International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1

A Web Semantic-Based Text Analysis Approach for Enhancing Named Entity Recognition Using PU-Learning and Negative Sampling

Shunqin Zhang, School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing, China
Sanguo Zhang, School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing, China
Wenduo He, Institute for Network Sciences and Cyberspace (INSC), Tsinghua University, Beijing, China
Xuan Zhang, Tsinghua University, China*

ABSTRACT

The NER task is largely developed based on well-annotated data. However, in many scenarios, the entities may not be fully annotated, leading to serious performance degradation. To address this issue, the authors propose a robust NER approach that combines a novel PU-learning algorithm and negative sampling. Unlike many existing studies, the proposed method adopts a two-step procedure for handling unlabeled entities, thereby enhancing its capability to mitigate the impact of such entities. Moreover, this algorithm demonstrates high versatility and can be integrated into any token-level NER model with ease. The effectiveness of the proposed method is verified on several classic NER models and datasets, demonstrating its strong ability to handle unlabeled entities. Finally, the authors achieve competitive performances on synthetic and real-world datasets.

KEYWORDS

Negative Sampling, NER, PU-Learning, Robustness, Self-Denoising, Token-Level, Two-Step Procedure, Unlabeled Entity Problem

INTRODUCTION

Named-entity recognition (NER) is a well-studied task in natural language processing (NLP) (Tekli et al., 2021; Barbosa et al., 2022; Ehrmann et al., 2023) that has received significant attention (Huang et al., 2015; Ma & Hovy, 2016; Akbik et al., 2018; Li et al., 2020a). In the area of NER, previous methods have had great success (Zhang & Yang, 2018; Gui et al., 2019; Jin et al., 2019; Wang et al., 2023). However, the majority of them rely on well-annotated data and ignore potential unlabeled entities, which are commonly encountered in many cases. Li et al. (2020c) discovered that NER models suffer significantly from the lack of annotations and referred to this as the unlabeled-entity problem.

DOI: 10.4018/IJSWIS.335113 *Corresponding Author

Unlabeled entities often arise from mistakes made by human annotators or the limitations of machine annotators. For instance, distant supervision is a classic method to produce labeled NER data automatically. However, owing to the limited coverage of knowledge resources, datasets generated through distant supervision often retain a significant number of unlabeled entities. Furthermore, enhancing performance with a small set of annotated data could significantly reduce costs. As such, developing an effective and versatile method for NER with unlabeled entities is of great research interest. However, there are several challenges. First, unlabeled entities will misguide the NER training process, causing the model to learn entities as negative instances. It is hard to identify unlabeled entities since they are always confused with negative instances. Second, the reduction of unlabeled entities results in a decrease in learnable data, making it challenging for the model to identify entities correctly. These challenges need to be effectively addressed.

Recently, numerous approaches to alleviate the unlabeled entity problem have been developed. To begin with, Li et al. (2020c) utilized a negative-sampling approach and trained a span-based model to mitigate the misguidance caused by unlabeled entities. They assumed that the unlabeled entities were unknown and thus applied random sampling to cover the unlabeled entities. This line of work was further extended by Li et al. (2022), who used a new weighted sampling distribution to perform a better sampling. Furthermore, Peng et al. (2021) considered reinforcement learning and trained a span selector to enhance the negative-sampling approach.

Another approach makes full use of the labeled data to approximate the true label sequences or detect the potential unlabeled entities. A classic algorithm called positive-unlabeled learning (PU learning) is designed for scenarios where some kinds of samples are easily obtained, but full labeling of all samples is either difficult to obtain or too costly. For instance, Mayhew et al. (2019) proposed the constrained binary learning method, which adaptively trained a binary classifier and assigned weights to each token using the CoDL framework (Chang et al., 2007). Peng et al. (2019) trained a PU-learning (Liu et al., 2002, 2003; Elkan & Noto, 2008; Shunxiang et al., 2023) classifier to perform label prediction; it can unbiasedly and consistently estimate the task loss. Zhang et al. (2022) proposed an adaptive PU-learning technology and then handled the unlabeled-entity problem by integrating it into a machine reading comprehension (MRC) framework. PU learning is widely applied in various fields where obtaining a comprehensive labeled dataset is hard or impractical, offering a solution to effectively utilize limited labeled data along with a larger pool of unlabeled data for better performance.

Another classic algorithm is partial conditional random fields (CRF) (Tsuboi et al., 2008), which is also an effective method for handling unlabeled entities (Yang et al., 2018; Jie et al., 2019; Ding et al., 2023). It functions by generating all potential label sequences for uncertain annotations and subsequently trains on these sequences.

The current methods have achieved great improvement in datasets with unlabeled entities. Despite their success, they also have some limitations. On the one hand, methods that rely on annotated data for self-denoising are heavily dependent on the quality of the available data. These methods often struggle to significantly reduce the impact of unlabeled entities. On the other hand, entirely ignoring annotated information might result in the underuse of valuable data (for example, the negative-sampling techniques). As such, we believe that strategies either entirely dependent on or independent of annotated data are suboptimal. We propose that some of the unlabeled entities can be identified through self-learning methods, yet a certain fraction remains undetected. Thus, combining the advantages of both techniques could offer a more effective solution to the problem of unlabeled entities.

To better solve the unlabeled-entity problem, we propose a robust NER approach that combines a novel PU-learning algorithm and negative sampling. From our empirical studies, we found that the annotated data can be utilized to identify some unlabeled entities in the early stages of model training. However, this capability is constrained, and, notably, a significant portion of unlabeled entities are still confused with negative instances. To address this problem, our approach adopts a two-step strategy for better handling unlabeled entities. Specifically, based on our empirical findings, we propose a

novel PU-learning algorithm to handle the unlabeled entities in the first step, significantly reducing the misguidance caused by unlabeled entities. Following this, we apply negative sampling to further diminish the influence of remaining potential unlabeled entities. Through these two steps, we aim to handle unlabeled entities more effectively and establish a more stable NER model. Furthermore, we use the angle-based technique (Zhang & Liu, 2014; Zhang et al., 2016; Fu et al., 2022) to improve the accuracy of our PU-learning algorithm, which is first encountered in deep neural networks.

The key contributions of this study are that, first, unlike previous studies, we innovatively integrate PU learning with negative sampling, enhancing robustness in the handling of unlabeled entities. Second, we further enhance the capability of our approach by leveraging the angle-based technique in the proposed PU-learning algorithm. Third, our method offers significant flexibility. It allows for straightforward application into any token-level NER models.

We evaluate the proposed method using four classic NER models and six benchmark NER datasets. The proposed method generally enables significant improvement of all baseline models on synthetic datasets. In a real-world situation, our approach also delivers competitive performance. Notably, the computational cost caused by our method is significantly small.

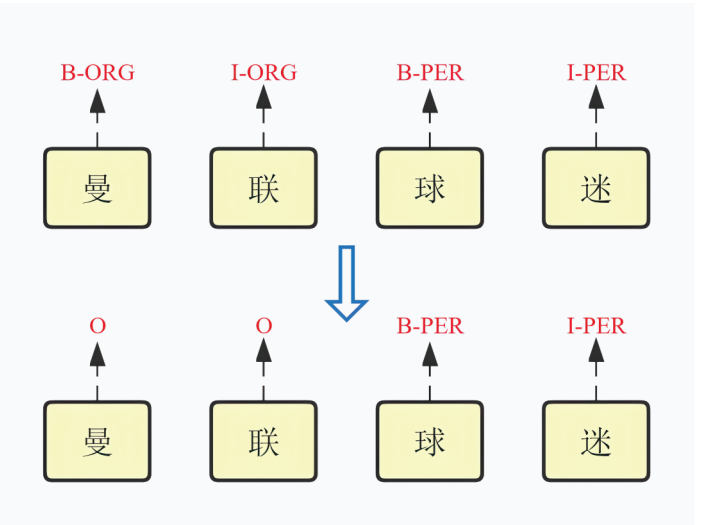
PRELIMINARIES

The Unlabeled-Entity Problem

The unlabeled-entity problem arises in scenarios where the entities are not fully annotated and are treated as negative instances. This problem may be caused by the negligence of the human annotator or the machine-based annotator with limited capabilities. In the setting of the unlabeled-entity problem, all (or most) observed entities are correct. However, limited entities are annotated, and no reliable sample is available.

For instance, consider a phrase (Manchester United) (football fan) which adopts beginning-inside-outside (BIO) tagging scheme. The true label sequence is {B-ORG,I-ORG,B-PER,I-PER}. As shown in Figure 1, when the unlabeled-entity problem occurs, the entity is not annotated, and the corresponding labels are replaced with the tag O.

Figure 1. Unlabeled-Entity problem



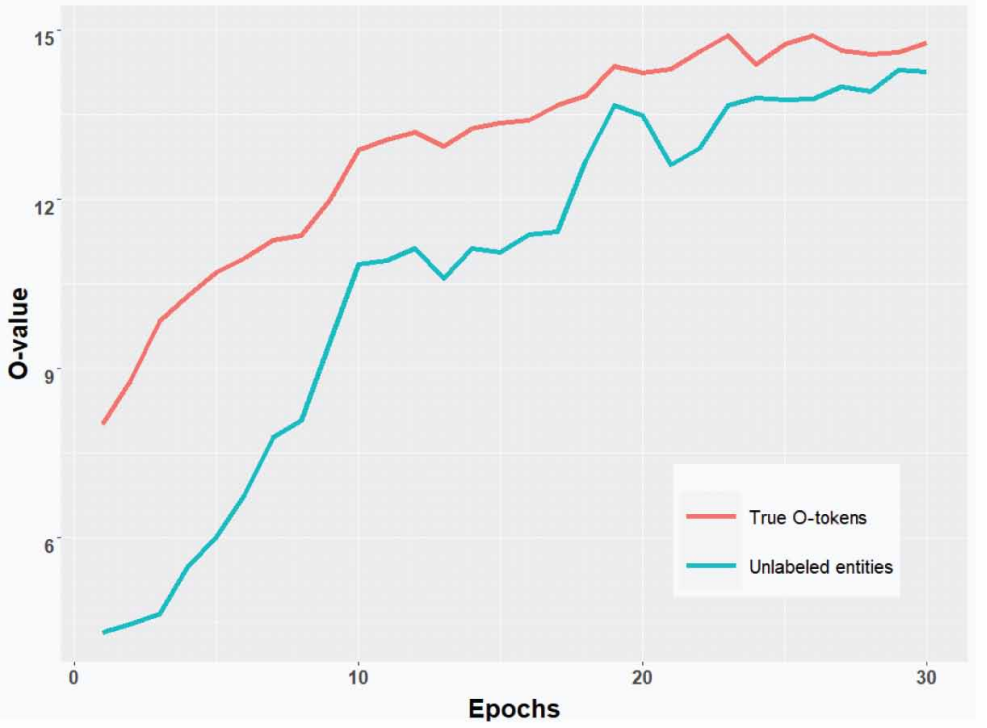
Motivation

As discussed in Li et al. (2020c), there are two causes for performance degradation in the unlabeled-entity problem: the reduction of annotated entities and the misguidance of unlabeled entities. The second cause has far more influence than the first one, and it can be mitigated by removing all unlabeled entities (Li et al., 2020c). Ideally, we would be able to detect all unlabeled entities correctly. However, the unlabeled entities are always confused with true negative instances. Therefore, the current challenge is how to discriminate between unlabeled entities and negative instances.

Arpit et al. (2017) showed that deep neural networks learn simple patterns first and subsequently fit the noise. Inspired by this, we performed a simple study to understand the training process for unlabeled entities. To begin with, we introduced the definition of O-value. For token-level sequence-labeling tasks, we always generated a k-dimensional decision vector for one token to match the classification, where k is the number of tags. Consider the decision vector $\boldsymbol{h} \in \mathbb{R}^k$ of token c. The O-value of token c is defined as $\boldsymbol{h}[o]$, which is the value in $\boldsymbol{h}$ corresponding to the O tag. Then we trained 30 epochs in the synthetic Weibo dataset with 50% unlabeled entities and plotted the average O-values for unlabeled entities and true negative instances. In this synthetic dataset, we randomly selected 50% of the labeled entities and relabeled them as O. Note that we counted every token to calculate the average. As shown in Figure 2, the average O-value for unlabeled entities is much smaller than the true negative instances in the first few epochs. With the increase of epochs, their difference becomes smaller and almost disappears. The result indicates that the unlabeled entities are learned by steps, which confirms the point of Arpit et al. (2017).

Since the difference in average O-values is significant in the early stage of training, we tried to detect unlabeled entities by their O-values. We fixed the number of training epochs at two and selected a portion of tokens labeled O, from those with the smallest to the largest O-values, to analyze their

Figure 2. Average O-value for true O-tokens and unlabeled entities (Note: O-tokens are the tokens tagged as O)



efficacy in detecting unlabeled entities. The results are in Figure 3. In general, the precision keeps decreasing while the recall keeps increasing as we pick more tokens. As a result, we can detect only a portion of unlabeled entities correctly. The others are still confused with the negative instances after training, and we have to pay more to cover them. Therefore, we can achieve only limited improvement in the unlabeled-entity problem by utilizing labeled data.

METHODOLOGY

In this section, we will introduce: (1) the proposed two-step learning approach, which encounters both PU learning and negative sampling; (2) the application of the angle-based technique; (3) the implemention of the CRF layer.

Robust Two-Step Learning

The core framework of our approach is depicted in Figure 4. In general, we detect a part of the unlabeled entities by self-learning in the first step and then apply negative sampling to further enhance our model. In the first step, we aim to detect and then remove some unlabeled entities with high

Figure 3. Precision (left) and recall (right) for detecting unlabeled entities

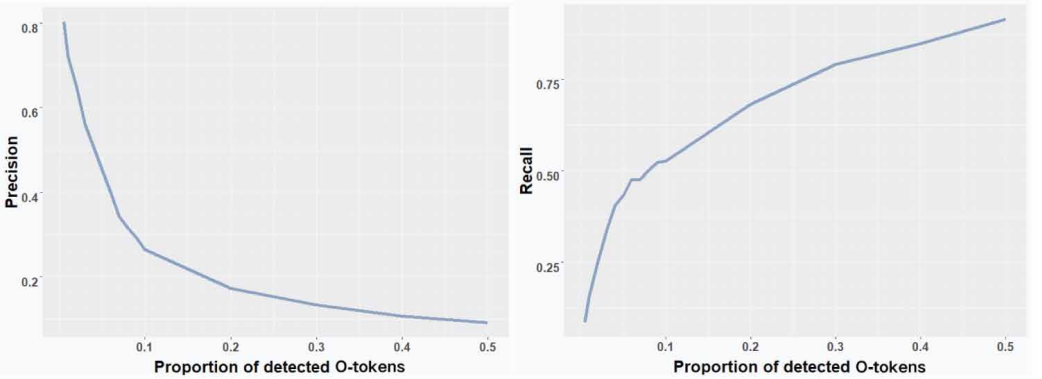
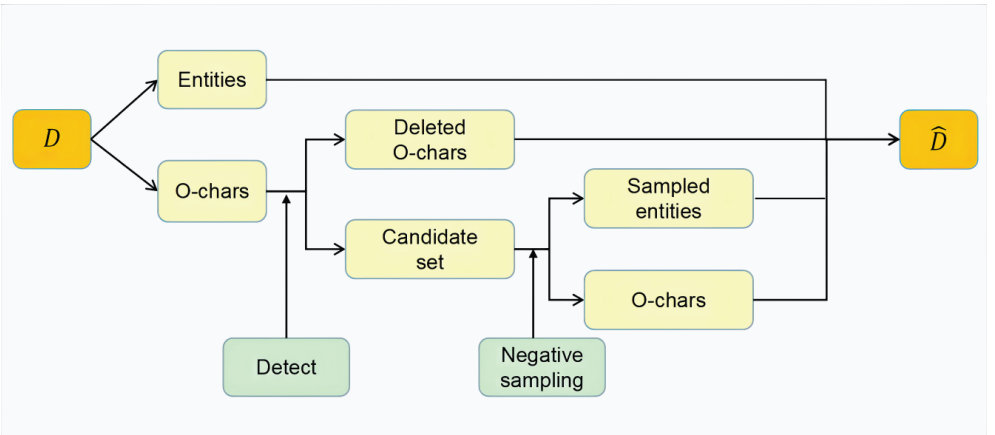


Figure 4. The architecture of our two-step method, in which the pu-learning and negative-sampling approaches are encountered (Note: D is the original dataset and $\hat{D}$ is the modified dataset)



confidence in preparation for the following negative-sampling step. Inspired by the empirical studies in the Preliminaries section, we propose our novel PU-learning algorithm. The details are as follows.

Consider a NER model and the training set $D = \left\{\left(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right), i = 1,\ldots,N\right\}$, where $\boldsymbol{y}^{(i)} = \left(y_1^{(i)},\ldots,y_{s_i}^{(i)}\right)$ and s_i is the length of ith sentence. In empirical studies, we observed that characters with smaller O-values tend to represent unlabeled entities, especially in the initial training epochs. As such, we first train the model for n epochs, where n is a small hyperparameter. After training, we fix the trained model and calculate the decision vectors for all tokens at this point, denoted as $\boldsymbol{f}\left(\boldsymbol{x}^{(i)}\right) = \boldsymbol{h}^{(i)} = \left(\boldsymbol{h}_1^{(i)},\ldots,\boldsymbol{h}_{\boldsymbol{s_i}}^{(i)}\right) \in \mathbb{R}^{s_i \times k}$ for the jth sentence. Afterward, we summarize O-values for all tokens tagged as O in:

$$\left\{\boldsymbol{h}_j^{(i)}\left[o\right] 1 \le i \le N, 1 \le j \le s_i, y_j^{(i)} = O\right\} \tag{1}$$

Here, we employ an intuitive and efficient PU-learning algorithm. Specifically, we select the smallest $\lceil \lambda * m \rceil$ O-values in (1) and will remove their corresponding losses in the subsequent training process. We define $\mathcal{A}_j$ as the set of index for the deleted tokens in the jth sentence. Here, $0 < \lambda < 1$ is a hyperparameter, m is the number of O-values, and $\lceil\ \rceil$ is the ceiling function. Since the smaller λ yields higher precision, we prefer choosing a smaller λ in practice. It is noted that we restart our training process using the new loss function after n epochs. It means the initial n epochs serve as exploratory training to identify a subset of unlabeled entities with higher confidence, followed by retraining. This is the core idea of our PU-learning approach.

Furthermore, we did not change the computations within the input layer and hidden layer. We only selectively discarded losses corresponding to the selected tokens. Therefore, the PU-learning approach can be easily extended to any token-level NER models. However, this is challenging to implement in loss functions with successive structures, such as CRF. We will discuss this in the following sections. Take the cross-entropy loss as example. After removing some potential unlabeled entities, the training loss becomes:

$$\left(\sum_i \sum_j -\log \boldsymbol{q}_j^{(i)}\left[y_j\right]\right) - \left(\sum_i \sum_{j \in \mathcal{A}_j} -\log \boldsymbol{q}_j^{(i)}\left[y_j\right]\right) \tag{2}$$

where $\boldsymbol{q}_j^{(i)} = \mathrm{Softmax}\left(\boldsymbol{h}_j^{(i)}\right)$.

In PU learning, we successfully identify and remove a portion of unlabeled entities, but this is not sufficient. In the next step, we will further enhance the algorithm using negative sampling. It is important to note that this negative sampling approach is applied after the initial exploratory training process. We will apply random sampling on the remaining tokens tagged as O. Inspired by Li et al. (2020c), we generate a candidate set of all potential unlabeled entities for each sentence $\boldsymbol{x}^{(i)}$:

$$\mathcal{L}_i = \left\{(j,k) | \forall j \le l \le k, y_l^{(i)} = O, l \notin \mathcal{A}_i\right\} \tag{3}$$

Then we randomly sample $\left\lceil \gamma * s_i \right\rceil$ spans from $\mathcal{L}_i$ and give these spans a special non-entity label. Here, γ is an important hyperparameter used to control the degree of sampling. The labels of these spans are replaced with the corresponding beginning-middle-end-single-outside (BMESO) or BIO tags, and the modified label sequences are defined as $\hat{\boldsymbol{y}}^{(i)}$. Ultimately, the dataset we used is $\hat{D} = \left\{ \left(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \mathcal{A}_i \right), i = 1, \ldots, N \right\}$. Taking cross-entropy loss as an example, our loss function then becomes:

$$\left(\sum_i \sum_j -\log \boldsymbol{q}_j^{(i)} \left[\widehat{y_j} \right] \right) - \left(\sum_i \sum_{j \in \mathcal{A}_j} -\log \boldsymbol{q}_j^{(i)} \left[y_j \right] \right) \tag{4}$$

The details of our method are presented in Algorithm 1. Note that given any token-level NER model, we can conduct exploratory training, remove O-tokens, apply negative sampling, and then retrain the model. As such, it is evident that the proposed method can be transferred into any token-level NER model. The practical models and scenarios will be applied in the Experiments section.

Li et al. (2020c) used negative sampling in a span-level NER model, which treated a span as the basic unit for labeling. Unlike Li et al. (2020c), we still treat the token as the basic unit. Thus, our method can be easily applied to many existing NER models since we modify only the training dataset and do no harm to the model. At the inference step, we treat the predicted non-entity spans as negative instances.

The proposed method is trained using the Adam optimizer. The hyperparameters that need to be determined include the initial training epoch n, the selection ratio », the sampling rate ³, and the learning rate ±. The optimal hyperparameters are determined based on the prediction accuracy in the validation set (if available). We also apply early stopping to avoid overfitting. The training process is terminated if the performance on the validation set does not improve across multiple epochs or if it reaches the maximum epoch.

Angle-Based Decision Vector

We note that the current decision vector used in our PU-learning algorithm is suboptimal. For instance, the decision vector we have always used is undefinable since it does not satisfy the sum-to-zero constraint, which will degrade the performance. Take $\boldsymbol{h}_j^{(i)}$ as an example. If we add a constant to each value of $\boldsymbol{h}_j^{(i)}$, the classification results remain unchanged, but the accuracy of detecting small O-values is significantly reduced. Thus, we would like to implement the proposed method using the decision vector with sum-to-zero constraint. However, adding such a constraint directly in the deep neural network would be challenging to optimize.

Note that a single decision vector can be used to separate two classes. Analogously, a $\left(k-1\right)$-dimensional decision vector should be sufficient for a k-category classification problem (Zhang & Liu, 2014). As a result, using a k-dimensional decision vector to match a k-category classification problem is redundant. To handle these difficulties, we will then introduce the angle-based technique. In the previous study of machine learning, Zhang and Liu (2014) proposed an angle-based technique to solve this problem, in which the angle-based classifiers can automatically satisfy the sum-to-zero constraint with fewer parameters.

To begin with, consider a centered simplex in $\mathbb{R}^{k-1}$ with elements:

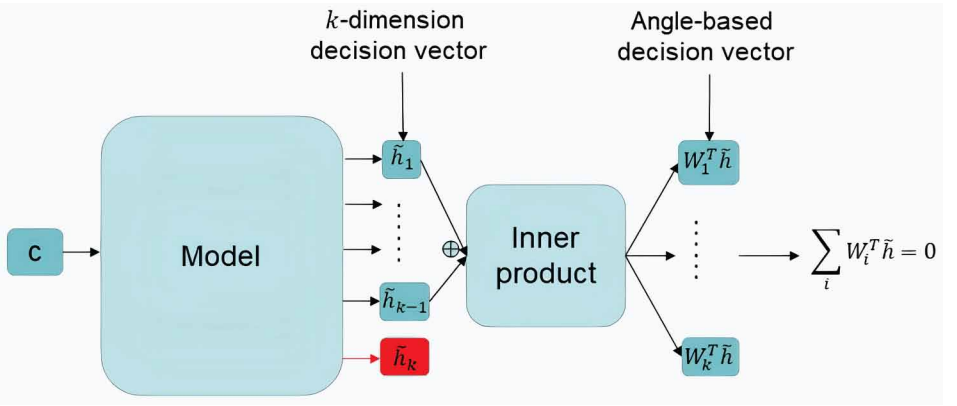
Algorithm 1. The two-step algorithm

Input: $\{\left(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right), i = 1,\ldots,N\}$, n, λ, γ, initial learning rate α

1: $\boldsymbol{f} \leftarrow$ Initialize the NER model
2: **for** $i \leftarrow 1$ to n **do**
3: $\hat{\boldsymbol{f}} \leftarrow$ Update NER model with original dataset $\{\left(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right), i = 1,\ldots,N\}$ with the Adam optimizer
4: Calculate O-values for all tokens using $\hat{\boldsymbol{f}}$ and generate the index set $\mathcal{A}_i$ using λ
5: Generate the candidate sets $\mathcal{L}_i$ for all sentences
6: Apply negative sampling using γ and generate $\left\{\left(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \mathcal{A}_i\right), i = 1,\ldots,N\right\}$
7: $\boldsymbol{f} \leftarrow$ Initialize the NER model
8: **while** Overall stopping criterion is not met **do**
9: $\boldsymbol{f}^* \leftarrow$ Update NER model using $\left\{\left(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \mathcal{A}_i\right), i = 1,\ldots,N\right\}$ with the Adam optimizer
10: Apply negative sampling using γ and generate $\left\{\left(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \mathcal{A}_i\right), i = 1,\ldots,N\right\}$
11: **end while**

Output: The optimal model $\boldsymbol{f}^*$

Figure 5. How the angle-based technique works



$$\mathbf{W}_j = \begin{cases} \left(k-1\right)^{-1/2} 1, & j = 1 \\ -\dfrac{1+\sqrt{k}}{\left(k-1\right)^{3/2}} 1 + \sqrt{\dfrac{k}{k-1}} e_{j-1}, & 2 \le j \le k \end{cases} \tag{5}$$

where $e_j \in \mathbb{R}^{k-1}$ is a vector of zeros, with the j th element as 1. We only require a $\left(k-1\right)$ -dimensional output to match the k -category classification in our tasks by the angle-based setting. For illustration,

we define the required output for one token as $\tilde{\boldsymbol{h}}$. Then we generate the k-dimensional angle-based decision vector by inner product, namely $\left(\tilde{\boldsymbol{h}}^T\mathbf{W}_1, \tilde{\boldsymbol{h}}^T\mathbf{W}_2, \ldots, \tilde{\boldsymbol{h}}^T\mathbf{W}_k\right)$. One can verify that the sum-to-zero constraint, $\sum_i \tilde{\boldsymbol{h}}^T\mathbf{W}_i$, is automatically satisfied. Moreover, the needed hidden outputs are $\left(k-1\right)$-dimensional, which can reduce the parameter size to a certain extent. Figure 5 shows the details about how the angle-based technique works. See Zhang et al. (2016, 2018), Yang et al. (2021), and Fu et al. (2022) for more implementations of the angle-based technique in large-margin classifiers. To show effectiveness, we conducted a study similar to that in the Preliminaries section. From Figure 6, we find that using an angle-based decision vector can achieve higher precision and recall in detecting unlabeled entities than the original.

Implementation of CRF

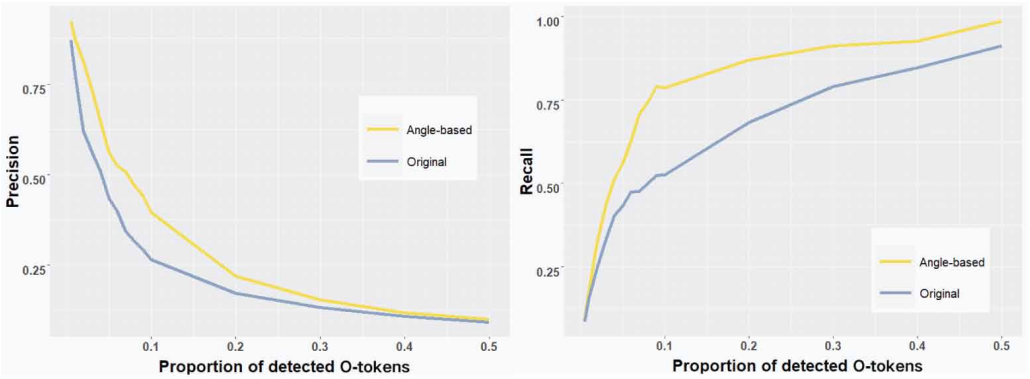
Due to the successive structure of the CRF layer, we cannot directly remove specific tokens in the first step. Thus, we will introduce how to implement our method in the CRF layer. Consider the decision vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n$ for sentence $\boldsymbol{x}$ with length n. For the original CRF, the probability of a label sequence $\boldsymbol{y} = \left\{y_1, \ldots, y_n\right\}$ is:

$$p\left(\boldsymbol{y} \mid \boldsymbol{x}\right) = \frac{\exp\left(\sum_i \left(\boldsymbol{h}_{i,y_i} + \boldsymbol{T}_{y_{i-1},y_i}\right)\right)}{\sum_{\tilde{\boldsymbol{y}}} \exp\left(\sum_i \left(\boldsymbol{h}_{i,\tilde{y}_i} + \boldsymbol{T}_{\tilde{y}_{i-1},\tilde{y}_i}\right)\right)} \tag{6}$$

where $\tilde{\boldsymbol{y}}$ represents an arbitrary label sequence and $\boldsymbol{T}_{y_{i-1},y_i}$ is the transition score from y_{i-1} to y_i. Utilizing the CRF structure, we remove the tokens by disregarding their scores, and the modified probability is:

$$\hat{p}\left(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{A}\right) = \frac{\exp\left(\sum_{i \notin \mathcal{A}} \left(\boldsymbol{h}_{i,y_i} + \boldsymbol{T}_{y_{i-1},y_i}\right)\right)}{\sum_{\tilde{\boldsymbol{y}}} \exp\left(\sum_{i \notin \mathcal{A}} \left(\boldsymbol{h}_{i,\tilde{y}_i} + \boldsymbol{T}_{\tilde{y}_{i-1},\tilde{y}_i}\right)\right)} \tag{7}$$

Figure 6. Precision (left) and recall (right) for detecting unlabeled entities



where $\mathcal{A}$ is the set of tokens deleted in the first step.

For training set $\hat{D} = \left\{\left(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \mathcal{A}_i\right), i = 1,\ldots,N\right\}$, we use the sentence-level log-likelihood loss to train the model, which is commonly applied in many scenarios (Daou et al., 2021):

$$L = -\sum_j \log\left(\hat{p}\left(\hat{\boldsymbol{y}}^{(j)} \middle| \boldsymbol{x}^{(j)}, \mathcal{A}_j\right)\right) \tag{8}$$

For inference, the Viterbi algorithm is employed to identify the best label sequence.

EXPERIMENTS

We conducted an extensive set of experiments on multiple classical NER models to investigate the effectiveness of our method. Experiments on real-world datasets and synthetic datasets are available. We will demonstrate that our method can be robust against unlabeled entities in the synthetic datasets. For real-world datasets we also achieve competitive performance. We use the standard F1-score (F1) as the evaluation metric.

Experimental Setup

Datasets

We conducted our experiments on six benchmark NER datasets, which were (1) Weibo (Peng & Dredze, 2015); (2) Resume (Zhang & Yang, 2018); (3) OntoNotes 4.0 (Weischedel et al., 2011); (4) MSRA (Levow, 2006); (5) EC (Yang et al., 2018); and (6) NEWS (Yang et al., 2018). The training data for EC and NEWS comprise two components: (1) a rigorously annotated set generated by human annotators; (2) a subset derived from distant supervision, marked by suboptimal annotations with abundant unlabeled entities. The evaluation datasets for both EC and NEWS are well-annotated. The statistics of the datasets are shown in Table A1. We constructed the synthetic datasets by contaminating two small datasets, Weibo and Resume. Specifically, we randomly selected at least 10% and up to 70% of entities in a training set and flipped their labels to O tags.

Baselines

To evaluate our method, we chose four different classical NER models:

- **Bi-LSTM:** A common Bi-LSTM+CRF (Huang et al., 2015) structure using the word2vec (Mikolov et al., 2013; Gu et al., 2022) embedding pretrained by Zhang and Yang (2018).
- **FLAT:** The Flat Lattice Transformer (Li et al., 2020b) using the same embedding as Bi-LSTM.
- **BERT+Word:** A strong BERT base model proposed by Liu et al. (2021), which uses a bilinear attention-weighted word vector as a supplement to the BERT input and uses LSTM and CRF as the fusion layer and inference layer, respectively.
- **LEBERT:** The recommended model in Liu et al. (2021), which is a combination of Lexicon Adapter and Transformer.

Overall, our choice of models was diverse, with two using BERT, two using Transformer, and one using just Bi-LSTM. For synthetic datasets, we report the F1-scores of each model with and without using our method. In addition, we carried out ablation studies to examine the contribution of the angle-based decision vector in synthetic Weibo NER. For real-world datasets, we compare the performance of BERT+Word and LEBERT using our method to their original procedure.

Hyperparameters

Recall that the parameter n is the number of epochs for the first training, λ stands for the proportion of removed tokens, and ratio γ represents the degree of negative sampling. We have found that the difference between unlabeled entities and true negative instances is huge in the first few epochs. Thus, the candidate set we used for n is $\{1,2,3\}$. Note that it is inappropriate to set λ as too small or too large values. Empirically speaking, we selected λ in $\{0.1\times 2^{-5}, 0.1\times 2^{-4},\ldots, 0.1\}$. We tended to use a large λ when the proportion of unlabeled entities increased. For real-world datasets, we always used the minimum value, 0.1×2^{-5}. To select the best parameter γ, we used the grid search in $\{0, 0.1, 0.2, 0.3\}$. Other hyperparameters were the same as the original method.

RESULTS

Synthetic Datasets

Tables 1–4 summarize the results of synthetic datasets. Figures 7 and 8 show the results of synthetic Weibo datasets, and the results for Resume can be found in the Appendix. For clarity reasons, /A indicates that we did not employ the angle-based decision vector. In general, each baseline method achieved better performance in handling unlabeled entities. For instance, when the proportion of unlabeled entities in Resume increased from 0.1 to 0.7, the F1-score of the original LEBERT model decreased by 90.67. After applying the proposed method, the F1-score decreased by only 3. This demonstrates the effectiveness of our method, even with a very few labeled entities. Furthermore, the ablation studies show that the angle-based decision vector is a beneficial addition for our method. Comparison of the results indicates that our method is more effective on BERT+Word and LEBERT than on FLAT and Bi-LSTM. This is likely because a strong NER model can acquire more precise underlying information in the first training, thus improving the performance of our method.

Table 1. The experimental results (F1-Score) for BERT+Word (Prob. represents the probability of labeled entities that be flipped to 'O' tags)

Prob.	BERT+Word				
	Weibo			Resume	
	Original	Ours	Ours/A	Original	Ours
0.1	64.85	67.03	66.67	94.96	95.66
0.2	60.95	66.36	65.39	94.56	95.36
0.3	56.24	65.39	64.87	94.18	95.18
0.4	53.08	65.09	63.82	92.88	95.09
0.5	49.10	63.88	62.73	64.78	94.45
0.6	43.29	63.63	61.31	13.24	94.13
0.7	31.75	63.38	59.68	1.63	85.10

Table 2. The experiment results (F1-Score) for LEBERT

Prob.	LEBERT				
	Weibo			Resume	
	Original	Ours	Ours/A	Original	Ours
0.1	65.80	69.34	68.28	94.50	95.16
0.2	65.31	68.75	67.45	94.16	94.91
0.3	62.31	68.17	67.04	93.17	94.95
0.4	56.76	67.71	66.59	91.11	94.46
0.5	54.37	65.73	65.07	70.07	94.02
0.6	43.87	64.32	63.92	43.54	93.27
0.7	29.34	62.88	61.19	3.82	92.40

Table 3. The experiment results (F1-Score) for FLAT

Prob.	FLAT				
	Weibo			Resume	
	Original	Ours	Ours/A	Original	Ours
0.1	57.38	59.97	59.62	95.10	95.27
0.2	57.18	59.68	59.10	94.95	95.22
0.3	53.90	59.00	57.99	94.81	95.14
0.4	49.25	56.09	55.12	94.16	94.70
0.5	47.30	53.53	53.25	89.52	92.50
0.6	46.46	49.85	49.45	57.89	72.39
0.7	42.14	49.19	48.82	24.29	68.58

Real-World Datasets

Table 5 shows the results for four well-annotated datasets. Specifically, we improved the F1-score in Weibo by 2.84% and Resume by 0.73%. For the LEBERT model, we achieved better results in Weibo, Resume, and MSRA. However, the F1-score on OntoNotes 4.0 is 0.12% worse than the original model. This demonstrates that our method can enhance performance in well-annotated datasets. There are two possible reasons that account for this. First, the well-annotated datasets may also have unlabeled entities, which can also lead to misguidance. Second, some true negative instances may also misguide the NER models. We will further discuss these factors in the following section.

The results for EC and NEWS are summarized in Table 6. As mentioned before, the two datasets are generated by distant supervision and thus have numerous unlabeled entities. In general, compared

Table 4. The experiment results (F1-Score) for Bi-LSTM

Prob.	Bi-LSTM				
	Weibo			Resume	
	Original	Ours	Ours/A	Original	Ours
0.1	47.73	49.47	48.67	93.96	94.13
0.2	46.32	48.63	46.71	93.93	94.07
0.3	41.91	45.03	44.78	93.10	93.72
0.4	33.58	37.75	36.56	92.01	93.02
0.5	28.68	35.10	33.22	83.53	87.66
0.6	19.54	27.52	27.26	39.81	47.85
0.7	8.24	12.76	9.83	5.33	17.85

Figure 7. Plots of the experimental results (F1-Score) on synthetic weibo datasets for BERT+Word (left) and LEBERT (right)

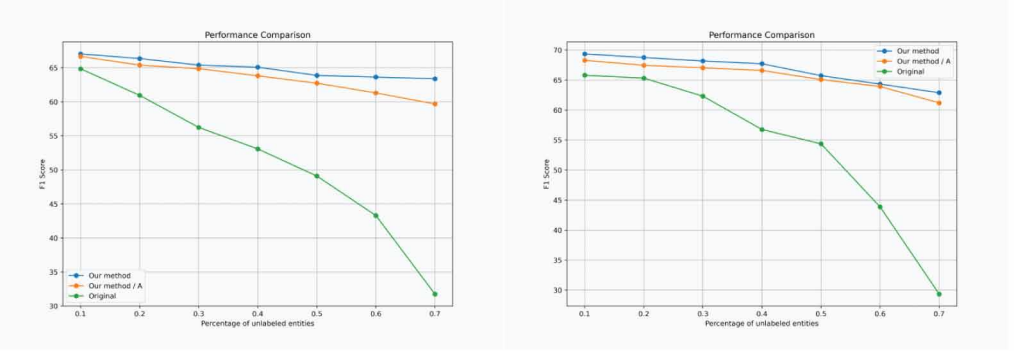
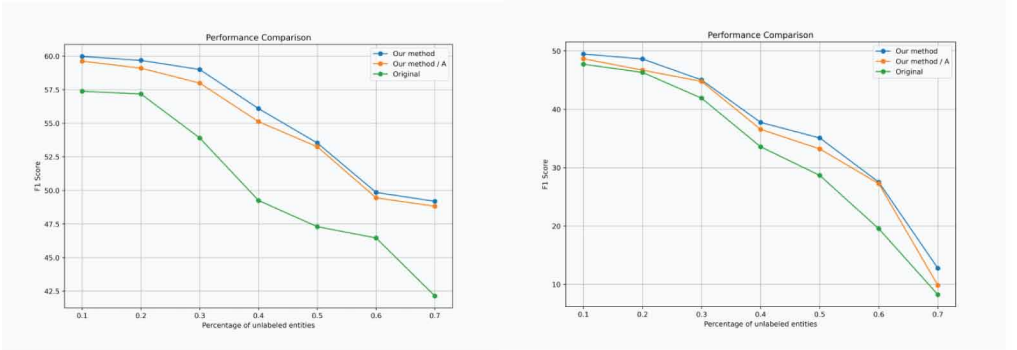


Figure 8. Plots of the experimental results (F1-Score) on synthetic weibo datasets for FLAT (Left) and Bi-LSTM (Right)



to the baseline models, we achieved significant improvement in both EC and NEWS datasets. For instance, the F1-score in EC is 70.18, compared to 61.75 (Partial CRF), 55.72 (Bert-MRC), 66.17

(negative sampling 1), and 67.03 (negative sampling 2). These results indicate that our method is not only robust on synthetic datasets, but is also competitive on the real-world datasets.

Compared to the non-robust baselines, we have achieved significant improvement in two kinds of synthetic datasets. This indicates that our method enhances robustness across various NER models. As stated before, the baselines we selected are diverse token-level NER models, demonstrating the versatility and scalability of our approach. In both well-annotated datasets and distant-supervision datasets, we also obtained competitive performance. This reflects the practical significance of our method.

Case Study

We have shown the validity of our method in real-world datasets. Note that the real-world datasets may contain very few or even no unlabeled entities. One natural question is how our method improves

Table 5. Real-World datasets results (F1-Score) for BERT+Word and LEBERT

Model	Weibo	Resume	OntoNotes 4.0	MSRA
Lattice LSTM (Zhang & Yang, 2018)	63.34	94.51	75.49	92.84
CAN (Zhu et al., 2019)	59.31	94.94	73.64	92.97
WC-LSTM (Liu et al., 2019)	65.30	94.49	75.79	93.50
SoftLexicon (Ma et al., 2019)	69.11	95.35	81.34	95.54
FLAT	68.07	95.78	80.56	95.46
BERT+Word	68.32	95.46	81.03	95.32
LEBERT	70.75	96.08	82.08	95.70
Our Bert+Word	70.26	96.16	81.34	95.44
Our LEBERT	71.00	96.24	81.98	95.73

Table 6. Real-World datasets results (F1-Score) for EC and NEWS

Model	EC	NEWS
Weighted Partial CRF (Jie et al., 2019)	61.75	78.64
Bert-MRC (Li et al., 2020a)	55.72	74.55
Negative Sampling (Li et al., 2020c)	66.17	85.39
Negative Sampling (Li et al., 2021)	67.03	86.15
BERT+Word	68.02	93.21
LEBERT	68.35	93.24
Our Bert+Word	69.72	93.88
Our LEBERT	70.18	94.05

the performance of BERT+Word and LEBERT. This is demonstrated by analyzing some removed tokens on Weibo. We show a portion of the removed tokens and the corresponding sentence fragments in Table 7. These removed tokens can be categorized into three types. To begin with, they might be entities from other undefined categories, such as 上海国际车站 and 卫生院. Second, they may be close to the existing entities, such as 何主席 and 沈太太. Note that substituting any other person-type entity for these tokens would still result in coherent sentences. Third, they may be fabricated or erroneous, such as 淘宝元 and 吴小. To conclude, such tokens may also misguide the NER model. Therefore, our method is also significant in real-world datasets.

Efficiency of Our Method

Table 8 reports the extra training time required after applying our method in LEBERT. Note that parameters λ and γ have a negligible effect on computational speed. Hence, we report the results only when n is varied. According to the table, our method adds no more than 8% additional training time when $n = 1$. If we use $n = 3$, the extra training time required can still be under 20%. Thus, the computational cost of our method is significantly small.

RELATED WORK

In recent years, we have witnessed the rapid advancement of deep learning and its successful applications (Sarivougioukas & Vagelatos, 2022; Lv et al., 2022; Zhou et al., 2022b; Jiao et al., 2023). NLP is a pivotal domain within deep learning and encompasses various developmental trajectories and applications (Ismail et al., 2022; Vats et al., 2023), such as text classification (Singh & Sachan, 2021; Miri et al., 2022), text-to-image synthesis (Chopra et al., 2022), and unsupervised information

Table 7. Examples of removed tokens

Sentence fragments	Removed tokens
上海国际车展	上海国际车展 Shanghai International Auto Show
卫生院的那个	卫生院 Health Center
探访一下何主席	何主席 Chairman He
想当沈太太	沈太太 Mrs. Shen
淘宝元专区	淘宝元 Tao Bao Yuan
吴小公民的微博	吴小 Wu Xiao

Table 8. The percentage of extra training time due to our method

n	Weibo	Resume	OntoNotes 4.0	MSRA
1	2.21%	5.15%	6.29%	7.31%
2	4.17%	11.97%	13.56%	14.47%
3	6.30%	15.73%	17.79%	19.72%

extraction (Sarkissian & Tekli, 2021; Hajjar & Tekli, 2022). Among these domains, knowledge graphs are an indispensable component and have extensive applications in various fields (Zhao et al., 2022; Zhou et al., 2022a; Li et al., 2023), such as health care and cybersecurity (Gou et al., 2017; Sahoo & Gupta, 2019).

NER is a crucial task in knowledge graphs and has received significant attention. In Chinese NER, leveraging the word information can significantly improve the performance. A possible strategy is to perform word segmentation first, followed by the NER task. Unfortunately, because the cross-domain word segmentation is still an unsolved problem (Liu & Zhang, 2012; Jiang et al., 2013; Liu et al., 2014; Qiu & Zhang, 2015; Chen et al., 2017; Huang et al., 2017), this strategy may result in error propagation. Another line of work is enhancing lexicon information in character-based models such as Lattice LSTM (Zhang & Yang, 2018), FLAT (Li et al., 2020b), LEBERT (Liu et al., 2021), and MCL (Zhao et al., 2023), which has demonstrated a significant benefit in merging the word information and preventing error propagation.

However, the NER models suffer from the unlabeled-entity problem in many scenarios (Zhang et al., 2020). Recently, numerous ways to address this issue have been proposed. Fuzzy CRF and AutoNER (Shang et al., 2018) handle the unlabeled entities by learning from high-quality phrases. Another approach for solving this problem involves the use of PU learning (Mayhew et al., 2019; Peng et al., 2019), which builds a distinct binary classifier to detect unlabeled entities. Partial CRF (Yang et al., 2018; Jie et al., 2019) is an extension of common CRF that allows NER models to learn from incomplete annotations. Li et al. (2020c) discovered that the primary cause of performance degradation is misguidance of unlabeled entities and employed a span-level negative sampling model to mitigate the misguidance. However, the current methods either rely entirely on the labeled data or do not encounter labeled data at all, which may be suboptimal for handling the unlabeled-entity problem.

CONCLUSION

In this work, we propose a two-step method to handle the unlabeled-entity problem. Our first step is based on the finding from empirical studies, which is proven to be effective in detecting unlabeled entities. The second step borrows the negative sampling method in the sequence labeling task, which is a helpful aid to the first step. As previously mentioned, methods that either entirely rely on or completely ignore annotated data are suboptimal. Hence, our method integrates the advantages of both and achieves superior results. Furthermore, we are the first to employ the angle-based technique in deep neural networks, which certainly enhances the effectiveness of our first step.

Compared to many existing methods, the proposed method is more scalable and efficient. Any token-level NER model can utilize our method to enhance its effectiveness and gain robustness in dealing with unlabeled entities. Similarly, our approach can be extended to other domains such as relation classification and text analysis, addressing issues of sparse or noisy annotations. Through empirical research, we have conducted an in-depth analysis of the noise-training process in NER models, which is beneficial for both this and future studies.

The experimental results demonstrate that the proposed method is effective in distant-supervised datasets and partially annotated datasets. This offers many meaningful applications. For example, in medical-document analysis, medical records and clinical reports often contain numerous new technical terms and entities. Moreover, in data sources such as news, scientific papers, and social media, the complexity of knowledge often makes it challenging to fully annotate all entities. In these cases, our model can be of significant assistance.

There may be multiple directions for future research. One interesting future direction is to extend our method to other scenarios or models. For example, the proposed method can be appropriately modified for the span-level NER models or to address incorrect entities in NER datasets. It can also

be of interest to further enhance the proposed PU-learning algorithm or develop more effective PU-learning algorithms. Note that the proposed approach is helpful in the NER domain. It may be desirable to apply the idea to other tasks such as relation classification and text analysis. Finally, it may be of interest to conduct more experimental analysis.

AUTHOR NOTE

The authors declare that there is no competing interest for this work.

This study was partially supported by Key R&D Program of Guangxi (2020AB10023), the National Natural Science Foundation of China (12171454 and U19B2040), and Fundamental Research Funds for the Central Universities.

The authors thank the editor and anonymous reviewers for their contributions toward improving the quality of this paper.

REFERENCES

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638–1649). Association for Computational Linguistics.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., & Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. In D. Precup and Y. W. The (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (pp. 233–242). JMLR. doi:10.48550/arXiv.1706.05394

Barbosa, A., Bittencourt, I. I., Siqueira, S. W., Dermeval, D., & Cruz, N. J. (2022). A context-independent ontological linked data alignment approach to instance matching. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–29. doi:10.4018/IJSWIS.295977

Chang, M. W., Ratinov, L., & Roth, D. (2007). Guiding semi-supervision with constraint-driven learning. In A. Zaenen, & A. van den Bosch (Eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 280–287). ACL.

Chen, X., Shi, Z., Qiu, X., & Huang, X. (2017). Adversarial multi-criteria learning for Chinese word segmentation. In R. Barzilay & M. Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (*Volume 1*: Long Papers)* (pp. 1193–1203). ACL. doi:10.18653/v1/P17-1110

Chopra, M., Singh, S. K., Sharma, A., & Gill, S. S. (2022). A comparative study of generative adversarial networks for text-to-image synthesis. [IJSSCI]. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–12. doi:10.4018/IJSSCI.300364

Daou, G., Maroun, C. B., & Hammoud, B. (2021). Advanced iterative multi-frequency algorithm used by radar remote-sensing systems for oil-spill thickness estimation. In *International Conference on Electrical, Computer and Energy Technologies (ICECET 2021)* (pp. 1–6). IEEE. doi:10.1109/ICECET52533.2021.9698742

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, *56*(2), 1–47. doi:10.1145/3604931

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 213–220). ACM. doi:10.1145/1401890.1401920

Fu, S., Chen, P., Liu, Y., & Ye, Z. (2022). Simplex-based multinomial logistic regression with diverging numbers of categories and covariates. *Statistica Sinica*. doi:10.5705/ss.202021.0082

Gou, Z., Yamaguchi, S., & Gupta, B. B. (2017). Analysis of various security issues and challenges in cloud computing environment: A survey. In Identity Theft: Breakthroughs in Research and Practice (pp. 221–247). IGI Global. doi:10.4018/978-1-5225-0808-3.ch011

Gu, J., Li, G., Vo, N. D., & Jung, J. J. (2022). Contextual Word2Vec model for understanding Chinese out of vocabularies on online social media. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–14. doi:10.4018/IJSWIS.309428

Gui, T., Ma, R., Zhang, Q., Zhao, L., Jiang, Y. G., & Huang, X. (2019). CNN-based Chinese NER with lexicon rethinking. In S. Kraus (Ed.), *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (pp. 4982–4988). IJCAI. doi:10.24963/ijcai.2019/692

Hajjar, A., & Tekli, J. (2022). Unsupervised extractive text summarization using frequency-based sentence clustering. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, & E. Zumpano (Eds.), *European Conference on Advances in Databases and Information Systems* (pp. 245–255). Springer. doi:10.1007/978-3-031-15743-1_23

Huang, S., Sun, X., & Wang, H. (2017). Addressing domain adaptation for Chinese word segmentation with global recurrent structure. In G. Kondrak & T. Watanabe (Eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing (*Volume 1*: Long Papers)* (pp. 184–193). Asian Federation of Natural Language Processing.

Ismail, S., Shishtawy, T. E., & Alsammak, A. K. (2022). A new alignment word-space approach for measuring semantic similarity for Arabic text. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–18. doi:10.4018/IJSWIS.297036

Jiang, W., Sun, M., Lü, Y., Yang, Y., & Liu, Q. (2013). Discriminative learning with natural annotations: Word segmentation as a case study. In H. Schuetze, P. Fund, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (*Volume 1*: Long Papers)* (pp. 761–769). ACL.

Jiao, R., Li, C., Xun, G., Zhang, T., Gupta, B. B., & Yan, G. (2023). A context-aware multi-event identification method for nonintrusive load monitoring. *IEEE Transactions on Consumer Electronics*, *69*(2), 194–204. doi:10.1109/TCE.2023.3236452

Jie, Z., Xie, P., Lu, W., Ding, R., & Li, L. (2019). Better modeling of incomplete annotations for named entity recognition. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 *(Long and Short Papers)* (pp. 729–734). ACL. doi:10.18653/v1/N19-1079

Jin, Y., Xie, J., Guo, W., Luo, C., Wu, D., & Wang, R. (2019). LSTM-CRF neural network with gated self attention for Chinese NER. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 136694–136703. doi:10.1109/ACCESS.2019.2942433

Levow, G. A. (2006). The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In H. T. Ng, & O. O. Y. Kwong (Eds.), *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing* (pp. 108–117). ACL.

Li, P., Zhou, G., Yin, Z., Chen, R., & Zhang, S. (2023). A semantically enhanced knowledge discovery method for knowledge graph based on adjacency fuzzy predicates reasoning. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *19*(1), 1–24. doi:10.4018/IJSWIS.323921

Li, X., Yan, H., Qiu, X., & Huang, X. (2020b). FLAT: Chinese NER using flat-lattice transformer. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6836–6842). ACL. doi:10.18653/v1/2020.acl-main.611

Li, Y., Liu, L., & Shi, S. (2022). Rethinking negative sampling for handling missing entity annotations. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (*Volume 1*: Long Papers)* (pp. 7188–7197). ACL. doi:10.18653/v1/2022.acl-long.497

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining* (pp. 179–186). IEEE. doi:10.1109/ICDM.2003.1250918

Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002). Partially supervised classification of text documents. In C. Sammut & A. G. Hoffmann (Eds.), *Machine learning: Proceedings of the 19th International Conference.* Morgan Kaufmann.

Liu, W., Fu, X., Zhang, Y., & Xiao, W. (2021). Lexicon enhanced Chinese sequence labeling using BERT adapter. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (*Volume 1*, Long Papers)* (pp. 5847–5858). ACL. doi:10.18653/v1/2021.acl-long.454

Liu, W., Xu, T., Xu, Q., Song, J., & Zu, Y. (2019). An encoding strategy based word-character LSTM for Chinese NER. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 *(Long and Short Papers)* (pp. 2379–2389). ACL. doi:10.18653/v1/N19-1247

Liu, Y., & Zhang, Y. (2012). Unsupervised domain adaptation for joint segmentation and POS-tagging. In *Proceedings of COLING 2012: Posters* (pp. 745–754). The COLING 2012 Organizing Committee.

Liu, Y., Zhang, Y., Che, W., Liu, T., & Wu, F. (2014). Domain adaptation for CRF-based Chinese word segmentation using free annotations. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 864–874). ACL. doi:10.3115/v1/D14-1093

Lv, L., Wu, Z., Zhang, L., Gupta, B. B., & Tian, Z. (2022). An edge-AI based forecasting approach for improving smart microgrid efficiency. *IEEE Transactions on Industrial Informatics*, *18*(11), 7946–7954. doi:10.1109/TII.2022.3163137

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (*Volume 1*: Long Papers)* (pp. 1064–1074). ACL. doi:10.18653/v1/P16-1101

Mayhew, S., Chaturvedi, S., Tsai, C. T., & Roth, D. (2019). Named entity recognition with partially annotated training data. In M. Bansal, & A. Villavicencio (Eds.), *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 645–655). ACL. doi:10.18653/v1/K19-1060

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems: Vol. 26. *Curran*.

Miri, M., Dowlatshahi, M. B., Hashemi, A., Rafsanjani, M. K., Gupta, B. B., & Alhalabi, W. (2022). Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, *37*(12), 11319–11341. doi:10.1002/int.23044

Peng, M., Xing, X., Zhang, Q., Fu, J., & Huang, X. (2019). Distantly supervised named entity recognition using positive-unlabeled learning. In A. Korhonen, D. Traum, & L. Márquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2409–2419). ACL. doi:10.18653/v1/P19-1231

Peng, N., & Dredze, M. (2015). Named entity recognition for Chinese social media with jointly trained embeddings. In L. Márquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 548–554). ACL. doi:10.18653/v1/D15-1064

Peng, S., Zhang, Y., Wang, Z., Gao, D., Xiong, F., & Zuo, H. (2021). Named entity recognition using negative sampling and reinforcement learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 714–719). IEEE. doi:10.1109/BIBM52615.2021.9669602

Qiu, L., & Zhang, Y. (2015). Word segmentation for Chinese novels. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(1). Advance online publication. doi:10.1609/aaai.v29i1.9523

Sahoo, S. R., & Gupta, B. B. (2019). Classification of various attacks and their defence mechanism in online social networks: A survey. *Enterprise Information Systems*, *13*(6), 832–864. doi:10.1080/17517575.2019.1605542

Sarivougioukas, J., & Vagelatos, A. (2022). Fused contextual data with threading technology to accelerate processing in home UbiHealth. [IJSSCI]. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–14. doi:10.4018/IJSSCI.285590

Sarkissian, S., & Tekli, J. (2021). Unsupervised topical organization of documents using corpus-based text analysis. In *Proceedings of the 13th International Conference on Management of Digital EcoSystems* (pp. 87–94). ACM. doi:10.1145/3444757.3485078

Shang, J., Liu, L., Gu, X., Ren, X., Ren, T., & Han, J. (2018). Learning named entity tagger using domain-specific dictionary. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2054–2064). ACL. doi:10.18653/v1/D18-1230

Shunxiang, Z., Aoqiang, Z., Guangli, Z., Zhongliang, W., & KuanChing, L. (2023). Building fake review detection model based on sentiment intensity and PU learning. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 6926–6939. doi:10.1109/TNNLS.2023.3234427 PMID:37018666

Singh, S. K., & Sachan, M. K. (2021). Classification of code-mixed bilingual phonetic text using sentiment analysis. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *17*(2), 59–78. doi:10.4018/IJSWIS.2021040104

Tekli, J., Tekli, G., & Chbeir, R. (2021). Almost linear semantic XML keyword search. In *Proceedings of the 13th International Conference on Management of Digital EcoSystems* (pp. 129–138). ACM. doi:10.1145/3444757.3485079

Tsuboi, Y., Kashima, H., Mori, S., Oda, H., & Matsumoto, Y. (2008). Training conditional random fields using incomplete annotations. In D. Scott & H. Uszkoreit (Eds.), *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 897–904). Coling 2008 Organizing Committee. doi:10.3115/1599081.1599194

Vats, T., Singh, S. K., Kumar, S., Gupta, B. B., Gill, S. S., Arya, V., & Alhalabi, W. (2023). Explainable context-aware IoT framework using human digital twin for healthcare. *Multimedia Tools and Applications*, 1–25. doi:10.1007/s11042-023-16922-5

Yang, Y., Chen, W., Li, Z., He, Z., & Zhang, M. (2018). Distantly supervised NER with partial annotation learning and reinforcement learning. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2159–2169). ACL.

Yang, Y., Guo, Y., & Chang, X. (2021). Angle-based cost-sensitive multicategory classification. *Computational Statistics & Data Analysis*, *156*, 107107. doi:10.1016/j.csda.2020.107107

Zhang, C., & Liu, Y. (2014). Multicategory angle-based large-margin classification. *Biometrika*, *101*(3), 625–640. doi:10.1093/biomet/asu017 PMID:26538663

Zhang, C., Liu, Y., Wang, J., & Zhu, H. (2016). Reinforced angle-based multicategory support vector machines. *Journal of Computational and Graphical Statistics*, *25*(3), 806–825. doi:10.1080/10618600.2015.1043010 PMID:27891045

Zhang, C., Pham, M., Fu, S., & Liu, Y. (2018). Robust multicategory support vector machines using difference convex algorithm. *Mathematical Programming*, *169*(4), 277–305. doi:10.1007/s10107-017-1209-5 PMID:29736090

Zhang, F., Ma, L., Wang, J., & Cheng, J. (2022). An MRC and adaptive positive-unlabeled learning framework for incompletely labeled named entity recognition. *International Journal of Intelligent Systems*, *37*(11), 9580–9597. doi:10.1002/int.23015

Zhang, H., Liu, L., Jiang, H., Li, Y., Zhao, E., Xu, K., Song, L., Zheng, S., Zhou, B., Zhu, J., Feng, X., Chen, T., Yang, T., Yu, D., Zhang, F., Kang, Z., & Shi, S. (2020). *TexSmart: A text understanding system for fine-grained NER and enhanced semantic analysis*. arXiv.

Zhang, Y., & Yang, J. (2018). Chinese NER using lattice LSTM. In I. Gurevynch & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (*Volume 1*, Long Papers)* (pp. 1554–1564). ACL. doi:10.18653/v1/P18-1144

Zhao, S., Wang, C., Hu, M., Yan, T., & Wang, M. (2023). MCL: Multi-granularity contrastive learning framework for Chinese NER. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, 37*(11), 14011–14019. ACL. doi:10.1609/aaai.v37i11.26640

Zhao, W., Huang, J., Fan, T., Wu, Y., & Liu, K. (2022). A novel compressed sensing-based graph isomorphic network for key node recognition and entity alignment. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(2), 1–17. doi:10.4018/IJSWIS.315600

Zhou, L., Li, J., Gu, Z., Qiu, J., Gupta, B. B., & Tian, Z. (2022a). PANNER: POS-aware nested named entity recognition through heterogeneous graph neural network. [early access]. *IEEE Transactions on Computational Social Systems*, 1–9. doi:10.1109/TCSS.2022.3159366

Zhou, Z., Li, Y., Li, J., Yu, K., Kou, G., Wang, M., & Gupta, B. B. (2022b). GAN-Siamese network for cross-domain vehicle re-identification in intelligent transport systems. *IEEE Transactions on Network Science and Engineering*, *10*(5), 1–12. doi:10.1109/TNSE.2022.3199919

APPENDIX

Table 9. The statistics of the datasets

Dataset	Type	Train	Dev	Test
Weibo	Sentence	1.4k	0.27k	0.27k
	Char	73.8k	14.5k	14.8k
Resume	Sentence	3.8k	0.46k	0.48k
	Char	124.1k	13.9k	15.1k
OntoNotes	Sentence	15.7k	4.3k	4.3k
	Char	491.9k	200.5k	208.1k
MSRA	Sentence	46.4k	–	4.4k
	Char	2169.9k	–	172.6k
EC	Sentence	3.66k	0.4k	0.8k
	Char	30.6k	3.1k	6.1k
NEWS	Sentence	6.72k	3.33k	3.19k
	Char	326.5k	149.0k	132.1k

Figure 9. Plots of the experimental results (F1-Score) on synthetic resume datasets for BERT+Word (left) and LEBERT (right)

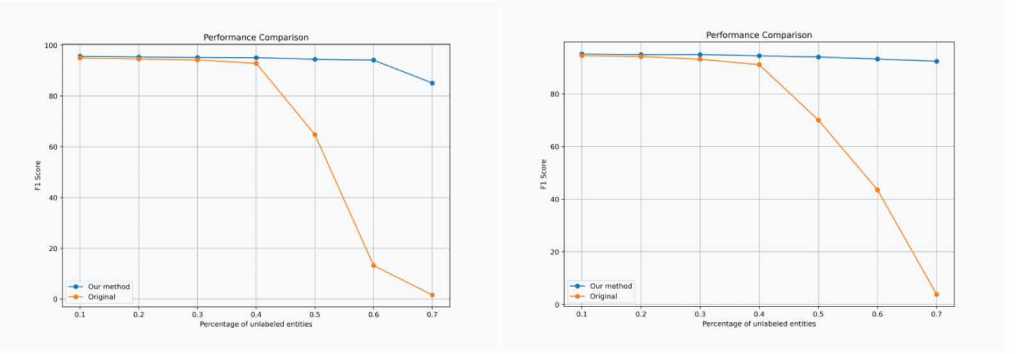
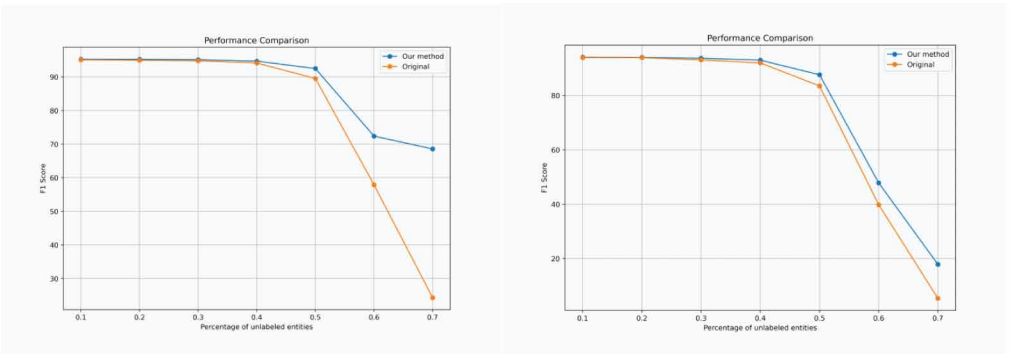


Figure 10. Plots of the experimental results (F1-Score) on synthetic resume datasets for FLAT (left) and Bi-LSTM (right)



Shunqin Zhang, a Ph.D student in University of Chinese Academy of Sciences. His research interests include natural language processing and machine learning.

Sanguo Zhang, a professor at University of Chinese Academy of Sciences. He graduated from University of Science and Technology of China in 2002 with a Ph.D degree. His research interests include high-dimensional statistical inference, heterogeneous data analysis.

Wenduo He, Engineer, Bachelor, Graduated from Huazhong University of Science and Technology in 2013. Worked in Tsinghua University. His research interests include Deep learning, natural language processing, Image and video analysis.

Xuan Zhang, an associate professor at Institute for Network Sciences and Cyberspace (INSC), Tsinghua University. He received the Ph.D. degree from the Department of Electronic Engineering at Tsinghua University in 2009 under the supervision of Prof. Xing Li. His research interests include image and video analysis, deep learning, natural language processing, and edge computing.