# A Hybrid Generative/Discriminative Model for Rapid Prototyping of Domain-Specific Named Entity Recognition

Suzushi Tomori[1], Yugo Murawaki[2] and Shinsuke Mori[3]

[1,2]Graduate School of Informatics, Kyoto University
[3]Academic Center for Computing and Media Studies, Kyoto University
[1]tomori.suzushi.72e@st.kyoto-u.ac.jp
[2]murawaki@i.kyoto-u.ac.jp
[3]forest@i.kyoto-u.ac.jp

**Abstract.** We propose PYHSCRF, a novel tagger for domain-specific named entity recognition that only requires a few seed terms, in addition to unannotated corpora, and thus permits the iterative and incremental design of named entity (NE) classes for new domains. The proposed model is a hybrid of a generative model named PYHSMM and a semi-Markov CRF-based discriminative model, which play complementary roles in generalizing seed terms and in distinguishing between NE chunks and non-NE words. It also allows a smooth transition to full-scale annotation because the discriminative model makes effective use of annotated data when available. Experiments involving two languages and three domains demonstrate that the proposed method outperforms baselines.

## 1 Introduction

Named entity recognition (NER) is the task of extracting named entity (NE) chunks from texts and classifying them into predefined classes. It has a wide range of NLP applications such as information retrieval [1], relation extraction [2], and coreference resolution [3]. While the standard classes of NEs are PERSON, LOCATION, and ORGANIZATION among others, domain-specific NER with specialized classes has proven to be useful in downstream tasks [4].

A major challenge in developing a domain-specific NER system lies in the fact that a large amount of annotated data is needed to train high-performance systems, and even larger amounts are needed for neural models [5]. In many domains, however, domain-specific NE corpora are small in size or even non-existent because manual corpus annotation is costly and time-consuming. What is worse, domain-specific NE classes cannot be designed without specialized knowledge of the target domain, and even with expert knowledge, a trial-and-error process is inevitable, especially in the early stage of development.

In this paper, we propose PYHSCRF, a novel NE tagger that facilitates rapid prototyping of domain-specific NER. All we need to run the tagger is a few seed terms per NE class, in addition to an unannotated target domain corpus and a general domain corpus. Even with minimal supervision, it yields reasonable performance, allowing us to go back-and-forth between different NE definitions. It also enables a smooth transition to full scale annotation because it can straightforwardly incorporate labeled instances.

Regarding the technical aspects, the proposed tagger is a hybrid of a generative model and a discriminative model. The generative model called the Pitman-Yor hidden semi-Markov model (PYHSMM) [6] recognizes high-frequency word sequences as NE chunks and identifies their classes. The discriminative model, semi-Markov CRF (semiCRF) [7], initializes the learning process using the seed terms and generalizes to other NEs of the same classes. It also exploits labeled instances more powerfully when they are available. The two models are combined into one using a framework known as JESS-CM [8].

Generative and discriminative models have mutually complementary strengths. PYHSMM exploits frequency while semiCRF does not, at least explicitly. SemiCRF exploits contextual information more efficiently, but its high expressiveness is sometimes harmful. Because of this, it has difficulty in balancing between positive and negative examples. We treat the seed terms as positive examples and the general corpus as proxy data for negative examples. While semiCRF is too sensitive to use the general corpus as negative examples, PYHSMM utilizes them in a softer manner. We conducted extensive experiments on three domains in two languages and demonstrated that the proposed method outperformed baselines.

## 2 Related Work

### 2.1 General and Domain-Specific NER

NER is one of the fundamental tasks in NLP and has been applied not only to English but to a variety of languages such as Spanish, Dutch [9], and Japanese [10, 11]. NER can be classified into general NER and domain-specific NER. Typical NE classes in general NER are PERSON, LOCATION, and ORGANIZATION.

In domain-specific NER, special NE classes are defined to facilitate the development of downstream applications. For example, the GENIA corpus for the biomedical domain has five NE classes, such as DNA and PROTEIN, to organize research papers [12] and to extract semantic relations [13]. Disease corpora [14–16], which are annotated with the disease class and the treatment class, are used to solve disease-treatment relation extraction. However, domain-specific NER is not limited to only the biomedical domain; it also covers recipes [17] and game commentaries [18], to name a few examples. In addition, recognition of brand names and product names [19], recognition of the names of tasks, materials, and processes in science texts [20] can be seen as domain-specific NER.

### 2.2 Types of Supervision in NER

The standard approach to NER is supervised learning. Early studies used the hidden Markov model [21], the maximum entropy model [22], and support vector machines [23] before conditional random fields (CRFs) [24, 25] dominated. A CRF can be built on top of neural network components such as a bidirectional LSTM and convolutional neural networks [26].

Although modern high-performance NER systems require a large amount of annotated data in the form of labeled training examples, annotated corpora for domain-specific NER are usually of limited size because building NE corpora is costly and time-consuming. Tang et al. [5] proposed a transfer learning model for domain-specific NER with a medium-sized annotated corpus (about 6,000 sentences).

Several methods have been proposed to get around costly annotation and they can be classified into rule-based, heuristic feature-based, and weakly supervised methods. Rau [27] proposed a system to extract company names while Sekine and Nobata [28] proposed a rule-based NE tagger. Settles [29] proposed a CRF model with hand-crafted features for biomedical NER. These methods are time-consuming to develop and need specialized knowledge. Collins and Singer [30] proposed bootstrap methods for NE classification that exploited a small amount of seed data to classify NE chunks into typical NE classes. Nadeau et al. [31] proposed a two-step NER system in which NE extraction followed NE classification. Since their seed-based NE list generation from Web pages exploited HTML tree structures, it cannot be applied to plain text. Zhang and Elhadad [32] proposed another two-step NER method for the biomedical domain which first uses a noun phrase chunker to extract NE chunks and then classifies them using TF-IDF and biomedical terminology. Shang et al. [33] and Yang et al. [34] proposed weakly supervised methods by using domain-specific terminologies and unannotated target domain corpus. Shang et al. [33] automatically build a partially labeled corpus and then train a model by using it. Yang et al. [34] also use automatically labeled corpus and then select sentences to eliminate incomplete and noisy labeled sentences. The selector is trained on a human-labeled corpus. We also use automatically labeled corpus but there is a major difference. We focus on rapid prototyping of domain-specific NER that only requires a few seed terms because domain-specific terminologies are not necessarily available in other domains.

### 2.3 Unsupervised Word Segmentation and Part-of-Speech Induction

The model proposed in this paper has a close connection to unsupervised word segmentation and part-of-speech (POS) induction [6]. A key difference is that, while they use characters as the unit for the input sequence, we utilize word sequences.

Uchiumi et al. [6] can be seen as an extension to Mochihashi et al. [35], who focused on unsupervised word segmentation. They proposed a nonparametric Bayesian $n$-gram language model based on Pitman-Yor processes. Given an unsegmented corpus, the model infers word segmentation using Gibbs sampling. Uchiumi et al. [6] worked on the joint task of unsupervised word segmentation and POS induction. We employ their model, PYHSMM, for our task. However, instead of combining character sequences into words and assigning POS tags to them, we group word sequences into NE chunks and give NE classes to them.

To efficiently exploit annotated data when available, Fujii et al. [36] extended Mochihashi et al. [35] by integrating the generative word segmentation model into a CRF-based discriminative model. Our model, PYHSCRF, is also a hybrid generative/discriminative model but there are two major differences. First, to extend the approach to NER, we combine PYHSMM with a semiCRF, not an $n$-gram model with a plain CRF. Second, since our goal is to facilitate rapid prototyping of domain-specific

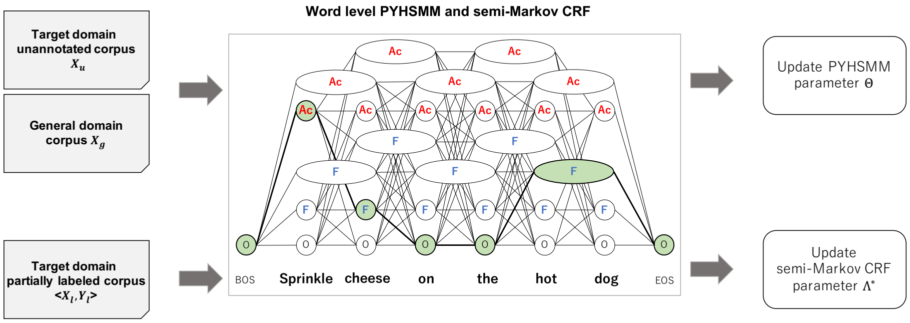
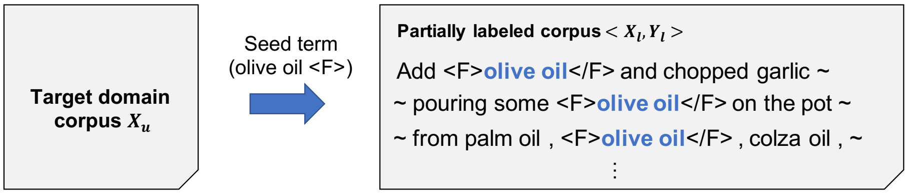

Word level PYHSMM and semi-Markov CRF
Target domain unannotated corpus $X_u$
General domain corpus $X_g$
Target domain partially labeled corpus $<X_l, Y_l>$
Update PYHSMM parameter Θ
Update semi-Markov CRF parameter Λ*
Ac
F
O
BOS
Sprinkle
cheese
on
the
hot
dog
EOS

**Fig. 1.** The overall architecture of PYHSCRF for domain-specific NER. Here, the maximum length of NE chunks $L = 2$. F and Ac stand for FOOD and ACTION, respectively, while O indicates a word outside of any NE chunks.


Target domain corpus $X_u$
Seed term (olive oil <F>)
Partially labeled corpus $< X_l, Y_l >$

Add <F>olive oil</F> and chopped garlic ~
~ pouring some <F>olive oil</F> on the pot ~
~ from palm oil , <F>olive oil</F> , colza oil , ~
⋮

**Fig. 2.** Partially labeled sentences. F stands for FOOD.

NER, we consider a much weaker type of supervision than fully annotated sentences: a few seed terms per NE class. This is challenging partly because seed terms can only be seen as implicit positive examples although most text fragments are outside of NE chunks (i.e., the O class). Our solution is to use a general domain corpus as implicit negative examples.

## 3 Proposed Method

### 3.1 Task Setting

NER is often formalized as a sequence labeling task. Given a word sequence $\boldsymbol{x} = (x_1, x_2, ..., x_N) \in \boldsymbol{X_l}$, our system outputs a label sequence $\boldsymbol{y} = (y_1, y_2, .., y_{N'}) \in \boldsymbol{Y_l}$, where $y_i = (z_i, b_i, e_i)$ means that a chunk starting at the $b_i$-th word and ending at the $e_i$-th word belongs to class $z_i$. The special O class is assigned to any word that is not part of an NE (if $z_i = \mathsf{O}$, then $b_i = e_i$). In the recipe domain, for example, the word sequence "*Sprinkle cheese on the hot dog*" contains an NE in the F (FOOD) class, "*hot dog*," which corresponds to $y_5 = (\mathsf{F}, 5, 6)$. Likewise the third word "*on*" is mapped to $y_3 = (\mathsf{O}, 3, 3)$.

We assume that we are given a few typical NEs per class (e.g., "*olive oil*" for the F class). Since choosing seed terms is by far less laborious than corpus annotation, our task settings allow us to design domain-specific NER in an exploratory manner. In addition to seed terms, an unannotated target domain corpus $\boldsymbol{X_u}$ and an unannotated general domain corpus $\boldsymbol{X_g}$ are provided. The underlying assumption is that domain-specific NEs are observed characteristically in $\boldsymbol{X_u}$. Contrasting $\boldsymbol{X_u}$ with $\boldsymbol{X_g}$ helps distinguishing NEs from the O class.

### 3.2 Model Overview

Figure 3.1 illustrates our approach. We use seed terms as implicit positive examples. We first automatically build a partially labeled corpus $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$ using seed terms. For example, if "*olive oil*" is selected as a seed term of class F, sentences in the target domain corpus $\boldsymbol{X_u}$ that contain the term are marked with its NE chunks and the class as in Figure 2. We train semiCRF using the partially labeled corpus (Section 3.3). To recognize high-frequency word sequences as NE chunks, we apply PYHSMM to the unannotated corpus $\boldsymbol{X_u}$ (Section 3.4). The general domain corpus $\boldsymbol{X_g}$ is also provided to the generative model as proxy data for the O class, with the assumption that domain-specific NE chunks should appear more frequently in the target domain corpus than in the general domain corpus. PYHSMM is expected to extract high-frequency word sequences in the target domain as NE chunks. Note that we do not train semiCRF with the implicit negative examples because the discriminative model is too sensitive to noise inherent to them. We combine the discriminative and generative models using JESS-CM [8] (Section 3.5).

### 3.3 Semi-Markov CRF with a Partially Labeled Corpus

We use semiCRF as the discriminative model, although Markov CRF is more often used as an NE tagger. Markov CRF employs the BIO tagging scheme or variants of it to identify NE chunks. Since each NE class is divided into multiple tags (e.g., B-PERSON and I-PERSON), it is unsuitable for our task, which is characterized by the scarcity of supervision. For this reason, we chose semiCRF.

SemiCRF is a log-linear model that directly infers NE chunks and classes. The probability of $\boldsymbol{y}$ given $\boldsymbol{x}$ is defined as:

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathit{\Lambda}) = \frac{\exp(\mathit{\Lambda} \cdot F(\boldsymbol{x}, \boldsymbol{y}))}{Z(\boldsymbol{x})},$$

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \boldsymbol{Y}} \exp(\mathit{\Lambda} \cdot F(\boldsymbol{x}, \boldsymbol{y})),$$

where $F(\boldsymbol{y}, \boldsymbol{x}) = (f_1, f_2, \cdots, f_M)$ are features, $\mathit{\Lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$ are the corresponding weights, and $\boldsymbol{Y}$ is the set of all possible label sequences. The feature function can be expressed as the combination of $F(b_i, e_i, z_i, z_{i-1})$ in relation to $x_i, y_i$, and $y_{i-1}$.

The training process is different from standard supervised learning because we use partially labeled corpus $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$. Following Tsuboi et al. [37], we marginalize the probabilities of words that are not labeled. Instead of using the full log likelihood

$$LL = F(\boldsymbol{x}, \boldsymbol{y}) - \sum_{\boldsymbol{y} \in \boldsymbol{Y}} p(\boldsymbol{y}|\boldsymbol{x}) F(\boldsymbol{x}, \boldsymbol{y})$$

as the objective function, we use the following marginalized log likelihood

$$MLL = \sum_{\boldsymbol{y} \in \boldsymbol{Y_p}} p(\boldsymbol{y}|\boldsymbol{Y_p}, \boldsymbol{x})F(\boldsymbol{x}, \boldsymbol{y}) - \sum_{\boldsymbol{y} \in \boldsymbol{Y}} p(\boldsymbol{y}|\boldsymbol{x})F(\boldsymbol{x}, \boldsymbol{y}),$$

where $\boldsymbol{Y_p}$ is the set of all possible label sequences in which labeled chunks are fixed.

### 3.4 PYHSMM

The generative model, PYHSMM, was originally proposed for joint unsupervised word segmentation and POS induction. While it was used to group character sequences into words and assign POS tags to them, here we extend it to word-level modeling. In our case, PYHSMM consists of 1) transitions between NE classes and 2) the emission of each NE chunk $x'_i = x_{b_i}, ..., x_{e_i}$ from its class $z_i$. As a semi-Markov model, it employs $n$-grams not only for calculating transition probabilities but also for computing emission probabilities.

The building blocks of PYHSMM are hierarchical Pitman-Yor processes, which can be seen as a back-off $n$-gram model. To calculate the transition and emission probabilities, we need to keep track of latent *table assignments* [38]. For notational brevity, let $\Theta$ be the set of the model's parameters. The joint probability of the $i$-th chunk $x'_i$ and its class $z_i$ conditioned on history $h_{xz}$ is given by

$$p(x'_i, z_i|h_{xz}; \Theta) = p(x'_i|h^n_x, z_i; \Theta)p(z_i|h^n_z; \Theta),$$

where $h^n_x = x'_{i-1}, x'_{i-2}, ..., x'_{i-(n-1)}$ and $h^n_z = z_{i-1}, z_{i-2}, ..., z_{i-(n-1)}$. $p(x'_i|h_x, z_i)$ is the chunk $n$-gram probability given its class $z_i$, and $p(z_i|h_z)$ is the class $n$-gram probability. The posterior predictive probability of the $i$-th chunk is

$$p(x'_i|h^n_x, z_i) = \frac{freq(x'_i|h^n_x) - d \cdot t_{x'_i, h^n_x}}{\theta + freq(h^n_x)} + \frac{\theta + d \cdot t_{h^n_x}}{\theta + freq(h^n_x)} p(x'_i|h^{n-1}_x, z_i), \qquad (1)$$

where $h^{n-1}_x$ is the shorter history of $(n-1)$-gram, $\theta$ and $d$ are hyperparameters, $freq(x'_i|h^n_x)$ is $n$-gram frequency, $t_{h^n_x, x'_i}$ is a count related to table assignments, $freq(h^n_x) = \sum_{x'_i} freq(x'_i|h^n_x)$, and $t_{h^n_x} = \sum_{x'_i} t_{h^n_x, x'_i}$. The class $n$-gram probability is computed in a similar manner.

Gibbs sampling is used to infer PYHSMM's parameters [35]. During training, we randomly select a sentence and remove it from the parameters (e.g., we subtract $n$-gram counts from $freq(x'_i|h^n_x)$). We sample a new label sequence using forward filtering-backward sampling. We then update the model parameters by adding the corresponding $n$-gram counts. We repeat the process until convergence.

Now we explain the sampling procedure in detail. We consider the bigram case for simplicity. The forward score $\alpha[t][k][z]$ is the probability that a sub-sequence $(x_1, x_2, ..., x_t)$ of a word sequence $\boldsymbol{x} = (x_1, x_2, ..., x_N)$ is generated with its last $k$ words being a chunk $(x^t_{t-k+1} = x_{t-k+1}, ..., x_t)$ which is generated from class $z$. Let $L$ be maximum length of a chunk and $Z$ be the number of classes. $\alpha[t][k][z]$ is recursively computed as follows:

$$\alpha[t][k][z] = \sum_{j=1}^{L}\sum_{r=1}^{Z}\left[p(x_{t-k+1}^{t}|x_{t-k-j+1}^{t-k}, z)p(z|r)\alpha[t-k][j][r]\right]. \tag{2}$$

The forward scores are calculated from the beginning to the end of the sentence. Chunks and classes are sampled in the reverse direction by using the forward score. There is always the special token EOS and its class $z_{\mathrm{EOS}}$ at the end of the sequence. The final chunk and its class in the sequence is sampled with the score proportional to

$$p(\mathrm{EOS}|w_{N-k}^{N}, z_{\mathrm{EOS}}) \cdot p(z_{\mathrm{EOS}}|z) \cdot \alpha[N][k][z].$$

The second-to-last chunk is sampled similarly using the score of the last chunk. We continue this process unti we reach the beginning of the sequence. To update the parameters in Equation (1), we add $n$-gram counts to $freq(x'_i|h_x^n)$ and $freq(h_x^n)$, and also update the table assignment count $t_{h_x^n,x'_i}$. Parameters related to the class $n$-gram model are updated in the same manner.

Recall that we use the general domain corpus $\boldsymbol{X_g}$ to learn the O class. We assume that $\boldsymbol{X_g}$ consists entirely of single-word chunks in the O class. Although the general domain corpus might contain some domain-specific NE chunks, most words indeed belong to the O class. During training, we add and remove sentences in $\boldsymbol{X_g}$ without performing sampling. Thus these sentences can be seen as implicit negative samples.

### 3.5 PYHSCRF

PYHSCRF combines discriminative semiCRF with generative PYHSMM in a similar manner to the model presented in Fujii et al. [36]. The probability of label sequence $\boldsymbol{y}$ given word sequence $\boldsymbol{x}$ is written as follows:

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto p_{\mathrm{DISC}}(\boldsymbol{y}|\boldsymbol{x}; \mathit{\Lambda})\; p_{\mathrm{GEN}}(\boldsymbol{y}, \boldsymbol{x}; \mathit{\Theta})^{\lambda_0},$$

where $p_{\mathrm{DISC}}$ and $p_{\mathrm{GEN}}$ are the discriminative and generative models, respectively. $\mathit{\Lambda}$ and $\mathit{\Theta}$ are their corresponding parameters. When $p_{\mathrm{DISC}}$ is a log-linear model like semiCRF, $p(\boldsymbol{y}|\boldsymbol{x})$ can be expressed as a log-linear model:

$$\begin{aligned}
p_{\mathrm{DISC}}(\boldsymbol{y}|\boldsymbol{x}) &\propto \exp\left[\sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{y}, \boldsymbol{x})\right], \\
p(\boldsymbol{y}|\boldsymbol{x}) &\propto \exp\left[\lambda_0 \log(p_{\mathrm{GEN}}(\boldsymbol{y}, \boldsymbol{x})) + \sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{y}, \boldsymbol{x})\right] \\
&= \exp(\mathit{\Lambda}^* \cdot F^*(\boldsymbol{y}, \boldsymbol{x})),
\end{aligned} \tag{3}$$

where

$$\begin{aligned}
\mathit{\Lambda}^* &= (\lambda_0, \lambda_1, \lambda_2 ..., \lambda_M), \\
F^*(\boldsymbol{y}, \boldsymbol{x}) &= (\log(p_{\mathrm{GEN}}), f_1, f_2, ..., f_M).
\end{aligned}$$

**Algorithm 1** Learning algorithm for PYHSCRF. $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$ is a partially labeled corpus and $\boldsymbol{X_u}$ is an unannotated corpus in the target domain. $\boldsymbol{X_g}$ is the general domain corpus used as implicit negative examples.

```
for $epoch = 1, 2, ..., E$ do
    for $\boldsymbol{x}$ $in$ randperm($\boldsymbol{X_u}$, $\boldsymbol{X_g}$) do
        if $epoch > 1$ then
            Remove parameters of $\boldsymbol{y}$ from $\mathit{\Theta}$
        end if
        if $\boldsymbol{x} \in X_u$ then
            Sample $\boldsymbol{y}$ according to $p(\boldsymbol{y}|\boldsymbol{x}; \mathit{\Lambda}^*, \mathit{\Theta})$
        else
            Determine $\boldsymbol{y}$ according to $\boldsymbol{X_g}$
        end if
        Add parameters of $\boldsymbol{y}$ to $\mathit{\Theta}$
    end for
    Optimize $\mathit{\Lambda}^*$ on $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$
end for
```

In other words, PYHCRF is another semiCRF in which PYHSMM is added to the original semiCRF as a feature. The objective function is

$$p(\boldsymbol{Y_l}|\boldsymbol{X_l}; \mathit{\Lambda}^*)\; p(\boldsymbol{X_u}, \boldsymbol{X_g}; \mathit{\Theta}).$$

Algorithm 1 shows our training algorithm. During training, PYHSCRF repeats the following two steps:

1. fixing $\mathit{\Theta}$ and optimizing $\mathit{\Lambda}^*$ of semiCRF on $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$,
2. fixing $\mathit{\Lambda}^*$ and optimizing $\mathit{\Theta}$ of PYHSMM on $\boldsymbol{X_u}, \boldsymbol{X_g}$

until convergence. When updating $\mathit{\Lambda}^*$, we use the marginalized log likelihood of the partially labeled data. When updating $\mathit{\Theta}$, we sample chunks and their classes from unlabeled sentences in the same manner as in PYSHMM. In PYHSCRF, a modification is needed to Equation (2) because forward score $\alpha[t][k][z]$ incorporates the semiCRF score:

$$\alpha[t][k][z] = \sum_{j=1}^{L} \sum_{r=1}^{Z} \exp\Bigg[\lambda_0 \log(p(x_{t-k+1}^{t}|x_{t-k-j+1}^{t-k}, z)p(z|r)) \\ + \mathit{\Lambda} \cdot F(t-k+1, t, z, r)\Bigg] \alpha[t-k][j][r],$$

where $F(t-k+1, t, z, r)$ is a feature function in relation to chunk candidate $x_{t-k+1}^{t}$, its class $z$, and class $r$ of the preceding chunk candidate $x_{t-k-j+1}^{t-k}$.

## 4 Experimentals

### 4.1 Data

Table 1 summarizes the specifications of three domain-specific NER datasets used in our experiments: the GENIA corpus, the recipe corpus, and the game commentary cor-

**Table 1.** Statistics of the datasets for the experiments.

| Language | Corpus (#NE classes) | #Sentences | #Words | #NE instances |
|---|---|---|---|---|
| English | Target | | | |
| | **GENIA corpus** (5) | | | |
| | train | 10,000 | 264,743 | - |
| | test | 3,856 | 101,039 | 90,309 |
| | General | | | |
| | **Brown** (-) | 50,000 | 1039,886 | - |
| Japanese | Target | | | |
| | **Recipe corpus** (8) | | | |
| | train | 10,000 | 244,648 | - |
| | test | 148 | 2,667 | 869 |
| | **Game commentary corpus** (21) | | | |
| | train | 10,000 | 398,947 | - |
| | test | 491 | 7,161 | 2,365 |
| | General | | | |
| | **BCCWJ** (-) | 40,000 | 936,498 | - |
| | **Oral communication corpus** (-) | 10,000 | 124,031 | - |

pus. We used the GENIA corpus, together with its test script in the BioNLP/NLPBA 2004 shared task [39], as an English corpus for the biomedical domain. It contains five biological NE classes such as DNA and PROTEIN in addition to the O class. The corresponding general domain corpus was the Brown corpus [40], which consists of one million words and ranges over 15 domains.

The recipe corpus [17] and the game commentary corpus [18] are both in Japanese. The recipe corpus consists of procedural texts from recipes for cooking. The game commentary corpus consists of commentaries on professional matches of Japanese chess (*shogi*) given by professional players and writers. We used gold-standard word segmentation for both corpora. As NEs, eight classes such as FOOD, TOOL, and ACTION were defined for the recipe corpus, while the game commentary corpus was annotated with 21 classes such as PERSON, STRATEGY, and ACTION. Note that NE chunks were not necessarily noun phrases. For example, most NE chunks labeled with ACTION in the two corpora were verbal phrases. The combination of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [41] and the oral communication corpus [42] were used as the general domain corpus. We automatically segmented sentences in these corpora using KyTea[1] [43]. (The segmentation accuracy was higher than $98\%$.)

### 4.2 Training Settings

Although PYHSMM can theoretically handle arbitrarily long $n$-grams, we limited our scope to bigrams to reduce computational costs. To initialize PYHSMM's parameter, $\Theta$, we treated each word in a given sentence as an O-class chunk. Just like Uchiumi et al. [6] modeled expected word length with negative binomial distributions for the tasks of Japanese word segmentation and POS induction, chunk length was drawn from a

[1] http://www.phontron.com/kytea/ (accessed on March 15, 2017)

**Table 2.** Feature templates for semiCRF. $chunk_i$ consists of word $n$-grams $\boldsymbol{w}_{b_i}^{e_i} = w_{b_i} w_{b_i+1} ... w_{e_i}$, $w_{i-1}$. $w_{i-1}$ and $w_{i+1}$ are the preceding word and the following word, respectively. BoW is a set of words (bag-of-words) in $chunk_i$.

| Semi-Markov CRF features |
|---|
| $chunk_i(w_{b_i} w_{b_1+1} ... w_{e_i})$ |
| $w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}$ |
| BoW($w_{b_i}$, $w_{b_i+1}$, ..., $w_{e_i}$) |

**Table 3.** Precision, recall, and F-measure of various systems.

| Target | | | |
|---|---|---|---|
| Method | Precision | Recall | F-measure |
| GENIA | | | |
| MetaMap [44] | N/A | N/A | 7.70 |
| weakly supervised biomedical NER [32] | 15.40 | 15.00 | 15.20 |
| PYHSCRF | **19.20** | **23.50** | **21.13** |
| Recipe | | | |
| Baseline | **49.78** | 25.89 | 34.07 |
| PYHSCRF | 38.45 | **42.58** | **40.41** |
| Game | | | |
| Baseline | 52.75 | 29.18 | 37.57 |
| PYHSCRF | **75.57** | **35.05** | **47.89** |

negative binomial distribution. Uchiumi et al. [6] set different parameters for character types such as *hiragana* and *kanji*, but we used a single parameter. We constrained the maximum length of chunk $L$ to be 6 for computational efficiency.

We used the normal priors of truncated $N(\mu, \sigma^2)$ to initialize PYHSMM's weight $\lambda_0$ and semiCRF's weights $\lambda_1, \lambda_2, \cdots, \lambda_M$. We set $\mu = 1.0$ and $\sigma = 1.0$. We fixed the L2 regularization parameter $C$ of semiCRF to 1.0. We used stochastic gradient descent for optimization of semiCRF. The number of iterations $J$ was set to 300. Table 2 shows the feature templates for semiCRF.

Each target domain corpus was divided into a training set and a test set. For each NE class, the 2 most frequent chunks according to the training set were selected as seed terms. In the GENIA corpus, for example, we automatically chose "IL-2" and "LTR" as seed terms for the DNA class.

### 4.3 Baselines

In biomedical NER, the proposed model was compared with two baselines. MetaMap is based on a dictionary matching approach with biomedical terminology [44]. The other baseline model is a weakly supervised biomedical NER system proposed by Zhang and Elhadad [32]. To our knowledge, there was no weakly supervised domain-specific NER tool in the recipe and game commentary domains. For these domains, we created a baseline model as follows: We first used a Japanese term extractor[2] to extract NE chunks

[2] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html (accessed on March 15, 2017)

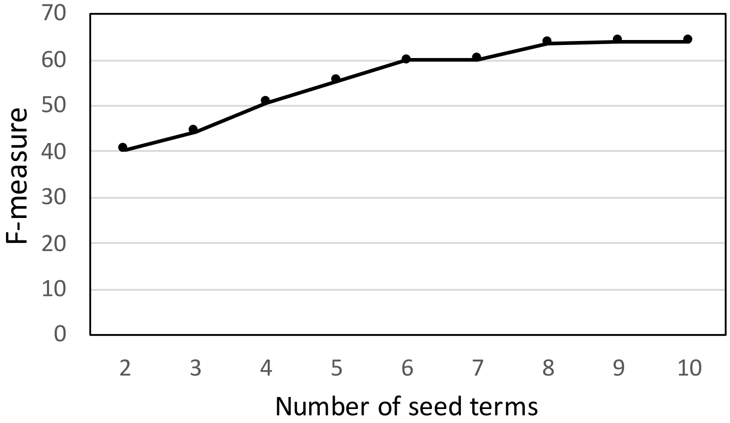


**Fig. 3.** Learning curve for recipe NER. The horizontal axis shows number of seed terms in each NE class.

and then classified them with seed terms using a Bayesian HMM originally proposed for unsupervised POS induction [45]. Note that only noun phrases were extracted by the term extractor.

### 4.4 Results and Discussion

Table 3 compares the proposed method with baselines in terms of precision, recall, and F-measure. We can see that PYHSCRF consistently outperformed the baselines.

Taking a closer look at the results, we found that the model successfully inferred NE classes from their contexts. For example, the NE chunk “水” (water) can be both FOOD and TOOL in the recipe domain. It was correctly identified as TOOL when it was part of the phrase “水で洗い流す” (wash with water) while the phrase “水を鍋に加える” (add water in the pot) was identified as the FOOD class.

We conducted a series of additional experiments. First, we changed the number of seed terms to examine their effects. Figure 3 shows F-measure as a function of the number of seed terms per NE class in the recipe domain. The F-measure increased almost monotonically as more seed terms became available.

A major advantage of PYHSCRF over other seed-based weakly supervised methods for NER [32, 31] is that it can straightforwardly exploit labeled instances. To see this, we trained PYHSCRF with fully annotated data (about 2,000 sentences) in the recipe domain and compared it with vanilla semiCRF. We found that they achieve competitive performance (the F-measure was $90.01$ for PYHSCRF and $89.98$ for vanilla semiCRF). In this setting, PYHSCRF ended up simply ignoring PYHSMM ($-0.1 < \lambda_0 < 0.0$).

Next, we reduced the size of the general domain corpus. Figure 4 shows how F-measure changes with the size of the general domain corpus in recipe NER. We can confirm that PYHSCRF cannot be trained without the general domain corpus because it is a vital source for distinguishing NE chunks from the O class.

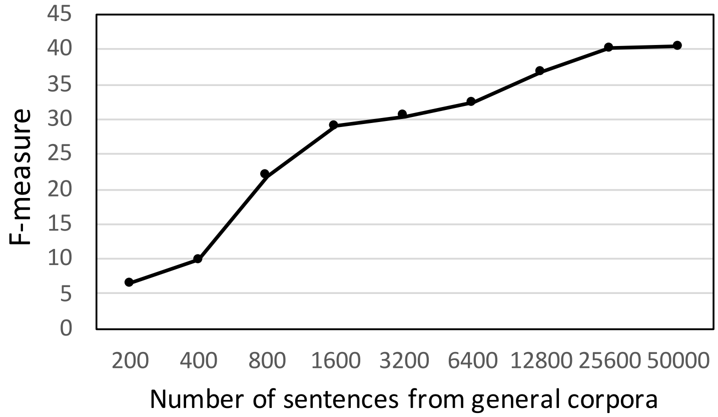


**Fig. 4.** Learning curve for recipe NER. The horizontal axis shows number of general domain sentences.

Finally, we evaluated NE classification performance. Collins and Singer [30] focused on weakly supervised NE classification, in which given NE chunks were classified into three classes (PERSON, LOCATION, and ORGANIZATION) by bootstrapping with seven seed terms and hand-crafted features. We tested PYHSCRF with the CoNLL 2003 dataset [46] in the same settings. We did not use a general corpus because NE chunks are given a priori. PYHSCRF achieved competitive performance (over $93\%$ accuracy, compared to over $91\%$ accuracy for Collins and Singer [30]) although the use of different datasets makes direct comparison difficult.

The semiCRF feature templates in our experiments are simple. Though not explored here, the accuracies can probably be improved by a wider window size or richer feature sets such as character type and POS. Word embeddings [47, 48], character embeddings [49], and $n$-gram embeddings [50] are other possible improvements because domain-specific NE chunks exhibit spelling variants. For example, in the Japanese recipe corpus, the NE chunk “玉ねぎ” (onion, *kanji* followed by *hiragana*) can also be written as “たまねぎ” (*hiragana*), “タマネギ” (*katakana*), and “玉葱” (*kanji*).

## 5 Conclusion

We proposed PYHSCRF, a nonparametric Bayesian method for distant supervised NER in specialized domains. PYHSCRF is useful for rapid prototyping domain-specific NER because it does not need texts annotated with NE tags and boundaries. We only need a few seed terms as typical NEs in each NE class, an unannotated corpus in the target domain, and a general domain corpus. PYHSCRF incorporates word level PYHSMM and semiCRF. In addition, we use implicit negative examples from the general domain corpus to train the O class.

In our experiments, we used a biomedical corpus in English, and a recipe corpus and a game commentary corpus in Japanese as examples. We conducted domain-specific NER experiments and showed that PYHSCRF achieved higher accuracy than the baselines. Therefore we can build a domain-specific NE recognizer with much less cost. Additionally, PYHSCRF can be easily applied to other domains for domain-specific NER and is useful for low-resource languages and domains.

In the future, we would like to investigate the effectiveness of the proposed method for downstream tasks of domain-specific NER such as relation extraction and knowledge base population.

## Acknowledgement

In this paper, we used recipe data provided by Cookpad and the National Institute of Informatics.

## References

1. Thompson, P., Dozier, C.C.: Name searching and information retrieval. CoRR **cmp-lg/9706017** (1997)
2. Feldman, R., Rosenfeld, B.: Boosting unsupervised relation extraction by using NER. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. (2006) 473–481
3. Lee, H., Recasens, M., Chang, A., Surdeanu, M., Jurafsky, D.: Joint entity and event coreference resolution across documents. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. (2012) 489–500
4. Shahab, E.: A short survey of biomedical relation extraction techniques. CoRR **abs/1707.05850** (2017)
5. Tang, S., Zhang, N., Zhang, J., Wu, F., Zhuang, Y.: NITE: A neural inductive teaching framework for domain specific NER. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017) 2642–2647
6. Uchiumi, K., Tsukahara, H., Mochihashi, D.: Inducing word and part-of-speech with Pitman-Yor hidden semi-Markov models. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). (2015) 1774–1782
7. Sarawagi, S., Cohen, W.W.: Semi-Markov conditional random fields for information extraction. In: Advances in Neural Information Processing Systems 17. (2005) 1185–1192
8. Suzuki, J., Isozaki, H.: Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In: Proceedings of ACL-08: HLT, Association for Computational Linguistics (2008) 665–673
9. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, proceedings of the 6th conference on natural language learning. August **31** (2002) 1–4
10. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. Volume 1. (1996)
11. Sekine, S., Isahara, H.: IREX: IR and IE evaluation project in Japanese. In: Proceedings of International Conference on Language Resources & Evaluation. (2000)

12. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus: A semantically annotated corpus for bio-textmining. **19 Suppl 1** (2003) i180–2
13. Ciaramita, M., Gangemi, A., Ratsch, E., Šaric, J., Rojas, I.: Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. (2005) 659–664
14. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18** (2011) 552–556
15. Doğan, R.I., Lu, Z.: An improved corpus of disease mentions in PubMed citations. In: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. (2012) 91–99
16. Doğan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics **47** (2014) 1–10
17. Mori, S., Maeta, H., Yamakata, Y., Sasada, T.: Flow graph corpus from recipe texts. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). (2014) 2370–2377
18. Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H., Tsuruoka, Y.: A Japanese chess commentary corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluatio. (2016) 1415–1420
19. Bick, E.: A named entity recognizer for Danish. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). (2004)
20. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). (2017) 546–555
21. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Proceedings of the fifth conference on Applied natural language processing. (1997) 194–201
22. Borthwick, A.E.: A Maximum Entropy Approach to Named Entity Recognition. PhD thesis (1999) AAI9945252.
23. Asahara, M., Matsumoto, Y.: Japanese named entity extraction with redundant morphological analysis. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. (2003) 8–15
24. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01 (2001) 282–289
25. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. (2003) 188–191
26. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2016) 1064–1074
27. Rau, L.F.: Extracting company names from text. In: Proceedings of the Seventh Conference on Artificial Intelligence Applications CAIA-91 (Volume II: Visuals). (1991) 189–194
28. Sekine, S., Nobata, C.: Definition, dictionaries and tagger for extended named entity hierarchy. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). (2004)
29. Settles, B.: Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004) 33–38

30. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora. (1999)
31. Nadeau, D., Turney, P.D., Matwin, S.: Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In: Conference of the Canadian Society for Computational Studies of Intelligence. (2006) 266–277
32. Zhang, S., Elhadad, N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics **46** (2013) 1088–1098
33. Shang, J., Liu, L., Gu, X., Ren, X., Ren, T., Han, J.: Learning named entity tagger using domain-specific dictionary. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2018) 2054–2064
34. Yang, Y., Chen, W., Li, Z., He, Z., Zhang, M.: Distantly supervised ner with partial annotation learning and reinforcement learning. In: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics (2018) 2159–2169
35. Mochihashi, D., Yamada, T., Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. (2009) 100–108
36. Fujii, R., Domoto, R., Mochihashi, D.: Nonparametric Bayesian semi-supervised word segmentation. Transactions of the Association for Computational Linguistics **5** (2017) 179–189
37. Tsuboi, Y., Kashima, H., Mori, S., Oda, H., Matsumoto, Y.: Training conditional random fields using incomplete annotations. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). (2008) 897–904
38. Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. (2006) 985–992
39. Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04). (2004) 70–75
40. Francis, W.N., Kucera, H.: Brown corpus manual. Brown University **2** (1979)
41. Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., Den, Y.: Balanced corpus of contemporary written Japanese. Language Resources and Evaluation **48** (2014) 345–371
42. Keene, D., Hatori, H., Yamada, H., Irabu, S.: Japanese-English Sentence Equivalents. Electronic book edn. Asahi Press (1992)
43. Neubig, G., Nakata, Y., Mori, S.: Pointwise prediction for robust, adaptable Japanese morphological analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. (2011) 529–533
44. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. (2001) 17
45. Goldwater, S., Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. (2007) 744–751
46. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. CoNLL '03 (2003) 142–147
47. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

48. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems **26** (2013) 3111–3119
49. Wieting, J., Bansal, M., Gimpel, K., Livescu, K.: Charagram: Embedding words and sentences via character n-grams. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016) 1504–1515
50. Zhao, Z., Liu, T., Li, S., Li, B., Du, X.: Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017) 244–253