Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism

Pengfei Cao[1,2], **Yubo Chen**[1], **Kang Liu**[1,2], **Jun Zhao**[1,2] and **Shengping Liu**[3]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Beijing Unisound Information Technology Co., Ltd, Beijing, 100028, China
{pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn, liushengping@unisound.com

Abstract

Named entity recognition (NER) is an important task in natural language processing area, which needs to determine entities boundaries and classify them into pre-defined categories. For Chinese NER task, there is only a very small amount of annotated data available. Chinese NER task and Chinese word segmentation (CWS) task have many similar word boundaries. There are also specificities in each task. However, existing methods for Chinese NER either do not exploit word boundary information from CWS or cannot filter the specific information of CWS. In this paper, we propose a novel adversarial transfer learning framework to make full use of task-shared boundaries information and prevent the task-specific features of CWS. Besides, since arbitrary character can provide important cues when predicting entity type, we exploit self-attention to explicitly capture long range dependencies between two tokens. Experimental results on two different widely used datasets show that our proposed model significantly and consistently outperforms other state-of-the-art methods.

1 Introduction

The task of named entity recognition (NER) is to recognize the named entities in given text. NER is a preliminary and important task in natural language processing (NLP) area and can be used in many downstream NLP tasks, such as relation extraction (Bunescu and Mooney, 2005), event extraction (Chen et al., 2015) and question answering (Yao and Van Durme, 2014). In recent years, numerous methods have been carefully studied for NER task, including Hidden Markov Models (HMMs) (Bikel et al., 1997), Support Vector Machines (SVMs) (Isozaki and Kazawa, 2002) and Conditional Random Fields (CRFs) (Lafferty et al., 2001). Currently, with the development of deep learning, neural networks (Lample et al., 2016; Peng and Dredze, 2016; Luo and Yang, 2016) have been introduced to NER task. All these methods need to **determine entities boundaries** and **classify them into pre-defined categories**.

Task	Hilton 希尔顿			leaves 离开		Houston 休斯顿			Airport 机场	
Chinese NER	希	尔	顿	离	开	休	斯	顿	机	场
	B-PER	I-PER	I-PER	O	O	B-LOC	I-LOC	I-LOC	I-LOC	I-LOC
CWS	希	尔	顿	离	开	休	斯	顿	机	场
	B	I	E	S	S	B	I	E	B	E

Figure 1: An example of illustrating the similarities and specificities between Chinese NER and CWS.

Although great improvements have been achieved by these methods on Chinese NER task, some issues still have not been well addressed. One significant drawback is that there is only a very small amount of annotated data available. Weibo NER dataset (Peng and Dredze, 2015; He and Sun, 2017a) and Sighan2006 NER dataset (Levow, 2006) are two widely used datasets for Chinese NER task, containing 1.3k and 45k training examples, respectively. On the two datasets, the highest F1 scores are 48.41% and 89.21%, respectively. As a basic task in NLP area, the performance is not satisfactory. Fortunately, Chinese word segmentation (CWS) task is to recognize word boundaries and the amount of supervised training data for CWS is abundant compared with NER. There are many similarities between Chinese NER task and CWS task, which we call task-shared information. As shown in Figure 1, given a sentence "希尔顿离开休斯顿机场 (Hilton leaves Houston Airport)", the two tasks have the same boundaries for some words such as "希尔顿 (Hilton)" and "离开 (leaves)", while Chinese NER has more coarse-grained boundaries

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 182–192
Brussels, Belgium, October 31 - November 4, 2018.

than CWS task for certain word such as "休斯顿机场 (Houston Airport)" in the example of Figure 1, which we call task-specific information. In order to incorporate word boundary information from CWS task into NER task, Peng and Dredze (2016) propose a joint model that performs Chinese NER with CWS task. However, their proposed model only focuses on task-shared information between Chinese NER and CWS, and ignores filtering the specificities of each task, which will bring noise for both of the tasks. For example, the CWS task splits "休斯顿机场 (Houston Airport)" into "休斯顿 (Houston)" and "机场 (Airport)", while the NER task takes "休斯顿机场 (Houston Airport)" as a whole entity. Thus, how to exploit task-shared information and prevent the noise brought by CWS task to Chinese NER task is a challenging problem.

Another issue is that most proposed models cannot explicitly model long range dependencies when predicting entity type. Though bidirectional long short term memory (BiLSTM) can learn long-distance dependencies, it cannot conduct direct connections between arbitrary two characters. As shown in Figure 1, if the model only focuses on the word "希尔顿 (Hilton)", it can be a person or organization. However, when the model explicitly captures the dependencies between "希尔顿 (Hilton)" and "离开 (leaves)", it is easy to classify "希尔顿 (Hilton)" into "person" category. Context information is very crucial for determining the entity type. While in the sentence "我将住在希尔顿 (I will be staying at the Hilton)", the entity type of "希尔顿 (Hilton)" is "organization". Thus, how to better capture the global dependencies of the whole sentence is another challenging problem.

To address the above problems, we propose an adversarial transfer learning framework to integrate the task-shared word boundary information into Chinese NER task in this paper. The adversarial transfer learning is incorporating adversarial training into transfer learning. To better capture long range dependencies and synthesize the information of the sentence, we extend self-attention mechanism into the framework. Specifically, we try to improve Chinese NER task performance by incorporating shared boundary information from CWS task. To prevent the specific information of CWS task from lowering the performance of the Chinese NER task, we introduce adversarial training to ensure that the Chinese NER task only exploits task-shared word boundary information. Then, for tackling the long range dependency problems, we utilize self-attention to synthesize the hidden representation of BiLSTM. Finally, we evaluate our model on two different widely used Chinese NER datasets. Experimental results show that our proposed model achieves better performance than other state-of-the-art methods and gains new benchmarks.

In summary, the contributions of this paper are as follows:

- We propose an adversarial transfer learning framework to incorporate task-shared word boundary information from CWS task into Chinese NER task. To our best knowledge, it is the first work to apply adversarial transfer learning method into NER task.
- We introduce self-attention mechanism into our model, which aims to capture the global dependencies of the whole sentence and learn inner structure features of sentence.
- We conduct our experiment on two different widely used Chinese NER datasets, and the experimental results demonstrate that our proposed model significantly and consistently outperforms previous state-of-the-art methods. We release the source code publicly for further research[1].

2 Related Work

NER Many methods have been proposed for NER task. Early studies on NER often exploit SVMs (Isozaki and Kazawa, 2002), HMMs (Bikel et al., 1997) and CRFs (Lafferty et al., 2001), heavily relying on feature engineering. Zhou et al. (2013) formulate Chinese NER as a joint identification and categorization task. In recent years, neural network models have been introduced to NER task (Collobert et al., 2011; Huang et al., 2015; Peng and Dredze, 2016). Huang et al. (2015) exploit BiLSTM to extract features and feed them into CRF decoder. After that, the BiLSTM-CRF model is usually exploited as the baseline. Lample et al. (2016) use a character LSTM to represent spelling characteristics. In addition, Wang et al. (2017) propose a gated convolutional neural network (GCNN) model for Chinese NER. Peng and Dredze (2016) propose a joint model for Chinese

[1] https://github.com/CPF-NLPR/AT4ChineseNER

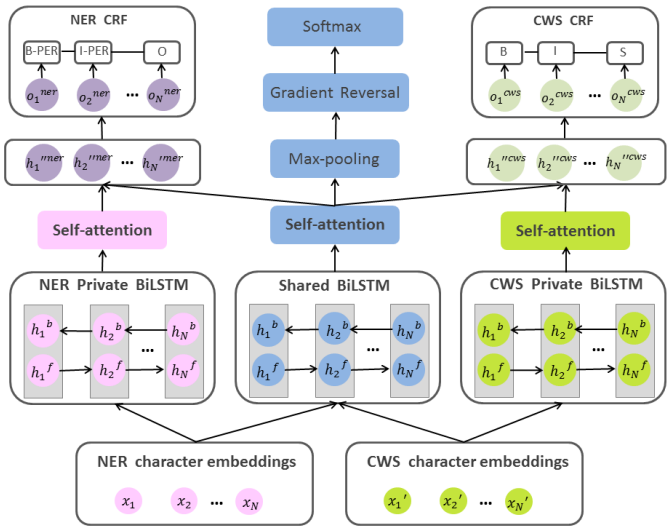

Figure 2: The general architecture of our proposed model. The left and right part are Chinese NER space and CWS private space, respectively, including embedding layer, feature extractor (Private BiLSTM), self-attention and CRF layer. The middle part is shared space consisting of feature extractor (Shared BiLSTM), self-attention and task discriminator.

NER, which are jointly trained with CWS task. However, the specific features brought by CWS task can lower the performance of the Chinese NER task.

Adversarial Training Adversarial networks have achieved great success in computer vision (Goodfellow et al., 2014; Denton et al., 2015). In NLP area, adversarial training has been introduced for domain adaptation (Ganin and Lempitsky, 2014; Zhang et al., 2017; Gui et al., 2017), cross-lingual transfer learning (Chen et al., 2016; Kim et al., 2017), multi-task learning (Chen et al., 2017; Liu et al., 2017) and crowdsourcing learning (Yang et al., 2018). Bousmalis et al. (2016) propose shared-private model in domain separation network. Different from these works, we exploit adversarial network to jointly train Chinese NER task and CWS task, aiming to extract task-shared word boundary information from CWS task. To our knowledge, it is the first work to apply adversarial transfer learning framework to Chinese NER task.

Self-Attention Self-attention has been introduced to machine translation by Vaswani et al. (2017) for capturing global dependencies between input and output and achieves state-of-the-art performance. For language understanding task, Shen et al. (2017) exploit self-attention to learn long range dependencies. Tan et al. (2017) apply self-attention to semantic role labelling task and achieve state-of-the-art results. We are the first to introduce self-attention mechanism to Chinese NER task.

3 Method

In this paper, we propose a novel adversarial transfer learning framework that will learn task-shared word boundary information from CWS task, filter specific information of CWS and explicitly capture the long range dependencies between arbitrary two characters in sentence. The architecture of our proposed model is illustrated in Figure 2. The model mainly consists of five components: embedding layer, shared-private feature extractor, self-attention, task-specific CRF and task discriminator. In the following sections, we will describe each part of our proposed model in detail.

3.1 Embedding Layer

Similar to other neural network models, the first step of our proposed model is to map discrete characters into the distributed representations. For a given Chinese sentence $\mathbf{x} = \{c_1, c_2, \ldots, c_N\}$ from Chinese NER dataset or CWS dataset, we lookup embedding vector from pre-trained embedding matrix for each character c_i as $\mathbf{x}_i \in \mathbb{R}^{d_e}$.

3.2 Shared-Private Feature Extractor

Long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a variant of recurrent neural network (RNN) (Elman, 1990), which enables to address the gradient vanishing and exploding

problems in RNN via introducing gate mechanism and memory cell. The unidirectional LSTM only leverages information from the past, ignoring the future information. In order to incorporate information from both sides of sequence, we adopt BiLSTM to extract features. Specially, the hidden state of BiLSTM could be expressed as follows:

$$\overrightarrow{\mathbf{h}_i} = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{x}_i) \quad (1)$$

$$\overleftarrow{\mathbf{h}_i} = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{\mathbf{h}}_{i+1}, \mathbf{x}_i) \quad (2)$$

$$\mathbf{h}_i = \overrightarrow{\mathbf{h}_i} \oplus \overleftarrow{\mathbf{h}_i} \quad (3)$$

where $\overrightarrow{\mathbf{h}_i} \in \mathbb{R}^{d_h}$ and $\overleftarrow{\mathbf{h}_i} \in \mathbb{R}^{d_h}$ are the hidden states of the forward and backward LSTM at position i, respectively. $\oplus$ denotes concatenation operation.

As shown in Figure 2, we propose a shared-private feature extractor, which assigns a private BiLSTM layer and shared BiLSTM layer for task $k \in \{NER, CWS\}$. The private BiLSTM layer is used to extract task-specific features, and the shared BiLSTM layer is used to learn task-shared word boundaries. Formally, for any sentence in dataset of task k, the hidden states of shared and private BiLSTM layer can be computed as follows:

$$\mathbf{s}_i^k = \mathrm{BiLSTM}(\mathbf{x}_i^k, \mathbf{s}_{i-1}^k; \theta_s) \quad (4)$$

$$\mathbf{h}_i^k = \mathrm{BiLSTM}(\mathbf{x}_i^k, \mathbf{h}_{i-1}^k; \theta_k) \quad (5)$$

where θ_s and θ_k are the shared BiLSTM parameters and private BiLSTM parameters of task k, respectively.

3.3 Self-Attention

Inspired by the self-attention applied to machine translation (Vaswani et al., 2017) and semantic role labelling (Tan et al., 2017), we exploit self-attention to explicitly learn the dependencies between any two characters in sentence and capture the inner structure information of sentence. In this paper, we adopt the multi-head self-attention mechanism. $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$ denotes the output of private BiLSTM. Correspondingly, $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ is the output of shared BiLSTM. We will take the self-attention in private space as example to illustrate how it works. The scaled dot-product attention can be precisely described as follows:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{\mathbf{d}}})\mathbf{V} \quad (6)$$

where $\mathbf{Q} \in \mathbb{R}^{N \times 2d_h}$, $\mathbf{K} \in \mathbb{R}^{N \times 2d_h}$ and $\mathbf{V} \in \mathbb{R}^{N \times 2d_h}$ are query matrix, keys matrix and value matrix, respectively. In our setting, $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{H}$. $\mathbf{d}$ is the dimension of hidden units of BiLSTM, which equals to $2d_h$.

Multi-head attention first linearly projects the queries, keys and values h times by using different linear projections. Then h projections perform the scaled dot-product attention in parallel. Finally, these results of attention are concatenated and once again projected to get the new representation. Formally, the multi-head attention can be expressed as follows:

$$head_i = \mathrm{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (7)$$

$$\mathbf{H}^{'} = (head_i \oplus \ldots \oplus head_h)\mathbf{W}_o \quad (8)$$

where $\mathbf{W}_i^Q \in \mathbb{R}^{2d_h \times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{2d_h \times d_k}$ and $\mathbf{W}_i^V \in \mathbb{R}^{2d_h \times d_k}$ are trainable projection parameters and $d_k = 2d_h/h$. $\mathbf{W}_o \in \mathbb{R}^{2d_h \times 2d_h}$ is also trainable parameter.

3.4 Task-Specific CRF

For a sentence in dataset of task k, we compute the final representation via concatenating the representations from private space and shared space after self-attention layer:

$$\mathbf{H}^{''k} = \mathbf{H}^{'k} \oplus \mathbf{S}^{'k} \quad (9)$$

where $\mathbf{H}^{'k}$ and $\mathbf{S}^{'k}$ are the outputs of private self-attention and shared self-attention of task k, respectively.

Considering the dependencies between successive labels, we exploit CRF (Lafferty et al., 2001) to inference tags instead of making tagging decisions using $\mathbf{h}_i^{''}$ independently. Due to the difference of labels, we introduce a specific CRF layer for each task. Given a sentence $\mathbf{x} = \{c_1, c_2, \ldots, c_N\}$ with a predicted tag sequence $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$, the CRF tagging process can be formalized as follows:

$$\mathbf{o}_i = \mathbf{W}_s\mathbf{h}_i^{''} + \mathbf{b}_s \quad (10)$$

$$s(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N}(\mathbf{o}_{i,y_i} + \mathbf{T}_{y_{i-1},y_i}) \quad (11)$$

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathbf{Y}_x} s(\mathbf{x}, \mathbf{y}) \quad (12)$$

where $\mathbf{W}_s \in \mathbb{R}^{|T| \times 4d_h}$ and $\mathbf{b}_s \in \mathbb{R}^{|T|}$ are trainable parameters. $|T|$ denotes the number of output labels. $\mathbf{o}_{i,y_i}$ represents the score of the y_i-th tag

of the character c_i. $\mathbf{T}$ is a transition score matrix which defines the scores of two successive labels. $\mathbf{Y}_x$ represents all candidate tag sequences for given sentence $\mathbf{x}$. In decoding, we use Viterbi algorithm to get the predicted tag sequence $\bar{\mathbf{y}}$.

For training, we exploit negative log-likelihood objective as the loss function. The probability of the ground-truth label sequence is computed by:

$$p(\hat{\mathbf{y}}|\mathbf{x}) = \frac{e^{s(\mathbf{x},\hat{\mathbf{y}})}}{\sum_{\tilde{\mathbf{y}}\in\mathbf{Y}_x} e^{s(\mathbf{x},\tilde{\mathbf{y}})}} \tag{13}$$

where $\hat{\mathbf{y}}$ denotes the ground-truth label sequence. Given T training examples $(\mathbf{x}^{(i)}; \hat{\mathbf{y}}^{(i)})$, the loss function L_{Task} can be defined as follows:

$$L_{Task} = -\sum_{i=1}^{T} \log p(\hat{\mathbf{y}}^{(i)}|\mathbf{x}^{(i)}) \tag{14}$$

We use gradient back-propagation method to minimize the loss function.

3.5 Task Discriminator

Inspired by adversarial networks (Goodfellow et al., 2014), we incorporate adversarial training into shared space to guarantee that specific features of tasks do not exist in shared space. We propose a task discriminator to estimate which task the sentence comes from. Formally, the task discriminator can be expressed as follows:

$$\mathbf{s}'^{k} = \text{Maxpooling}(\mathbf{S}'^{k}) \tag{15}$$

$$D(\mathbf{s}'^{k}; \theta_d) = \text{softmax}(\mathbf{W}_d \mathbf{s}'^{k} + \mathbf{b}_d) \tag{16}$$

where θ_d indicates the parameters of task discriminator. $\mathbf{W}_d \in \mathbb{R}^{K\times 2d_h}$ and $\mathbf{b}_d \in \mathbb{R}^K$ are trainable parameters. K is the number of tasks.

Besides the task loss L_{Task}, we introduce an adversarial loss L_{Adv} to prevent specific features of CWS task from creeping into shared space. The adversarial loss trains the shared model to produce shared features such that the task discriminator cannot reliably recognize which task the sentence comes from. The adversarial loss can be computed as follows:

$$L_{Adv} = \min_{\theta_s}(\max_{\theta_d} \sum_{k=1}^{K}\sum_{i=1}^{T_k} \log D(E_s(\mathbf{x}_k^{(i)}))) \tag{17}$$

where θ_s denotes the trainable parameters of shared BiLSTM. E_s denotes the shared feature extractor. T_k is the number of training examples of task k. $\mathbf{x}_k^{(i)}$ is the i-th example of task k. There is a minimax optimization that the shared BiLSTM generates a representation to mislead the task discriminator and the discriminator tries its best to correctly determine the type of task.

We add a gradient reversal layer (Ganin and Lempitsky, 2014) below the softmax layer to address the minimax optimization problem. In the training phrase, we minimize the task discriminator errors, and through gradient reversal layer the gradients will become opposed sign to adversarially encourage the shared feature extractor to learn task-shared word boundary information. After training phrase, the shared feature extractor and task discriminator reach a point where the discriminator cannot differentiate the tasks according to the representations learned from shared feature extractor.

3.6 Training

The final loss function of our proposed model can be written as follows:

$$L = L_{NER} \cdot I(\mathbf{x}) + L_{CWS} \cdot (1 - I(\mathbf{x})) + \lambda L_{Adv} \tag{18}$$

where λ is a hyper-parameter. L_{NER} and L_{CWS} can be computed via Eq.14. $I(\mathbf{x})$ is a switching function to identify which task the input comes from. It is defined as follows:

$$I(\mathbf{x}) = \begin{cases} 1, & if \quad \mathbf{x} \in \mathcal{D}_{NER} \\ 0, & if \quad \mathbf{x} \in \mathcal{D}_{CWS} \end{cases} \tag{19}$$

where $\mathcal{D}_{NER}$ and $\mathcal{D}_{CWS}$ are Chinese NER training corpora and CWS training corpora, respectively.

In the training phrase, at each iteration, we first select a task from $\{NER, CWS\}$ in turn. Then, we sample a batch of training instances from the given task to update the parameters. We use Adam (Kingma and Ba, 2014) algorithm to optimize the final loss function. Since Chinese NER task and CWS task may have different convergence rate, we repeat the above iterations until early stopping according to the Chinese NER task performance.

4 Experiments

4.1 Datasets

To evaluate our proposed model on Chinese NER, we experiment on two different widely used datasets, including Weibo NER dataset (WeiboNER) (Peng and Dredze, 2015; He and Sun,

Dataset	Task	# Train sent	# Dev sent	# Test sent
WeiboNER	Chinese NER	1350	270	270
SighanNER	Chinese NER	41728	4636	4365
MSR	CWS	86924	—	3985

Table 1: Statistics of the datasets.

Models	P(%)	R(%)	F1(%)
CRF (Peng and Dredze, 2015)	56.98	25.26	35.00
CRF+word (Peng and Dredze, 2015)	64.94	25.77	36.90
CRF+character (Peng and Dredze, 2015)	57.89	34.02	42.86
CRF+character+position (Peng and Dredze, 2015)	57.26	34.53	43.09
Joint(cp) (main) (Peng and Dredze, 2015)	57.98	35.57	44.09
Pipeline Seg.Repr.+NER (Peng and Dredze, 2016)	64.22	36.08	46.20
Jointly Train Char.Emb (Peng and Dredze, 2016)	63.16	37.11	46.75
Jointly Train LSTM Hidden (Peng and Dredze, 2016)	63.03	38.66	47.92
Jointly Train LSTM+Emb (main) (Peng and Dredze, 2016)	63.33	39.18	48.41
BiLSTM+CRF+adversarial+self-attention	55.72	**50.68**	**53.08**

Table 2: NER results for named entities on the original WeiboNER dataset (Peng and Dredze, 2015). There are three blocks. The first two blocks contain the main and simplified models proposed by Peng and Dredze (2015) and Peng and Dredze (2016), respectively. The last block lists the performance of our proposed model.

2017a) and SIGHAN2006 NER dataset (SighanNER) (Levow, 2006). We use the MSR dataset (from SIGHAN2005) for CWS task.

The WeiboNER is annotated with four entity types (person, location, organization and geo-political entities), including named entities and nominal mentions. The SighanNER is simplified Chinese, which contains three entity types (person, location and organization). For WeiboNER, we use the same training, development and testing splits as Peng and Dredze (2015). Since the SighanNER does not have development set, we sample 10% data of training set as development set. We use MSR dataset to improve the performance of the Chinese NER task. Table 1 gives the details of the three datasets.

4.2 Settings

For evaluation, we use the Precision (P), Recall (R) and F1 score as metrics in our experiment.

For hyper-parameter configurations, we adjust them according to the performance on development set of Chinese NER task. We set the character embedding size d_e to 100. The dimensionality of LSTM hidden states d_h is 120. The initial learning rate is set to 0.001. The loss weight coefficient λ is set to 0.06. We set the dropout rate to 0.3. The number of projections h is 8. We set the batch size of SighanNER and WeiboNER as 64 and 20, respectively.

For trainable parameters initialization, we use xavier initializer (Glorot and Bengio, 2010) to initialize parameters. The character embeddings used in our experiment are pre-trained on Baidu Encyclopedia corpus and Weibo corpus by using word2vec toolkit (Mikolov et al., 2013).

4.3 Compared with State-of-the-art Methods

In this section, we will give the experimental results of our proposed model and previous state-of-the-art methods on WeiboNER dataset and SighanNER dataset, respectively.

4.3.1 Evaluation on WeiboNER

We compare our proposed model with the latest models on WeiboNER dataset. Table 2 shows the experimental results for named entities on the original WeiboNER dataset.

In the first block of Table 2, we give the performance of the main model and baselines proposed by Peng and Dredze (2015). They propose a CRF-based model to jointly train the embeddings with NER task, which achieves better results than pipeline models. In addition, they consider the po-

Models	Named Entity			Nominal Mention			Overall
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	F1(%)
Peng and Dredze (2015)	74.78	39.81	51.96	71.92	53.03	61.05	56.05
Peng and Dredze (2016)	66.67	47.22	55.28	74.48	54.55	62.97	58.99
He and Sun (2017a)	66.93	40.67	50.60	66.46	53.57	59.32	54.82
He and Sun (2017b)	61.68	48.82	54.50	74.13	53.54	62.17	58.23
BiLSTM+CRF+adv+self-attention	59.51	**50.00**	54.34	71.43	47.90	57.35	58.70

Table 3: Experimental results on the updated WeiboNER dataset (He and Sun, 2017a). There are two blocks. The first block is the performance of latest models. The second block reports the performance of our proposed model. With the limited length of the page, we use "adv" to denote "adversarial".

Models	P(%)	R(%)	F1(%)
Chen et al. (2006)	91.22	81.71	86.20
Zhou et al. (2006)	88.94	84.20	86.51
Luo and Yang (2016)	91.30	87.22	89.21
BiLSTM+CRF+adversarial+self-attention	**91.73**	**89.58**	**90.64**

Table 4: Results on SighanNER dataset. There are two blocks. The first block reports the result of previous methods. The second block gives the performance of our proposed model.

sition of each character in a word to train character and position embeddings.

In the second block of Table 2, we report the performance of the main model and baselines proposed by Peng and Dredze (2016). Aiming to incorporate word boundary information into the NER task, they propose an integrated model that can joint training CWS task, improving the F1 score from 46.20% to 48.41% as compared with pipeline model (Pipeline Seg.Repr.+NER).

In the last block of Table 2, we give the experimental result of our proposed model (BiLSTM+CRF+adversarial+self-attention). We can observe that our proposed model significantly outperforms other models. Compared with the model proposed by Peng and Dredze (2016), our method gains 4.67% improvement in F1 score. Interestingly, WeiboNER dataset and MSR dataset are different domains. The WeiboNER dataset is social media domain, while the MSR dataset can be regard as news domain. The improvement of performance indicates that our proposed adversarial transfer learning framework may not only learn task-shared word boundary information from CWS task but also tackle the domain adaptation problem.

We also conduct an experiment on the updated WeiboNER dataset. Table 3 lists the performance of the latest models and our proposed model on the updated dataset. In the first block of Table 3, we report the performance of the latest models. The model proposed by Peng and Dredze (2015) achieves F1 score of 56.05% on overall performance. He and Sun (2017b) propose an unified model for Chinese NER task to exploit the data from out-of-domain corpus and in-domain unlabelled texts. The unified model improves the F1 score from 54.82% to 58.23% compared with the model proposed by He and Sun (2017a).

In the second block of Table 3, we give the result of our proposed model. It can be observed that our proposed model achieves a very competitive performance. Compared with the latest model proposed by He and Sun (2017b), our model improves the F1 score from 58.23% to 58.70% on overall performance. The improvement demonstrates the effectiveness of our proposed model.

4.3.2 Evaluation on SighanNER

Table 4 lists the comparisons on SighanNER dataset. We observe that our proposed model achieves new state-of-the-art performance.

In the first block, we give the performance of previous methods for Chinese NER task on SighanNER dataset. Chen et al. (2006) propose a character-based CRF model for Chinese NER task. Zhou et al. (2006) introduce a pipeline model, which first segments the text with character-level CRF model and then applies word-level CRF to tag. Luo and Yang (2016) first train a word segmenter and then use word segmentation as addi-

Models	SighanNER			WeiboNER		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BiLSTM+CRF	89.84	88.42	89.13	58.99	44.93	51.01
BiLSTM+CRF+transfer	90.60	89.19	89.89	60.00	46.03	52.09
BiLSTM+CRF+adversarial	90.52	89.56	90.04	61.94	45.48	52.45
BiLSTM+CRF+self-attention	90.62	88.81	89.71	57.81	47.67	52.25
BiLSTM+CRF+adversarial+self-attention	**91.73**	**89.58**	**90.64**	55.72	**50.68**	**53.08**

Table 5: Comparison between our proposed model and simplified models on SighanNER dataset and original WeiboNER dataset.

Task	Mom and dad			miss	you
	爸爸		妈妈	想	你们
CWS	爸 爸		妈 妈	想	你 们
	B	E	B E	S	B E
Baselines	爸 爸		妈 妈	想	你 们
	B-PER	I-PER	I-PER I-PER	O	O O
Ours	爸 爸		妈 妈	想	你 们
	B-PER	I-PER	B-PER I-PER	O	O O
Golds	爸 爸		妈 妈	想	你 们
	B-PER	I-PER	B-PER I-PER	O	O O

(a) Example for the effectiveness of boundary information.

Task	The boss	thinks	you	don't	respect	him
	上司	会以为	你	不	尊重	他
CWS	上 司	会	以 为	你	不	尊 重	他
	B E	S	B E	S	S	B E	S
Baselines	上 司	会	以 为	你	不	尊 重	他
	O O	O	O O	O	O	O O	O
Ours	上 司	会	以 为	你	不	尊 重	他
	B-PER I-PER	O	O O	O	O	O O	O
Golds	上 司	会	以 为	你	不	尊 重	他
	B-PER I-PER	O	O O	O	O	O O	O

(b) Example for the effectiveness of self-attention.

Figure 3: The analysis of Chinese NER cases from WeiboNER dataset.

tional features for sequence tagging. Although the model achieves competitive performance, giving the F1 score of 89.21%, it suffers from the error propagation problem.

In the second block, we report the result of our proposed model. Compared with the state-of-the-art model proposed by Luo and Yang (2016), our method improves the F1 score from 89.21% to 90.64% without any additional features, which demonstrates the effectiveness of our proposed model.

4.4 Effectiveness of Adversarial Transfer Learning and Self-Attention

Table 5 provides the experimental results of our proposed model and baseline as well as its simplified models on SighanNER dataset and WeiboNER dataset. The simplified models are described as follows:

- BiLSTM+CRF: The model is used as strong baseline in our work, which is trained using Chinese NER training data.
- BiLSTM+CRF+transfer: We apply transfer learning to BiLSTM+CRF model without adversarial loss and self-attention mechanism.
- BiLSTM+CRF+adversarial: Compared with BiLSTM+CRF+transfer model, the BiLSTM+CRF+adversarial model incorporates adversarial training.
- BiLSTM+CRF+self-attention: The model integrates the self-attention mechanism based on BiLSTM+CRF model.

From the experimental results of Table 5, we have following observations:

- **Effectiveness of transfer learning.** BiLSTM+CRF+transfer improves F1 score from 89.13% to 89.89% as compared with BiLSTM+CRF on SighanNER dataset and achieves 1.08% improvement on WeiboNER dataset, which indicates the word boundary information from CWS is very effective for Chinese NER task.
- **Effectiveness of adversarial training.** By introducing adversarial training, BiLSTM+CRF+adversarial boosts the performance as compared with BiLSTM+CRF+transfer model, showing 0.15% and 0.36% improvement on SighanNER dataset and WeiboNER dataset, respectively. It proves that adversarial training can prevent specific features of CWS task from creeping into shared space.
- **Effectiveness of self-attention mechanism.** When compared with BiLSTM+CRF, the

BiLSTM+CRF+self-attention significantly improves the performance on the two different datasets with the help of information learned from self-attention, which verifies that the self-attention mechanism is effective for Chinese NER task.

We observe that our proposed adversarial transfer learning framework and self-attention lead to noticeable improvements over the baseline, improving F1 score from 51.01% to 53.08% on WeiboNER dataset and giving 1.51% improvement on SighanNER dataset.

4.5 Detailed Analysis

4.5.1 Case Study

Word boundary information from CWS task is very important for Chinese NER task, especially when different entities appear together, . We take a sentence in WeiboNER test set as example for illustrating the effectiveness of our proposed model. As shown in Figure 4(a), when two "person" entities appearing together, our proposed method exploits word segmentation information to determine the boundary between them and then make correct taggings. In Figure 4(b), when labelling the word "上司 (the boss)", the self-attention explicitly learns the dependencies with "尊重 (respect)", therefore, our model enables to correctly classify the word into "person" category. It verifies that the self-attention is very effective for Chinese NER task.

4.5.2 Error Analysis

According to the results of Table 2 and Table 4, our proposed model achieves 4.67% and 1.43% improvement as compared with previous state-of-the-art methods on WeiboNER dataset and SighanNER dataset, respectively. However, the overall performance on WeiboNER dataset is relatively low. Two reasons can be explained for this issue. One reason is that the number of training examples in WeiboNER dataset is very limited as compared with SighanNER dataset. There are only 1.3k examples in WeiboNER training corpora, which is not enough to train deep neural networks. Another reason is that the expression is informal in social media, lowering the performance on WeiboNER dataset. While the greater improvement on WeiboNER dataset proves that our method is helpful to solve the problem.

5 Conclusions

In this paper, we propose a novel adversarial transfer learning framework for Chinese NER task, which can exploit task-shared word boundaries features and prevent the specific information of CWS task. Besides, we introduce self-attention mechanism to capture the dependencies of arbitrary two characters and learn the inner structure information of sentence. Experiments on two different widely used datasets demonstrate that our method significantly and consistently outperforms previous state-of-the-art models.

Acknowledgments

The research work is supported by the Natural Science Foundation of China (No.61533018 and No.61702512), and the independent research project of National Laboratory of Pattern Recognition. This work is also supported in part by Beijing Unisound Information Technology Co., Ltd.

References

Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351.

Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731. Association for Computational Linguistics.

Aitao Chen, Fuchun Peng, Roy Shan, and Gordon Sun. 2006. Chinese named entity recognition with conditional probabilistic models. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 173–176.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. *arXiv preprint arXiv:1704.07556*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. 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, pages 167–176.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. 2017. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420.

Hangfeng He and Xu Sun. 2017a. F-score driven max margin neural network for named entity recognition in chinese social media. *EACL 2017*, page 713.

Hangfeng He and Xu Sun. 2017b. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *AAAI*, pages 3216–3222.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Wencan Luo and Fan Yang. 2016. An empirical study of automatic chinese word segmentation for spoken language understanding and named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–248.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 149–155.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2017. Disan: Directional self-attention network for rnn/cnn-free language understanding. *arXiv preprint arXiv:1709.04696*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2017. Deep semantic role labeling with self-attention. *arXiv preprint arXiv:1712.01586*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Chunqi Wang, Wei Chen, and Bo Xu. 2017. Named entity recognition with gated convolutional neural networks. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 110–121. Springer.

YaoSheng Yang, Meishan Zhang, Wenliang Chen, Wei Zhang, Haofen Wang, and Min Zhang. 2018. Adversarial learning for chinese ner from crowd annotations. *arXiv preprint arXiv:1801.05147*.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 956–966.

Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2017. Aspect-augmented adversarial networks for domain adaptation. *arXiv preprint arXiv:1701.00188*.

Junsheng Zhou, Liang He, Xinyu Dai, and Jiajun Chen. 2006. Chinese named entity recognition with a multi-phase model. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 213–216.

Junsheng Zhou, Weiguang Qu, and Fen Zhang. 2013. Chinese named entity recognition via joint identification and categorization. *Chinese journal of electronics*, 22(2):225–230.