A Double Adversarial Network Model for Multi-Domain and Multi-Task Chinese Named Entity Recognition

Yun HU(a), Student Member and Changwen ZHENG††, Nonmember

SUMMARY Named Entity Recognition (NER) systems are often realized by supervised methods such as CRF and neural network methods, which require large annotated data. In some domains that small annotated training data is available, multi-domain or multi-task learning methods are often used. In this paper, we explore the methods that use news domain and Chinese Word Segmentation (CWS) task to improve the performance of Chinese named entity recognition in weibo domain. We first propose two baseline models combining multi-domain and multi-task information. The two baseline models share information between different domains and tasks through sharing parameters simply. Then, we propose a Double AdversarialVersarial model (DoubADV model). The model uses two adversarial networks considering the shared and private features in different domains and tasks. Experimental results show that our DoubADV model outperforms other baseline models and achieves state-of-the-art performance compared with previous works in multi-domain and multi-task situation.

key words: Chinese named entity recognition, multi-domain learning, multi-task learning

1. Introduction

Name entity recognition is a fundamental Natural Language Processing (NLP) task that labels each word in sentences with predefined types, such as Person (PER), Location (LOC), Organization (ORG) and so on. The results of NER can be used in many downstream NLP tasks, such as relation extraction [1] and question answering [2]. Although NER is considered as a solved task in some domains like the news domain. However, in some user-generated domains like weibo domain is still an active research area. The main reason is that the NER system is often realized by supervised methods like CRF [3] and neural network methods [4]–[7]. Supervised methods often require large annotated data. However, a small amount of annotated data is available in some domains, such as the Chinese weibo domain. In this paper, we use the NER in Chinese weibo domain as a case to study, and the methods can be used to other domains and languages.

To address the problem of lacking annotated data, multi-domain or multi-task methods are often used [8]–[10]. Multi-domain learning transfers the information from the source domain to the target domain. Multi-task-learning transfers the information from the source task to the target task. The Chinese word segmentation and Chinese news domain are used to help the Chinese weibo named entity recognition [11]–[13]. An example is shown in Fig. 1. For example, the NER system in weibo domain and news domain all need to recognize the person name like ‘Han Han’ (Han Han) and ‘Xianliang Zhang’ (Xianliang Zhang) which is an independent word like ‘张贤亮’ which is a common word in the third sentence. Previous works mainly research multi-domain learning and multi-task learning separately. However, how to combine multi-domain learning and multi-task learning is still not well investigated.

In this paper, we explore two baseline models to combine multi-domain learning and multi-task learning first. In the first baseline model (SHLSTM model, SHared bi-LSTM model), the domain and task information are simply shared through a shared feature extractor. In the second baseline model (SPLSTM model, SPerific bi-LSTM model), the domain information and task information are extracted and shared by separated feature extractors. Then, a DoubADV model is proposed based on two baseline models. We consider that different domains and tasks will have both shared and private information. For example, in Fig. 1, for CWS task and NER task, some named entities have the same boundaries as the word boundaries, such as ‘张贤亮’ (Xianliang Zhang). While some entities have different boundaries, such as ‘李玲’ (Peng Li). In weibo domain and news domain, the writing styles of the weibo sentences are similar to the writing styles of news sentences,
such as the style of the third sentences. While some weibo sentences may have different writing styles, such as the first sentence. Two adversarial networks are used to process the shared and private information in different domains and tasks. Experiments show that our DoubADV model outperforms the baseline models.

- We propose two baseline models which combine the multi-domain learning and multi-task learning in Chinese NER.
- We propose a DoubADV model based on two baseline models, which considers the shared and private information in different domains and tasks.
- Experiments show that our DoubADV model outperforms other multi-domain and multi-task models in weibo NER dataset. The source code will be released later for further research.

2. Models

In this part, we first introduce the LSTM-CRF model which is widely used in Chinese named entity recognition [6]. Then, we consider two types of baseline multi-domain and multi-task models. The first is the SHLSTM model, the second is the SPLSTM model. The models use four types of data as input: weibo NER, weibo CWS, news NER, news CWS. In this paper, the superscript \( d \) represents the domain, which is one of \{weibo, news\}, and \( t \) represents the task, which is one of \{NER, CWS\}. The superscript \( dt \) is one of \{weibo NER, weibo CWS, news NER, news CWS\}. Finally, we describe the DoubADV model which considers both shared and private information.

2.1 LSTM-CRF Model

The architecture of LSTM-CRF is shown in Fig. 2 (a). The input data of the model is weibo NER data. The model contains three sub-modules: embedding layer, feature extractor (Bi-LSTM), and CRF. The ‘Weibo NER’ in ‘Weibo NER Bi-LSTM’ and ‘Weibo NER CRF’ mean that information from weibo NER dataset will flow into the Bi-LSTM and CRF modules.

**Embedding layer** Similar to other neural network models, we use the embedding layer as the first step. The embedding layer maps the input character into a low dimensional dense vector by pre-trained word embedding. The embedding vector \( x_i \) can be obtained by:

\[
x_i = \text{Embedding}(c_i; \theta_e)
\]

where \( c_i \) is the character input, and \( \theta_e \) is the shared look up table.

**Feature extractor** The feature extractor is realized using Bi-LSTM. The LSTM can handle gradient vanishing/exploding problems well [14]. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The Bi-LSTM concatenates the forward LSTM output and backward LSTM output to capture the information of a word from context [15]. The feature \( h_i \) extracted by Bi-LSTM can be computed by:

\[
h_i = \text{Bi-LSTM}(x_i; \theta_l)
\]

where \( \theta_l \) is the Bi-LSTM parameters.

CRF We use CRF to predict the label sequence. The neighborhood information can be considered in CRF. For example, the ‘I-LOC’ (Inter middle of Location) can not behind the ‘B-PER’ (Beginning of Person). The final predicted sequence \( y_i \) can be obtained by:

\[
y_i = \text{CRF}(h_i; \theta_c)
\]

where \( \theta_c \) is the CRF parameters. The loss of the model is CRF loss, which is negative log-probability of the correct tag sequence. More details can be obtained from [5].

2.2 SHLSTM Model

In order to improve the performance of named entity recognition in low resource domain, we use multi-domain and multi-task data to assist. Compared with previous works using multi-domain or multi-task data alone, more information is considered. The simplest model to use multi-domain and multi-task data is the SHLSTM model. The architecture of the SHLSTM model is shown in Fig. 2 (b). The model contains three parts: shared embedding, shared feature extractor, and specific CRF. The information from different domains and different tasks will flow into a domain and task shared Bi-LSTM. The shared model parameters may retain the information from high resource domains and tasks which is helpful to low resource domains and tasks. Compared with LSTM-CRF model, the SHLSTM model is a multi-domain and multi-task model and has the specific CRF.

**Shared embedding** The SHLSTM model uses shared character embedding. The pretrained character embedding is obtained from a large annotated corpus. The features obtained by the character embedding can be reused by different domains and tasks.

**Shared feature extractor** The shared feature extractor can guarantee the information shared between the different domains and tasks. The feature extractor is realized by a Bi-LSTM, which uses the output of shared character embedding from different domains and tasks as input. The low resource domains and tasks such as weibo NER can be benefited from the features learned by high resource domains and tasks.

**Specific CRF** The different information between domains and tasks can be captured by different CRFs. For example, different domains and tasks have different label sets. The label sets of weibo CWS is \{B, M, E, S\}, and the label sets of weibo NER is \{LOC, PER, ORG, GPE (Geographical, Social, Political Entity), O\}. The final predict sequence \( y_i^d \) from domain \( d \) and tasks \( t \) can be computed by:

\[
y_i^d = \text{CRF}(h_i; \theta_c^d)
\]
where $h_i$ is the output of the shared Bi-LSTM, and $\theta_{c,t}$ is the CRFs parameters. For different domains $d$ and tasks $t$, the CRFs parameters are different. The loss of the SHLSTM model is CRF loss.

2.3 SPLSTM Model

The SHLSTM model may not consider the differences between different domains and tasks, and the SPLSTM model can deal with this situation well. The architecture of the SPLSTM model is shown in Fig. 2(c). The model uses four data sets as inputs. The model contains three parts: shared embedding, specific feature extractor, and specific CRF. The embedding is shared between different domains and tasks which is the same as SHLSTM model. The specific CRF is separated for different domains and tasks. Different from the SHLSTM model, the SPLSTM model supposes that a sentence has two types of features: domain feature and task feature. The SPLSTM model uses specific feature extractors to process the domain and task feature.

**Specific feature extractor** The specific feature extractor is realized by four Bi-LSTMs: CWS Bi-LSTM, NER Bi-LSTM, weibo Bi-LSTM and news Bi-LSTM. For a sentence from weibo NER, the sentence contains weibo feature and NER feature. The hypothesis is that the domain information is independent from task information. The domain feature extractor can be benefited from the different tasks in the same domain, and the task feature extractor can be benefited from the different domains in the same task. For example, when the weibo NER dataset is used as input, the information will first flow into the shared embedding, and then flow into the weibo Bi-LSTM and NER Bi-LSTM respectively. When the weibo CWS dataset is used as input, the information will first flow into the shared embedding, and then flow into the weibo Bi-LSTM and CWS Bi-LSTM respectively. The weibo Bi-LSTM can be benefited from the weibo NER data and weibo CWS data. The output of domain $d$ Bi-LSTM $h_i^d$ and the output of task $t$ Bi-LSTM $h_i^t$.
can be computed by:

$$h^d_t = \text{Bi-LSTM}(x^{dt}_t; \theta^d_t)$$  \hspace{1cm} (5) \\
$$h^t_t = \text{Bi-LSTM}(x^{tt}_t; \theta^t_t)$$  \hspace{1cm} (6)

where $x^{dt}_t$ is the embedding vector, $\theta^d_t$ is the parameters of domain $d$ Bi-LSTM, and $\theta^t_t$ is the parameters of task $t$ Bi-LSTM.

**Specific CRF** The two types of feature information for a sentence are aggregated before making the prediction. For example, the CRF of weibo NER uses the output of weibo Bi-LSTM and NER Bi-LSTM as input. The final predict sequence $y^{dt}_i$ from domain $d$ and task $t$ is computed by:

$$y^{dt}_i = \text{CRF}((h^d_t \oplus h^t_t); \theta^{dt}_i)$$  \hspace{1cm} (7)

where $\oplus$ is concatenation operation, and $\theta^{dt}_i$ is the CRF parameters. The loss of the SPLSTM model is CRF loss.

### 2.4 DoubADV Model

For the basic models proposed in Sects. 2.2 and 2.3, it is difficult to consider the shared and private features of different datasets at the same time. In this paper, we propose the DoubADV model, which is an extension of the work of [13]. Compared with [13], DoubADV model can be well applied to multi-domain and multi-task scenario. In terms of model architecture, the DoubADV model uses another adversarial network to deal with domain features. In addition, we do not use self-attention mechanism in the model, because the model without self-attention mechanism saves more memory and training time, and the performance is similar to the model using attention mechanism.

The detailed architecture of the DoubADV model is shown in Fig. 2 (d). The model contains 5 parts: shared embedding layer, specific feature extractor, DoubADV feature extractor, specific CRF layer, and DoubADV discriminator. The model uses shared embedding, specific feature extractor and specific CRF which are the same as SPLSTM. Compared with SPLSTM, the DoubADV model uses DoubADV feature extractor and DoubADV discriminator to consider the information shared between domains and tasks. For example, the word boundaries information is considered to help NER.

**DoubADV feature extractor** The DoubADV feature extractor contains 2 types of extractor: domain-shared Bi-LSTM and task-shared Bi-LSTM. The domain-shared Bi-LSTM and task-shared Bi-LSTM are shared between all domains and tasks. The difference between domain-shared Bi-LSTM and task-shared Bi-LSTM is that the output of domain-shared Bi-LSTM is used as the input of domain discriminator, and the output of task-shared Bi-LSTM is used as the input of task discriminator. For example, in Fig. 2 (d), when a sentence from weibo NER dataset is used as input, the information will first pass through the shared embedded layer, and then pass through the weibo Bi-LSTM, NER Bi-LSTM, domain-shared Bi-LSTM, and task-shared Bi-LSTM. The arrows indicate the direction of information flow. In the aspect of implementation, the output of the shared character embedding is used as the input of four Bi-LSTMs: weibo Bi-LSTM, NER Bi-LSTM, domain-shared Bi-LSTM, and task-shared Bi-LSTM. Compared with SPLSTM, the weibo NER not only benefit from private information through weibo Bi-LSTM and NER Bi-LSTM, but also benefit from news information through domain-shared Bi-LSTM and CWS information through task-shared Bi-LSTM. The output of domain-shared Bi-LSTM $h^{ds}_i$ and task-shared Bi-LSTM $h^{ts}_i$ can be computed as follow:

$$h^{ds}_i = \text{Bi-LSTM}(x^{ds}_i; \theta^{ds}_i)$$  \hspace{1cm} (8) \\
$$h^{ts}_i = \text{Bi-LSTM}(x^{ts}_i; \theta^{ts}_i)$$  \hspace{1cm} (9)

where $x^{dt}_i$ is the embedding vector, $\theta^{ds}_i$ is the parameters of domain-shared Bi-LSTM, and $\theta^{ts}_i$ is the parameters of task-shared Bi-LSTM.

**Specific CRF** The four features are merged before fed into CRF layer to make prediction:

$$h_i = h^d_t \oplus h^t_t \oplus h^{ds}_i \oplus h^{ts}_i$$  \hspace{1cm} (10) \\
$$y^{dt}_i = \text{CRF}(h_i; \theta^{dt}_i)$$  \hspace{1cm} (11)

where $\oplus$ is concatenation operation, $h_i$ is the feature merged the private and shared information, $\theta^{dt}_i$ is the CRF parameters, and $y^{dt}_i$ is the final predict sequence.

**DoubADV discriminator** The DoubADV discriminator contains two types of discriminators: domain discriminator and task discriminator. The domain discriminator is used to guarantee the domain-shared Bi-LSTM does not contain private domain features. The task discriminator is used to guarantee the task-shared Bi-LSTM does not contain private task features. The discriminators are standard feed-forward networks with a softmax layer for classification. The domain discriminator is used to classify that the sentence is from weibo domain or news domain, and the task discriminator is used to classify that the sentence is from NER task or CWS task. Besides, a maxpooling is used to gather the sentence information. A gradient reversal layer is added below the softmax layer to solve the minimax optimization problem. The DoubADV feature extractor and the DoubADV discriminator will reach a point where the discriminator cannot differentiate the domains or the tasks according to the representation learned from the DoubADV feature extractor. The formalization of Domain or task discriminator is shown as follow:

$$h = \text{Maxpooling}(H)$$  \hspace{1cm} (12) \\
$$D(h; \theta_d) = \text{softmax}(W_d h + b_d)$$  \hspace{1cm} (13)

where $H$ is the feature representation of the sentences, and $\theta_d$ is the parameters in softmax, including $W_d$ and $b_d$.

We can represent the final loss $L$ as follow:

$$L = L^{dt} + \alpha L_{task} + \beta L_{domain}$$  \hspace{1cm} (14)

where $L^{dt}$ is the loss from the CRF part when the input data.
is from domain \(d \) and task \(t \), \(L_{\text{task}} \) is the loss from task discriminator, and \(L_{\text{domain}} \) is the loss from domain discriminator. The \( \alpha \) and \( \beta \) are hyper-parameters.

3. Experiments and Results

3.1 Datasets

In the LSTM-CRF model, we only use Chinese weibo NER data as input data [16, 17]. In the SHLSTM model, SPLSTM model, and DoubADV model, we use four types of data as inputs. The Chinese weibo NER corpus is from [16, 17]. The Chinese news NER corpus is from SIGHAN 2006 [18]. The Chinese weibo word segmentation corpus is from NLPCC 2016 [19]. The Chinese news word segmentation corpus is from SIGHAN 2005 [20]. The number of sentences in the different corpora is shown in Table 1.

In this paper, we focus on weibo NER. We use the two weibo NER datasets: the original weibo NER dataset [16] and the updated weibo NER dataset [17], which is the same with [13]. Four types of named entity are considered in the corpora: PER, LOC, ORG, and GPE. The number of named entities in the Chinese weibo NER is shown in Table 2.

3.2 Parameters Setting

We use the character embedding which is pre-trained on weibo corpus and Baidu Encyclopedia using word2vec [23]. The embedding dimension is 100. In all models, the LSTM dimension is 100. The other trainable parameters are initialized by xavier initializer [24]. The initial learning rate is 0.001. The dropout rate is 0.3. The \( \alpha \) and \( \beta \) are set to 0.06. The optimization method is adam [25].

3.3 Training

In one iteration, we select a batch of data from one of the four datasets in turn to train the model and optimize the loss. The training process is stopped until the Chinese weibo NER task convergence. Compared with SHLSTM and SPLSTM model, our DoubADV model considers the domain discriminator loss and task discriminator loss.

3.4 Results

For evaluation, the P(precision), R(Recall), and F1 score are used as metrics. In Table 3, we show our model results in weibo NER test datasets. The DoubADV model achieves the best F1 score in both original and updated weibo NER datasets. Compared with the LSTM-CRF model, the DoubADV model gains 13.19% and 6.67% F1 score improvements in original and updated datasets respectively. The main reason for the improvements comes from introducing other domains and tasks information. The DoubADV model outperforms the SHLSTM and SPLSTM model, which shows that the DoubADV model can better capture the information from other domains and tasks. In Table 4, we show the F1 results of the model on each named entity category. The results show that the DoubADV model achieves the best results in all categories compared with other models. The results also show that the recognition performance of all models on ORG is very poor. We will explore how to improve the performance of models on ORG in the future.

We also compared our method with previous works. We first compare our methods with five supervised neural network methods, which only use weibo NER data as the input. The results are shown in the first five rows of Table 3. Peng and Dredze (a) used character and position embedding to incorporate the word information into a character-based NER model [16]. He and Sun (a) proposed an F-score driven max margin neural network to instead CRF as final output [17]. Zhang and Yang investigated a lattice-structured LSTM model to utilize word information [7]. Meng et al. proposed glyph-vectors for Chinese character representations and used the glyph vector for NER task [21]. Zhu and Wang used the convolutional attention network to solve the Chinese NER [22]. The model achieves the previous state-of-the-art performance in the updated dataset. The DoubADV model gains the 1.01% F1 score improvement compared with [22]. The DoubADV model outperforms the supervised neural network methods in F1 score, which shows that multi-domain learning and multi-task learning is helpful. The latest works [21] have introduced the BERT [26] to the Chinese NER, we will do the work in the future.

Then we compare our methods with works using multi-domain or multi-task learning methods. The results are shown in the middle three rows of Table 3. Peng and Dredze (b) jointly trained word segmentation and NER with an LSTM-CRF model [11]. He and Sun (b) proposed a unified model that learned from out-of-domain corpora and in-domain unannotated texts [12]. Cao et al. proposed an adversarial transfer learning framework to better utilize the information from CWS [13]. The model achieves the previous state-of-the-art performance in the original dataset. The DoubADV model gains the 1.71% F1 score improvement compared with [13]. Compared with works using multi-domain or multi-task learning methods, the DoubADV model uses two adversarial networks to combine multi-domain learning and multi-task learning.
Table 3 Test results for Chinese named entity recognition in weibo NER.

|          | Original |          |          | Updated |          |          |          |
|----------|----------|----------|----------|---------|----------|----------|----------|
|          | P        | R        | F        |         | P        | R        | F        |
| Peng and Dredze (a) [16] | 51.98    | 35.57    | 44.09    | 74.78   | 39.81    | 51.96    |
| He and Sun (a) [17]      | -        | -        | -        | 66.93   | 40.67    | 50.60    |
| Zhang and Yang [7]       | -        | -        | -        | -       | -        | 53.04    |
| Meng et al. [21]         | -        | -        | 53.69    | 55.30   | 54.32    |
| Zhu and Wang [22]        | -        | -        | -        | -       | -        | 55.38    |
| Peng and Dredze (b) [11] | 63.33    | 39.18    | 48.41    | 66.67   | 47.22    | 55.28    |
| He and Sun (b) [12]      | -        | -        | 61.68    | 48.82   | 54.50    |
| Cao et al. [13]          | -        | -        | 55.72    | 50.68   | 53.08    |
| LSTM-CRF                 | 48.99    | 52.38    | 20.00    | 14.04   | 56.57    | 68.13    |
| SHLSTM                   | 57.30    | 60.87    | 37.50    | 22.64   | 60.87    | 66.66    |
| SPLSTM                   | 56.85    | 60.00    | 42.42    | 25.93   | 60.77    | 68.18    |
| DoubADV                  | 59.46    | 69.44    | 43.75    | 26.41   | 61.11    | 68.89    |

Table 4 F1 results on different categories.

|          | Original |          |          | Updated |          |          |          |
|----------|----------|----------|----------|---------|----------|----------|----------|
|          | PER      | GPE      | LOC      | ORG     | PER      | GPE      | LOC      | ORG     |
| LSTM-CRF | 48.89    | 52.38    | 20.00    | 14.04   | 56.57    | 68.13    | 16.00    | 15.62   |
| SHLSTM   | 57.30    | 60.87    | 37.50    | 22.64   | 60.87    | 66.66    | 33.33    | 20.00   |
| SPLSTM   | 56.85    | 60.00    | 42.42    | 25.93   | 60.77    | 68.18    | 36.36    | 21.28   |
| DoubADV  | 59.46    | 69.44    | 43.75    | 26.41   | 61.11    | 68.89    | 41.67    | 24.00   |

Table 5 Comparisons with models using multi-domain learning or multi-task learning.

|          | P        | R        | F        |
|----------|----------|----------|----------|
| LSTM-CRF | 56.07    | 44.50    | 49.62    |
| ADV-domain | 57.06   | 52.06    | 54.45    |
| ADV-task  | 60.92    | 47.42    | 53.33    |
| DoubADV   | 64.66    | 50.00    | 56.39    |

4. Analysis

4.1 Different Component Learning

We compare our DoubADV model with the methods using one adversarial network. The results are shown in Table 5. For the ADV-domain model, only the multi-domain information is used. The weibo NER data and news NER data are used as inputs. For the ADV-task model, only the multi-task information is used. The weibo NER data and weibo CWS data are used as inputs. The architecture of the ADV-domain model and ADV task model is the same as [13]. Two private Bi-LSTMs and one shared Bi-LSTM are used. The DoubADV model obtains the best results, which shows that both multi-domain information and multi-task information are helpful for weibo NER.

4.2 Data Size Learning

We explore the performance of models when different sizes of weibo NER data are used. The results are shown in Fig. 3. In one part, The advantage of multi-domain and multi-task learning is obvious when small training data is available. The smaller training data is available, the larger improvement can the multi-domain and multi-task models obtained. In the other part, the DoubADV model achieves the best results in most cases (50%, 75%, and full data size).

4.3 Other Datasets Learning

To show our methods is useful, we test our models in different NER datasets: variant of updated weibo NER datasets, Ontonotes and high resource dataset.

Variant of updated weibo NER datasets Two types of variant are considered: merged NER [27], [28] and NOM [11], [16]. The results are shown in Table 7. In the merged NER, the ‘GPE’ and the ‘LOC’ label are merged in Weibo NER data, so the source domain NER and target domain NER have the same labeling sets. In the NOM, the nominal mention is considered [16]. In all situations, our DoubADV model outperforms LSTM-CRF, SHLSTM and SPLSTM model.

Ontonotes learning Ontonotes [29] is a widely used dataset that contains different domains of Chinese named entities.

Fig. 3 The results of models using different sizes of weibo NER training data. 25%, 50%, 75% and full size of weibo NER training data are explored. The sizes of the other three corpora are unchanged. The F1 score of LSTM-CRF is 20.23% when the data size is 25%.
Table 7  The F1 score in other datasets learning.

| Methods  | Merged NER | NOM  | BC   | TC   | MSRA  |
|----------|------------|------|------|------|-------|
| LSTM-CRF | 51.66      | 54.39| 47.12| 61.76| 89.36 |
| SHLSTM   | 56.79      | 54.54| 54.44| 66.17| 85.83 |
| SPLSTM   | 57.70      | 55.29| 54.41| 69.58| 87.34 |
| DoubADV  | 57.98      | 57.53| 54.59| 69.92| 83.38 |

Table 6  The number of sentences used in Ontonotes learning.

|                  | #Train | #Dev  | #Test |
|------------------|--------|-------|-------|
| BC/TC NER        | 6986/7282 | 2000/1000 | 2000/1000 |
| News NER         | 16814   | 1868  | 4636  |
| BC/TC CWS       | 31082/32785 | 3000/3000 | 3000/3000 |
| News CWS        | 17000   | 1900  | 1900  |

entity recognition annotations [7], [21]. We test our models on the Broadcast Conversation domain (BC) and Telephone Conversation domain (TC) in Ontonotes datasets. The details of the corpus used in Ontonotes learning are shown in Table 6. The results are shown in Table 7. The multi-domain and multi-task models outperform the baseline models, which demonstrates that combining multi-domain and multi-task is helpful in Ontonotes. Similar to previous works, the DoubADV model achieves the best results among multi-domain and multi-task models.

High resource dataset In this paper, we use the news dataset as the high resource dataset. The three models (SHLSTM, SPLSTM and DoubADV) all use MSRA NER datasets as inputs. The weibo dataset is used as the low resource dataset. The details of data are the same as Sect. 3.1. In Table 7, we show that the performance of models in MSRA. The results show that the models can even have a slight performance decrease in high resource situation. We test ADV-task model and ADV-domain models as in Sect. 4.1, the results show that the performance slightly improves in ADV-task model and decreases in ADV-domain model. The results also indicate that the performance improvement of multi-domain and multi-task methods depends on the increase of dataset to a great extent. For high resource corpus, it has a large amount of data, so it is difficult to improve performance by a small amount of multi-task or multi-domain data. For multi-task method, there may be a lot of data for related tasks (CWS task), which can improve performance. For multi-domain method, because the data in related domain (weibo domain) is much less than that in high resource domain (news domain), the performance may not be improved.

4.4 Case Study

In Fig. 4, we present two examples to show that multi-domain and multi-task learning are helpful for the Chinese weibo NER. In the first case, the ADV-domain model labels the ‘雅典’ (Athens) as GPE, while ADV-task model does not. We find that ‘雅典’ (Athens) appears as an entity in news NER dataset (SIGHAN 2006), which may help the recognition. In the second case, the ADV-task model can find the ‘邓超’ (Deng Chao) as PER. We find that ‘贴吧’ (Tieba) appears 3 times as a word in weibo CWS dataset (NLPCC 2016), while we do not find the continue character ‘超贴’ (Chao Tie). The DoubADV model combines the advantage of multi-domain model and multi-task model, which makes more generalized results.

5. Related Work

Recently, most NER systems are based on supervised methods, such as Conditional Random Fields (CRF) and Neural Networks model. CRF model relies heavily on hand-crafted features [3], [30], [31]. The features of the sequence are extracted automatically in the neural network model [4], [5], [17], [22]. For Chinese named entity recognition, some specific language information are considered, such as radical information [6], word information [7] and glyph information [21]. In this paper, our model is based on the basic neural network models in [5].

For multi-task learning, the chunk and POS tagging were used to jointly train with the NER task in English [4], [8], [9]. In Chinese, the Chinese word segmentation was used as the auxiliary task of Chinese NER [11], [13], [16], [27]. For multi-domain learning, the twitter was studied in English [8], [10] and the weibo was studied in Chinese [17], [27], [28]. In this paper, we extend the work to combine multi-domain and multi-task learning in Chinese NER.

Adversarial networks have achieved great success in computer vision [32]–[34]. In the natural language processing area, adversarial networks have been explored in crowdsourcing learning [35], cross-lingual learning [36], [37], cross-domain learning [38], [39], and cross-task learning [13], [40], [41]. Compared with these works, we intro-
duced a double adversarial to process the multi-domain and multi-task situations.

6. Conclusion

In this paper, we propose two baseline models to combine multi-domain and multi-task learning. Then, a DoubADV model is proposed based on two baseline models. The DoubADV model can better utilize the shared and private information between different domains and tasks. Experiments show that the DoubADV model achieves state-of-the-art performance in weibo NER dataset. In the future, we will explore the performance of multi-domain and multi-task models in other languages and introduce the BERT as the feature extractor part.

References

[1] R.C. Bunescu and R. Mooney, “A shortest path dependency kernel for relation extraction,” Proc. Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT ’05, pp.724–731, 2005.
[2] X. Yao and B. Van Durme, “Information extraction over structured data: Question answering with freebase,” Proc. 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.956–966, 2014.
[3] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Eighteenth International Conference on Machine Learning, pp.282–289, 2001.
[4] R. Collobert, J. Weston, L. Bottou, M. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” Journal of Machine Learning Research, vol.12, no.Aug, pp.2493–2537, 2011.
[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.260–270, 2016.
[6] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, “Character-based LSTM-CRF with radical-level features for chinese named entity recognition,” Natural Language Understanding and Intelligent Applications, Lecture Notes in Computer Science, vol.10102, pp.239–250, Springer, 2016.
[7] Y. Zhang and J. Yang, “Chinese NER using lattice LSTM,” Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1554–1564, 2018.
[8] Z. Yang, R. Salakhutdinov, and W.W. Cohen, “Transfer learning for sequence tagging with hierarchical recurrent networks,” 2016.
[9] S. Changpinyo, H. Hu, and F. Sha, “Multi-task learning for sequence tagging: An empirical study,” Proc. 27th International Conference on Computational Linguistics, pp.2965–2977, 2018.
[10] B.Y. Lin and W. Lu, “Neural adaptation layers for cross-domain named entity recognition,” Proc. 2018 Conference on Empirical Methods in Natural Language Processing, pp.2012–2022, 2018.
[11] N. Peng and M. Dredze, “Improving named entity recognition for chinese social media with word segmentation representation learning,” Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.149–155, 2016.
[12] H. He and X. Sun, “A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media,” AAAI, pp.3216–3222, 2017.
[13] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, “Adversarial transfer learning for chinese named entity recognition with self-attention mechanism,” Proc. 2018 Conference on Empirical Methods in Natural Language Processing, pp.182–192, 2018.
[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol.9, no.8, pp.1735–1780, 1997.
[15] A. Graves and J. Schmidhuber, “Framework phoneme classification with bidirectional LSTM networks,” Proceedings, 2005 IEEE International Joint Conference on Neural Networks, 2005, IJCNN ’05, vol.4, pp.2047–2052, 2005.
[16] N. Peng and M. Dredze, “Named entity recognition for chinese social media with jointly trained embeddings,” Proc. 2015 Conference on Empirical Methods in Natural Language Processing, pp.548–554, 2015.
[17] H. He and X. Sun, “F-score driven max margin neural network for named entity recognition in chinese social media,” Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp.713–718, 2017.
[18] G.A. Lewov, “The third international chinese language processing bakeoff: Word segmentation and named entity recognition,” Proc. Fifth SIGHAN Workshop on Chinese Language Processing, pp.108–117, 2006.
[19] X. Qiu, P. Qian, and Z. Shi, “Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word segmentation for micro-blog texts,” Natural Language Understanding and Intelligent Applications, Lecture Notes in Computer Science, vol.10102, pp.901–906, Springer International Publishing, Cham, 2016.
[20] T. Emerson, “The second international chinese word segmentation bakeoff,” Proc. Fourth SIGHAN Workshop on Chinese Language Processing, 2005.
[21] W. Wu, Y. Meng, Q. Han, M. Li, X. Li, J. Mei, P. Nie, X. Sun, and J. Li, “Glyce: Glyph-vectors for chinese character representations,” arXiv preprint arXiv:1901.10125, 2019.
[22] Y. Zhu and G. Wang, “CAN-NER: Convolutional attention network for Chinese named entity recognition,” Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp.3384–3393, June 2019.
[23] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems, pp.3111–3119, 2013.
[24] X. Glerot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” Proc. Thirteenth International Conference on Artificial Intelligence and Statistics, pp.249–256, 2010.
[25] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Computer Science, 2014.
[26] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[27] N. Peng and M. Dredze, “Multi-task domain adaptation for sequence tagging,” Proc. 2nd Workshop on Representation Learning for NLP, pp.91–100, 2017.
[28] Z. Wang, Y. Qu, L. Chen, J. Shen, W. Zhang, S. Zhang, Y. Gao, K. Chen, and Y. Yu, “Label-aware double transfer learning for cross-specialty medical named entity recognition,” Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp.1–15, 2018.
[29] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, et al., “Ontonotes release 5.0 ldc2013s19,” Linguistic Data Consortium, Philadelphia, PA, 2013.
[30] W. Chen, Y. Zhang, and H. Isahara, “Chinese named entity recognition with conditional random fields,” Proc. Fifth SIGHAN Workshop on Chinese Language Processing, pp.118–121, 2006.
[31] X. Mao, Y. Dong, S. He, S. Bao, and H. Wang, “Chinese word segmentation and named entity recognition based on conditional ran-
dom fields,” Proc. Sixth SIGHAN Workshop on Chinese Language Processing, 2008.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems, pp.2672–2680, 2014.

[33] E.L. Denton, S. Chintala, R. Fergus, et al., “Deep generative image models using alaplacian pyramid of adversarial networks,” Advances in Neural Information Processing Systems, pp.1486–1494, 2015.

[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” Domain Adaptation in Computer Vision Applications, Advances in Computer Vision and Pattern Recognition, pp.189–209, Springer International Publishing, Cham, 2017.

[35] Y. Yang, M. Zhang, W. Chen, W. Zhang, H. Wang, and M. Zhang, “Adversarial learning for Chinese ner from crowd annotations,” arXiv preprint arXiv:1801.05147, 2018.

[36] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, “Adversarial deep averaging networks for cross-lingual sentiment classification,” Transactions of the Association for Computational Linguistics, vol.6, pp.557–570, 2018.

[37] J.-K. Kim, Y.-B. Kim, R. Sarikaya, and E. Fosler-Lussier, “Cross-lingual transfer learning for pos tagging without cross-lingual resources,” Proc. 2017 Conference on Empirical Methods in Natural Language Processing, pp.2832–2838, 2017.

[38] Y. Zhang, R. Barzilay, and T. Jaakkola, “Aspect-augmented adversarial networks for domain adaptation,” Transactions of the Association for Computational Linguistics, vol.5, pp.515–528, 2017.

[39] T. Gui, Q. Zhang, H. Huang, M. Peng, and X. Huang, “Part-of-speech tagging for twitter with adversarial neural networks,” Proc. 2017 Conference on Empirical Methods in Natural Language Processing, pp.2411–2420, 2017.

[40] X. Chen, Z. Shi, X. Qiu, and X. Huang, “Adversarial multi-criteria learning for Chinese word segmentation,” Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp.1193–1203, July 2017.

[41] P. Liu, X. Qiu, and X. Huang, “Adversarial multi-task learning for text classification,” Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp.1–10, July 2017.

Yun Hu received the B.E. degree in University of Science and Technology Beijing. He is currently a Ph.D. student in Institute of Software, Chinese Academy of Science. His research interests include natural language processing and artificial intelligence.

Changwen Zheng received the Ph.D. degree in Huazhong University of Science and Technology. He is currently a professor in Institute of Software, Chinese Academy of Science. His research interests include computer graph and artificial intelligence.