to capture such entities separately.

**Table 2**. Corpus overall statistics

| | #(W) | #(E) | #(C) | #(S) |
|---|---|---|---|---|
| MSRA (Training) | 1.3M | 75060 | 10.93% | --- |
| MSRA(Testing) | 100k | 6190 | 19.68% | 71.86% |
| CityU (Training) | 1.6M | 112347 | 10.13% | --- |
| CityU (Testing) | 220k | 16407 | 9.60% | 73.53% |

#(W): the size of words; (E): the size of entities;#(C): the proportion of the consecutive entities; #(S): the proportion of the seen entities

### 4.2. Problems of NER with only local information

[1] For the string*两岸*, when it refers to Mainland and Taiwan, it is an entity, when it refers to the bank of rivers, it is a common word.

Table 3 displays the performance on MSRA and CityU NER closed track. The F0 row lists the precision, recall and F-measure (β=1) got by the first CRF (described in Section 2) using local features alone. The score makes the first CRF rank the top position on the MSRA and the second on the CityU in SIGHAN bakeoff (Levow, 2006)[2]. It shows that our baseline has achieved the state-of-the-art performance. However, comparing the recall with the precision on each dataset, we find that the performance is impaired by the relatively low recall. To investigate the causes of this problem, we analyze the missed entities further. We categorize them into two classes, seen and unseen in training data. Five kinds of statistic information are collected and listed in F0 column in table 4. (1) The number of different missed named entities; (2) The times of missing occurrences; (3) The number of different missed named entities which are detected correctly at least once; (4) The times missing occurrences under the case of (3).
From (1) and (2) measurements of the seen entities in the F0 column in table 4, we find that there are many seen named enmities are missed. Though identifying unseen named entities is more difficult than seen named entities, the boldfaced number indicates that about 10% (24 of 254) for MSRA and 23% (111 of 476) for CityU of unseen and missed named entities have been labeled out correctly for at least once. The difficulty to capture unseen named entities in training data is because of the nature of machine learning techniques. However, the statistical results in Table 4 show there is a great potential (200+48)/(200+330)=47% for MSRA and (384+396)/(384+1144)=51% for CityU, to improve recall by enhancing the capture of seen named entities and making use of labeled outputs from test data to capture more unseen named entities. What is more, performance can be improved further when more named entities are labeled correctly, because many models, such as CRF, assign labels according to the possibilities of whole sequence.

### 4.3. Influence of using non-local features in NER

After we feed the non-local features (described in Section 3) to the second CRF, we test it on the testing data of MSRA and CityU again. Table 3 lists the performance got by each kind of feature configurations. F0 means the first CRF (baseline) using local features alone, and the F0+Fi (i=1, 2, 3, 4) means the second CRF using local features (F0) as well as the non-local features Fi. From the table 3, we can conclude that exploiting non-local information is a good choice to recall more missed entities. Comparing with the baseline using only local context, the recalls of NER systems are improved after taking non-local information into account by -0.34%~3.76% on MSRA, 2.92%~3.68% on CityU. And the overall F-measures increase by -0.54%~2.19% and 0.72%~1.27% on MSRA and CityU each. The MSRA performance got by F0+F1 decrease slightly because there are many consecutive entities in the testing data. Since F1 does not encode boundary and class information, more entity tokens are recalled, but their boundaries or classes are wrong. After we implement a post-processing step with person name list extracted from the MSRA training data to separate the consecutive candidate entities, the performance lists with F0+F1 (PP) increases. The performance difference among F1, F2, F3 and F4 are mainly because they encode different useful non-local information as described in Section 3.2. For F1, it only encode whether a token sequence is an entity. No boundary and class are considered which are represented in F2 and F3 respectively, so F2 and F3 both achieve high performance than F0, and F4 consider both boundary and class simultaneously, so it is the best choice of exploiting non-local information to improve NER recall. We can not compare between F2 and F3 directly because boundary detection and semantic class classification are the two different sub-tasks in NER.

The performance difference between the performance on CityU and MSRA come from two folds. One is because CityU testing data contains more seen entities than that of MSRA since the seen entities can be captured easily by the non-local features. The other is because MSRA data sets contain much more consecutive named entities than CityU. Since NER with non-local information prefers to dig out more and thereby longer named entities, it may tend to label more

[2] The best F1-score on MSRA and CityU is 86.51% and 89.03% respectively.

continuous named entities as a single named entities and introduce more errors damaging both in recall and precision.

**Table 3**. NER performance on MSRA and CityU

| Corpus | System | P | R | F |
|---|---|---|---|---|
| MSRA | F0 | 90.58 | ***84.04*** | **87.19** |
| | F0+F1 | 89.81 | 83.70 | 86.65 |
| | F0+F1(PP) | 89.40 | 85.46 | 87.39 |
| | F0+F2 | 89.73 | 85.96 | 87.81 |
| | F0+F3 | 90.58 | 87.16 | 88.84 |
| | F0+F4 | 91.01 | ***87.80*** | **89.38** |
| CityU | F0 | 92.48 | ***85.43*** | **88.82** |
| | F0+F1 | 90.73 | 88.35 | 89.53 |
| | F0+F2 | 90.96 | 88.83 | 89.88 |
| | F0+F3 | 90.90 | 88.65 | 89.76 |
| | F0+F4 | 91.09 | ***89.11*** | **90.09** |

Then, we investigate the situation of missed seen and missed unseen named entity in NER with non-local information by filling the Table 4. F0 is the first CRF (baseline) using local features alone, and the F0+Fi (i=1, 2, 3, 4) means the second CRF using local features (F0) as well as the non-local features Fi. The four same measurements are used as described in Section 4.2. Compared with the numbers in F0 column, significant reduction of missing of seen entities is achieved by adding non-local features. What is more, the hit of unseen entities is also increased as we predicted in previous analysis.

**Table 4**. Analysis of missed named entities with non-local information

| Corpus | | F0 | F0+F1(PP)[3] | F0+F2 | F0+F3 | F0+F4 |
|---|---|---|---|---|---|---|
| Seen (MSRA) | 1 | 109 | 28 | 26 | 33 | 29 |
| | 2 | 200 | 158 | 74 | 83 | 75 |
| | 3 | ***45*** | *13* | *14* | *15* | *17* |
| | 4 | ***126*** | *86* | *45* | *45* | *47* |
| Unseen (MSRA) | 1 | 254 | 198 | 202 | 229 | 224 |
| | 2 | 330 | 484 | 246 | 275 | 270 |
| | 3 | *24* | *1* | *0* | *0* | *0* |
| | 4 | *48* | *2* | *0* | *0* | *0* |
| Seen (CityU) | 1 | 216 | 87 | 87 | 88 | 86 |
| | 2 | 384 | 149 | 149 | 152 | 148 |
| | 3 | ***118*** | 51 | 55 | 56 | 54 |
| | 4 | ***243*** | 94 | 105 | 109 | 119 |
| Unseen (CityU) | 1 | 476 | 356 | 348 | 355 | 350 |
| | 2 | 1144 | 704 | 696 | 702 | 693 |
| | 3 | 111 | 12 | 6 | 9 | 5 |
| | 4 | 396 | 17 | 12 | 15 | 11 |

## 5. Conclusions and Future Work

In this paper, we propose an approach of exploiting non-local information to improve NER recall. To our best knowledge, our work is the first attempt to utilize non-local information to improve NER recall, our work demonstrates that non-local information are effective to recall the missed entities which are seen in training data or unseen but some occurrences of these unseen

[3] We do not perform post-processing step on CityU testing data

entities have been recognized correctly with local context alone. We also compare the different kinds of non-local features which fit to different NER sub-tasks and find that non-local feature considering the boundary and class information simultaneously is the best. Our approach is language independent, due to lack of annotated corpora of other languages, the experiments have only been conducted on Chinese corpora, and related experiments on other languages can be done in the future.

## References


R. Bunescu and R. J. Mooney. 2004. Collective information extraction with relational Markov networks. *Proceedings of the 42nd ACL*, pp. 439–446.

R.H. Byrd, J. Nocedal and R.B. Schnabel. 1994. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, (63):129-156.

J.N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43 (5):1470-1480.

J. Finkel, T. Grenager, and C. D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 42nd ACL*, pp. 363–370.

N. Kambhatla. 2006. Minority Vote: At-Least-N Voting Improves Recall for Extracting Relations. *Proceeding of the 44th ACL*, pp. 460–466.

V. Krishnan and C. D Manning. 2006. An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. *Proceedings of the 44th ACL*, pp. 1121–1128.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th ICML*, pp. 282–289. Morgan Kaufmann, San Francisco, CA

G. Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. *Proceedings of SIGHAN-2006*, pp. 108-117. Sydney, Australia.

R. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. *Proc. of CoNLL-2002*, 49-55. Taipei, Taiwan.

S.D. Pietra, V. Della Pietra, and J. Lafferty, 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380-393, 1997.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. *Proc. of HLT/NAACL-2003*, 213-220. Edmonton, Canada.

C. Sutton and A. McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. *In ICML Workshop on Statistical Relational Learning and Its connections to Other Fields*.

Y. CH. Wong and H. T. Ng. 2007. One class per named entity: exploiting unlabeled text for named entity recognition. *Proc. of IJCAI-2007*. 1763-1768. India.