CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition

Yuying Zhu*
Nankai University
Tianjin, China
yuyzhu@mail.nankai.edu.cn

Guoxin Wang
Microsoft Research Asia
Beijing, China
guow@microsoft.com

Abstract

Named entity recognition (NER) is a common task in Natural Language Processing (NLP), but it remains more challenging in Chinese because of its lack of natural delimiters. Therefore, Chinese Word Segmentation (CWS) is usually necessary as the first step for Chinese NER. However, models based on word-level embeddings and lexicon features often suffer from segmentation errors and out-of-vocabulary (OOV) problems. In this paper, we investigate a **C**onvolutional **A**ttention **N**etwork (**CAN**) for Chinese NER, which consists of a character-based convolutional neural network (CNN) with local-attention layer and a gated recurrent unit (GRU) with global self-attention layer to capture the information from adjacent characters and sentence contexts. Moreover, differently from other approaches, CAN-NER does not depend on any external resources like lexicons and employing small-size char embeddings makes CAN-NER more practical for real systems scenarios. Extensive experimental results show that our approach outperforms state-of-the-art methods without word embedding and external lexicon resources on different domains datasets.

1 Introduction

Named Entity Recognition (NER) aims at identifying text spans which are associated with a specific semantic entity type such as person (PER), organization (ORG), location (LOC), and geopolitical entity (GPE). NER has received constant research attention as it is the first step in a wide range of downstream Natural Language Processing (NLP) tasks, e.g., entity linking (Gupta et al., 2017), relation extraction (Miwa and Bansal, 2016), event extraction (Chen et al., 2015), and coreference resolution (Fragkou, 2017). The standard approach in existing state-of-the-art models for English NER treats the problem as a word-by-word sequence labeling task and makes full use of the Recurrent Neural Network (RNN) and Conditional Random Field (CRF) to capture context information at the word level (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016; Liu et al., 2018). These models for English NER predict a tag for each word assuming that words can be separated clearly by explicit word separators, e.g., blank spaces. As the Chinese language has no natural delimiters, it would be intuitive to apply Chinese Word Segmentation (CWS) first to get word boundaries and then use a word-level sequence labeling model similar to the English NER models. However, word boundaries can be ambiguous in Chinese, which leads to the possibility that entity boundaries do not match word boundaries. For example, the term "西藏自治区 (Tibet Autonomous Region)" is a GPE-type entity in NER, but it could be segmented as a single word or as two words "西藏 (Tibet)" and "自治区 (autonomous region)" separately, depending on different granularity of segmentation tools. Most of the time, however, it is hard to determine the correct granularity for word segmentation. Also, as shown in Figure 1, different segmentation can lead to different sentence meanings in Chinese, which could even result in different named entities. Obviously, if entity boundaries are

Sentence:
南京市长江大桥

Segmentation 1:
南京市 | 长江大桥
Nanjing City, Yangtze River Bridge
Location, Location

Segmentation 2:
南京 | 市长 | 江大桥
Nanjing, Mayor, Jiang Daqiao
Location, Title, Person

Figure 1: Entity Ambiguity with Word Segmentation.

* This work was performed when the first author was an intern at Microsoft Research Asia.

Proceedings of NAACL-HLT 2019, pages 3384–3393
Minneapolis, Minnesota, June 2 - June 7, 2019.

mistakenly detected in segmentation, it will negatively affect entity tagging in word-based NER models. Furthermore, most recent neural network-based Chinese NER models rely heavily on word-level embeddings and external lexicon sets (Huang et al., 2017; Zhang and Yang, 2018). The quality of such models strongly relies on the different word embedding representations and lexicon features. Moreover, word-based models tend to suffer from OOV issues as Chinese words can be very diverse and named entities are an important source of OOV words. Other potential limitations are as follows: (1) Dependency on word embeddings increases model size and makes the fine-tuning process more costly during training (while negatively affecting latency in testing/decoding); (2) It is hard to learn word representation correctly without enough labeled utterances for named entities are usually rarer proper nouns. (3) Large lexicons are very costly for real NER systems as they greatly increase memory usage and latency in feature extraction (matching), which makes models inefficient; (4) It is very costly to remove noise from large lexicons and any update to pre-trained word embeddings or lexicons requires model re-training. Meanwhile, character-level embedding by itself can only carry limited information due to losing word and word-sequence information. For instance, the character "拍" in words "球拍" (bat) and "拍卖" (auction) has very different meanings. How to better integrate segmentation-related information and exploit local context information is the key feature in a character-based model. Zhang and Yang (2018) leverage lexicons to add all the embeddings of candidate word segmentation to their last character embeddings as soft features, and construct a convolutional neural network (CNN) to encode characters as word-level information. Cao et al. (2018) propose a multi-task architecture to learn NER tagging and Chinese word segmentation together, with each part using a character-based Bi-LSTM. In this paper, we propose a convolutional attention layer to capture the implicit relations within adjacent characters, in which the position features from word segmentation are soft hints for character combinations. With the segmentation vector softly concatenating into character embedding, the convolutional attention layer is able to group implicitly meaning-related characters and help bypass the impact of segmentation errors. A BiGRU structure with a global self-attention layer on the whole sentence is utilized to capture sentence-level dependencies. Extensive experimental results show that our approach outperforms state-of-the-art methods without relying on external resources (e.g. word embedding, external lexicon) across different corpora. The main contributions of this paper can be summarized as follows:

- We first combine CNNs with the local-attention mechanism to enhance the ability of the model to capture implicitly local context relations among character sequences. Compared with experimental results against a baseline with a regular CNN layer, our Convolutional Attention layer leads to substantial performance improvements.

- We introduce a character-based Chinese NER model that consists of combined CNN with local attention and BiGRU with global self-attention layers. Our model achieves state-of-the-art F1-scores without using any external resources like word embeddings and lexicon resources, which make it very practical for real-world NER systems.

2 Methodology

We utilize BiGRU-CRF as our basic model structure. Our model considers multi-level context features in three layers: i) convolutional attention layer, ii) GRU layer, and iii) global attention layer. The whole architecture of our proposed model is illustrated in Figure 2.

2.1 Formulation

In the Chinese NER task, we denote an input sentence as $\mathbf{X}_i = \{x_{i,1}, x_{i,2}, x_{i,3}, ..., x_{i,\tau}\}$, where $x_{i,\tau} \in \mathbb{R}^{d_e}$ represents the τ-th character in sentence $\mathbf{X}_i$ and d_e is the dimension of the input embeddings. Correspondingly, we denote the sentence label sequence as $\mathbf{Y}_i = \{y_{i,1}, y_{i,2}, y_{i,3}, ..., y_{i,\tau}\}$, where $y_{i,\tau} \in \mathcal{Y}$ belongs to the set of all possible labels. The objective is learning a function $f_\theta : \mathbf{X} \mapsto \mathbf{Y}$ to obtain the entity types including the 'O' type for all the characters in the input text. In the following text, we take one instance as the example and therefore omit subindex i in the formula.

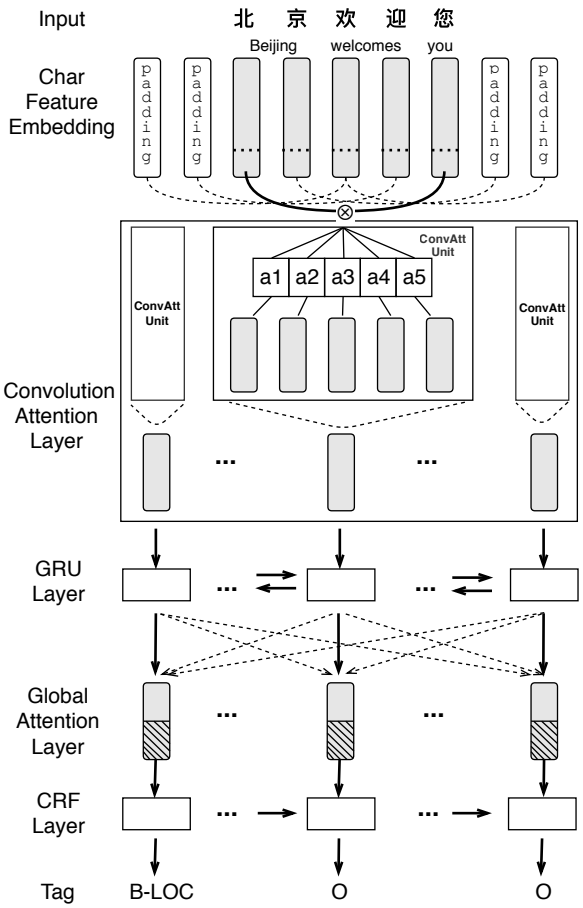


Figure 2: Overall model architecture. A convolutional attention layer is constructed to encode both character- and word-level information. The BiGRU-CRF layer is extended by a global self-attention layer to capture long sequential sentence-level relations.

2.2 Convolutional Attention Layer

The convolutional attention layer aims to encode the sequence of input characters and implicitly group meaning-related characters in the local context. The input representation for each character is constructed as $x = [x_{ch}; x_{seg}]$, where $x_{ch} \in \mathbb{R}^{d_{ch}}$ and $x_{seg} \in \mathbb{R}^{d_{seg}}$ are character embedding and segmentation mask, respectively. The segmentation information is encoded by BMES scheme (Wang and Xu, 2017).

For every window in the CNN, whose window size is k, we first concatenate a position embedding to each character embedding, helping to keep sequential relations in the local window context. The dimension of the position embedding equals to the window size k with the initial values of 1 at the position where the character lies in the window and 0 at other positions. So, the dimension of the concatenated embedding is $d_e = d_{ch} + d_{pos} + d_{seg}$. We then apply local attention inside the window to capture the relations between the center character and each context token, followed by a CNN with sum-pooling layer. We set the hidden dimension as d_h. For the j-th character, the local attention takes all the concatenated embeddings $x_{j-\frac{k-1}{2}}, ...x_j, ..., x_{j+\frac{k-1}{2}}$ in the window as its input and outputs k hidden vectors $h_{j-\frac{k-1}{2}}, ..., h_j, ..., h_{j+\frac{k-1}{2}}$. The hidden vectors are calculated as follows:

$$h_m = \alpha_m x_m, \tag{1}$$

where $m \in \{j - \frac{k-1}{2}, ..., j + \frac{k-1}{2}\}$ and α_m is the attention weight, which is calculated as:

$$\alpha_m = \frac{\exp s(x_j, x_m)}{\sum_{n \in \{j-\frac{k-1}{2}, ..., j+\frac{k-1}{2}\}} \exp s(x_j, x_n)}. \tag{2}$$

The score function s is defined as follows:

$$s(x_j, x_k) = v^\top \tanh(W_1 x_j + W_2 x_k), \tag{3}$$

where $v \in \mathbb{R}^{d_h}$ and $W_1, W_2 \in \mathbb{R}^{d_h, d_e}$.

The CNN layer contains d_h kernels on a context window of k tokens as:

$$h_j^c = \sum_k [W^c * h_{j-\frac{k-1}{2}:j+\frac{k-1}{2}} + b^c], \tag{4}$$

where $W^c \in \mathbb{R}^{k \times d_h \times d_e}$ and $b^c \in \mathbb{R}^{k \times d_h}$. The $*$ operation denotes element-wise product and $h_{j-\frac{k-1}{2}:j+\frac{k-1}{2}}$ means a concatenation of the hidden states $h_{j-\frac{k-1}{2}}, ..., h_{j+\frac{k-1}{2}}$, both of which are calculated at the first dimension. Finally sum-pooling is also conducted on the first dimension.

2.3 BiGRU-CRF with Global Attention

After extracting the local context features by the convolutional attention layer, we feed them into a BiGRU-CRF based model to predict final label for each character. This layer models the sequential sentence information and it is calculated as follows:

$$h_j^r = \text{BiGRU}(h_{j-1}^r, h_j^c; W^r, U^r), \tag{5}$$

where h_j^c is the output of the convolutional attention layer, h_{j-1}^r is the previous hidden state for the BiGRU layer, and $W^r, U^r \in \mathbb{R}^{d_h \times d_h}$ are its parameters.

A global self-attention layer is utilized to better handle sentence-level information, as:

$$h_j^g = \sum_{s=1}^{n} \alpha_{j,s}^g h_s^r \tag{6}$$

where $j = 1, ..., \tau$ denotes all characters in a sentence instance and $\alpha^g_{j,s}$ is calculated as:

$$\alpha^g_{j,s} = \frac{\exp s(h^r_j, h^r_s)}{\sum_{n\in\{1,...,\tau\}} \exp s(h^r_j, h^r_n)}. \quad (7)$$

The score function s is similar to Equation 3 with different parameters $v^g \in \mathbb{R}^{d_h}$ and $W^g_1, W^g_2 \in \mathbb{R}^{d_h, d_h}$ instead.

Finally, a standard CRF layer is used at the top of the concatenation of the output of the BiGRU and global attention layers, which is denoted as $H_\tau = [h^r_\tau; h^g_\tau]$. Given the predicted tag sequence $\mathbf{Y} = \{y_1, y_2, y_3, ..., y_\tau\}$, the probability of the ground-truth label sequence is computed by:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{exp(\sum_i(\mathbf{W}^{y_i}_{\text{CRF}} H_i + b^{(y_{i-1}, y_i)}_{\text{CRF}}))}{\sum_{y'} exp(\sum_i(\mathbf{W}^{y'_i}_{\text{CRF}} H_i + b^{(y'_{i-1}, y'_i)}_{\text{CRF}}))}, \quad (8)$$

where y' denotes an arbitrary label sequence, $\mathbf{W}^{y_i}_{\text{CRF}}$ and $b^{(y_{i-1}, y_i)}_{\text{CRF}}$ are trainable parameters. In decoding, we use the Viterbi algorithm to get the predicted tag sequence.

2.4 Training

For training, we exploit log-likelihood objective as the loss function. Given a set of training examples $\{(\mathbf{X}_i, \mathbf{Y}_i)\}|^K_{i=1}$, the loss function L can be defined as follows:

$$L = \sum_{i=1}^{K} logP(\mathbf{Y}_i|\mathbf{X}_i) \quad (9)$$

In the training phase, at each iteration, we first shuffle all the training instances, and then feed them to the model with batch updates. We use the AdaDelta (Zeiler, 2012) algorithm to optimize the final objective with all the parameters as described in Section 3.1.

Dataset	Type	Train	Test	Dev
OntoNotes	Sentences	15.7k	4.3k	4.3k
	Chars	491.9k	208.1k	200.5k
	Entities	13.4k	7.7k	6.95k
MSRA	Sentences	46.4k	4.4k	-
	Chars	2169.9k	172.6k	-
	Entities	74.8k	6.2k	-
Weibo	Sentences	1.4k	0.27k	0.27k
	Chars	73.8k	14.8k	14.5k
	Entities	1.89k	0.42k	0.39k
Resume	Sentences	3.8k	0.48k	0.46k
	Chars	124.1k	15.1k	13.9k
	Entities	1.34k	0.15k	0.16k

Table 1: Statistics of each dataset

3 Experiments

To demonstrate the effectiveness of our proposed model, we have run multiple experiments on Chinese NER datasets covering different domains. This section describes the details of each dataset, settings, and results in our experiments. Standard precision (P), recall (R) and F1-score (F1) are used as evaluation metrics.

3.1 Experimental Settings

Data We use four datasets in our experiments. For the news domain, we experiment on OntoNotes 4 (Weischedel et al., 2011) and MSRA NER dataset from SIGHAN Bakeoff 2006 (Levow, 2006). For the social media domain, we adopt the same annotated Weibo corpus as Peng and Dredze (2015) which is extracted from Sina Weibo[1]. For more variety in test domains, we also use a Chinese Resume dataset (Zhang and Yang, 2018) collected from Sina Finance[2].

The Weibo dataset is annotated with four entity types: PER (Person), ORG (Organization), LOC (Location), and GPE (Geo-Political Entity); and it includes both named and nominal mentions. This corpus is already divided into training, development, and test sets. The Chinese Resume dataset is annotated with eight types of named entities: CONT (Country), EDU (Educational Institution), LOC, PER, ORG, PRO (Profession), RACE (Ethnicity/Background), and TITLE (Job Title). OntoNotes 4 is annotated with four named entity categories: PER, ORG, LOC, and GPE. We follow the same data split method of Che et al. (2013) over OntoNotes 4. Lastly, the MSRA 2006 dataset contains three annotated named entities: ORG, PER and LOC. A development subset is

Models	NE	NM	Overall
Peng and Dredze (2015)	51.96	61.05	56.05
Peng and Dredze (2016)*	55.28	62.97	58.99
He and Sun (2017a)	50.60	59.32	54.82
He and Sun (2017b)*	54.50	62.17	58.23
Cao et al. (2018)	54.34	57.35	58.70
Zhang and Yang (2018)	53.04	62.25	58.79
Baseline	49.02	58.80	53.80
Baseline + CNN	53.86	58.05	55.91
CAN-NER Model	**55.38**	**62.98**	**59.31**

Table 2: Weibo NER results

[1]http://www.weibo.com/
[2]http://finance.sina.com.cn/stock/index.html

not available for the MSRA dataset. The detailed statistics of each datasets are shown in Table 1.

Gold segmentation is unavailable for Weibo, Chinese Resume, and MSRA test sections. We follow Zhang and Yang (2018) to automatically segment these by using the model described in Yang et al. (2017). We treat NER as a sequential labeling problem and adopt the BIOES tagging style since it has been shown to produce better results than straight BIO (Yang et al., 2018b).

Hyper-parameter settings For hyper-parameter configuration, we adjust them according to the performance on the described development sets for Chinese NER. We set the character embedding size, hidden sizes of CNN and BiGRU to 300 dims. After comparing experimental results with different CNN window sizes, we set the window size as 5. Adadelta is used for optimization, with an initial learning rate of 0.005. The character embeddings used in our experiments are from Li et al. (2018), which is trained by Skip-Gram with Negative Sampling (SGNS) on Baidu Encyclopedia.

3.2 Experimental Results

In this section, we describe the experimental results of our proposed model and previous state-of-the-art methods on four datasets: Weibo, Chinese Resume, OntoNotes 4, and MSRA. We propose two baselines for comparison, and show the CAN-NER model results. In the experiment results table, we use Baseline to represent a pure BiGRU + CRF model; and Baseline + CNN to indicate the base model with a CNN layer.

3.2.1 Weibo Dataset

Here we compare our proposed model with the latest models on the Weibo dataset.[3] Table 2 shows the F1-scores for named entities (NE), nominal entities (NM, excluding named entities), and both (Overall). We observe that our proposed model achieves state-of-the-art performance.

Existing state-of-the-art systems include Peng and Dredze (2016), He and Sun (2017b), Cao et al. (2018) and Zhang and Yang (2018), which leverage rich external data like cross-domain data, semi-supervised data, and lexicons, or joint-train NER and Chinese Word Segmentation (CWS).[4] In the first block of Table 2, we report the performance of the latest models. Peng and Dredze (2015) propose a model that jointly trains embeddings with NER and it achieves a F1-score of 56.05% on overall performance. The model (Peng and Dredze, 2016) that jointly trains NER and CWS reaches a F1-score of 58.99%. He and Sun (2017b) propose a unified model to exploit cross-domain and semi-supervised data, which improves the F1-score from 54.82% to 58.23% compared with the model proposed by He and Sun (2017a). Cao et al. (2018) use an adversarial transfer learning framework to incorporate task-shared word boundary information from CWS and achieves a F1-score of 58.70%. Zhang and Yang (2018) leverage a lattice structure to integrate lexicon information into their model and achieve a F1-score of 58.79%.

In the second block of Table 2, we give the results of our baselines and proposed models. While the BiGRU + CRF baseline only achieves a F1-score of 53.80%, adding a normal CNN layer as featurizer improves the score to 55.91%. Replacing the CNN with our convolutional attention layer greatly improves the F1-score to 59.31%, which outperforms other models. The improvement demonstrates the effectiveness of our proposed model.

Models	P	R	F1
Zhang and Yang (2018)[1†]	94.53	94.29	94.41
Zhang and Yang (2018)[2‡]	94.07	94.42	94.24
Zhang and Yang (2018)[3]	94.81	94.11	94.46
Baseline	93.71	93.74	93.73
Baseline + CNN	94.36	**94.85**	94.60
CAN-NER Model	**95.05**	94.82	**94.94**

Table 3: Results on Chinese Resume Dataset. For models proposed by Zhang and Yang (2018), 1 represents the char-based LSTM model, 2 indicates the word-based LSTM model and 3 is the Lattice model.

3.2.2 Chinese Resume Dataset

The Chinese Resume test results are shown in Table 3. Zhang and Yang (2018) released the Chinese Resume dataset and they achieve a F1-score of 94.46%. It can be seen that our proposed baseline (CNN + BiGRU + CRF) outperforms Zhang and Yang (2018) with F1-score of 94.60%.

[3] In Table 2,3, 4 and 5, we use $*$ to denote a model with external labeled data for semi-supervised learning. † denotes that the model use external lexicon data. Zhang and Yang (2018) with ‡ is the char-based model in the paper.

[4] The results of Peng and Dredze (2015, 2016) are taken from Peng and Dredze (2017)

Models	P	R	F1
Yang et al. (2016)	65.59	71.84	68.57
Yang et al. (2016)*	72.98	**80.15**	**76.40**
Che et al. (2013)*	**77.71**	72.51	75.02
Wang et al. (2013)*	76.43	72.32	74.32
Zhang and Yang (2018)†	76.35	71.56	73.88
Zhang and Yang (2018)‡	74.36	69.43	71.81
Baseline	70.67	71.64	71.15
Baseline + CNN	72.69	71.51	72.10
CAN-NER Model	75.05	72.29	73.64

Table 4: Results on OntoNotes

Models	P	R	F1
Chen et al. (2006)	91.22	81.71	86.20
Zhang et al. (2006)*	92.20	90.18	91.18
Zhou et al. (2013)	91.86	88.75	90.28
Lu et al. (2016)	-	-	87.94
Dong et al. (2016)	91.28	90.62	90.95
Cao et al. (2018)	91.30	89.58	90.64
Yang et al. (2018a)	92.04	91.31	91.67
Zhang and Yang (2018)†	**93.57**	**92.79**	**93.18**
Baseline	92.54	88.20	90.32
Baseline + CNN	92.57	92.11	92.34
CAN-NER Model	**93.53**	92.42	**92.97**

Table 5: Results on MSRA dataset

Adding our convolutional attention leads a further improvement and achieves state-of-the-art F1-score of 94.94%, which further demonstrates the effectiveness of our proposed model.

3.2.3 OntoNotes Dataset

Table 4 shows comparisons on the OntoNotes 4 dataset. The first block in the table lists the performance of previous methods for Chinese NER. Yang et al. (2016) propose a model combining neural and discrete feature, e.g., POS tagging features, CWS features and orthographic features, improving the F1-score from 68.57% to 76.40%. Leveraging bilingual data, Che et al. (2013) and Wang et al. (2013) achieves F1-scores of 74.32% and 73.88% respectively. Zhang and Yang (2018)‡ is a recent model that uses a character-based model with bichar and softword.

The second block of Table 4 shows the results of our baselines and proposed model. Consistently with observations on the Weibo and Resume datasets, our Convolutional Attention layer leads to a substantial increment on F1-score. Our proposed model achieves a competitive F1-score of 73.64% among character-based model without using external data (e.g., Zhang and Yang (2018)‡).

3.2.4 MSRA Dataset

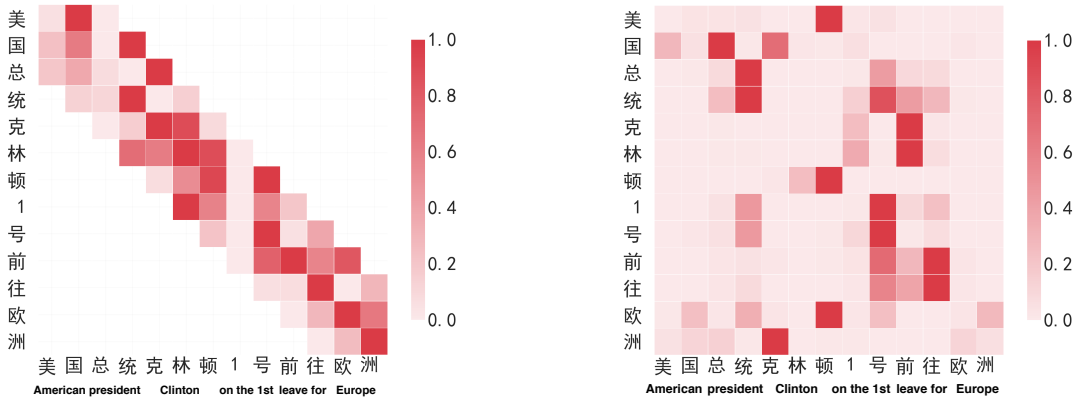
Table 5 shows experiment results on the MSRA 2006 dataset. Chen et al. (2006), Zhang et al. (2006), and Zhou et al. (2013) leverage rich handcrafted features and Lu et al. (2016) exploit multi-prototype embedding features. Dong et al. (2016) introduce radical features into LSTM-CRF. Cao et al. (2018) make use of Adversarial Transfer Learning and global self-attention to improve model performance. Yang et al. (2018a) propose a character-based CNN-BiLSTM-CRF model to incorporate stroke embeddings and generate n-gram features. Zhang and Yang (2018) introduce a lattice structure to incorporate lexicon information into the neural network, which actually includes word embedding information. Although this model achieves state-of-the-art F1-score at 93.18%, it leverages external lexicon data and thus the result is dependent on the quality of the lexicon. At the bottom section of the table, we can see that Baseline + CNN already outperforms most previous methods. Compared with Zhang and Yang (2018), our char-based method achieves a competitive F1-score of 92.97% without any additional lexicon data and word embedding information. Moreover, CAN-NER model achieves state-of-the-art result among the character-based models.

3.3 Discussion

This section discusses the model effectiveness and the experimental results.

3.3.1 Effectiveness of Convolutional Attention and Global Self-Attention

As shown in Tables 2, 3, and 5, our proposed model's performance demonstrates the effectiveness of the Convolutional Attention Network. To better evaluate the effect of the Attention Mechanism, we visualize the normalized attention weights α_m^l for each window from Eq. 2, as in Figure 3a. Each row of the matrix represents location attention weights in each window. For example, the third row indicates that the relationship between center character "总" and contexts "美 国 总 统 克". We can see from the Figure 3a that the word-level features can be extracted through the local attention. In the context, the center character "美" tends to have a stronger connection with



(a) Local attention.

(b) Global self-attention.

Figure 3: Attention visualization. The left-side image shows the normalized Convolutional Attention weights in each window in a sentence. The right-side indicates the global self-attention weights for the whole sentence. In both pictures, the x-axis represents context, while the y-axis represents the input query in the attention mechanism.

its related character "国", which means they have a higher probability of forming the Chinese word "美国 (American)". Also for characters "克", "林", and "顿", they tend to have a strong connection because "克林顿" means "Clinton". Characters "欧" and "洲" also have strong connections, as seen in Figure 3a, because "欧洲" represents "Europe" in Chinese. Therefore, both experiment results and visualization verifies that the Convolutional Attention is effective in obtaining phrase-level information between adjacent characters.

In Figure 3b, we visualize the global self-attention matrix. From the picture, we can find that global self-attention can capture the sentence context information from the long-distance relationship of words to overcome the limitation of Recurrent Neural Networks. For the word "克林顿 (Clinton)", the global self-attention learns the dependencies with "前往 (leave for)" and "1号 (on the 1st)". Distinguished by the red color, "克林顿 (Clinton)" has a stronger connection with "前往 (leave for)" than with "1号 (on the 1st)", which matches the expectation that the predicate in a sentence provides more information to the subject than adverbs of time.

3.3.2 Results Analysis

Our proposed model outperforms previous work on th eWeibo and Chinese Resume datasets and reaches competitive results on both MSRA and OntoNotes 4 datasets without using any external resources. The experiments results demonstrate the effectiveness of our proposed model, especially among char-based models. The performance improvement after adding Convolutional Attention Layer and Global Attention Layer verifies that our model can capture the relationship between character and its local context, as well as the relationship between word and global context. However, although we can obtain comparable or better results to other models that utilize no external resources, we find that our model performance on the OntoNotes 4 dataset still has room for improvement (2.76% F1-score gap to the best model that leverages additional data). This may be explained by specific discrete features and external resources (e.g., other labeled data or lexicons) having a more positive influence on this specific dataset, while CAN-NER cannot learn enough information from only the training set. However, we were not able to identify the precise contributors to the gap based on the available corresponding resources.

4 Related Work

4.1 Neural Network Models

Neural networks, such as LSTM and CNN, have been shown to outperform conventional machine learning methods without requiring handcrafted features. Collobert et al. (2011) describe a CNN-CRF model that reaches competitive results compared to the best statistical models at the time. More recently, the LSTM-CRF architecture has become a quasi-standard on NER tasks. Huang et al. (2015) employed BiLSTM to extract word-

level context information and Lample et al. (2016) further introduced a hierarchy structure by incorporating BiLSTM-based character embeddings. Multiple recent works integrating word-level information and character-level information have been found to achieve improved performance (dos Santos et al., 2015; Chiu and Nichols, 2016; Ma and Hovy, 2016; Lample et al., 2016; Chen et al., 2019). Moreover, external knowledge has also been exploited for NER, as has character-level knowledge, both pre-trained (Peters et al., 2017) and co-trained (Liu et al., 2018). More recently, large-scale pre-trained language representations with deep language models have been proposed to help improve the performance of downstream NLP tasks. E.g., ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).

4.2 Attention Mechanism

Also, Attention Mechanisms have shown very good performance on a variety of tasks including machine translation, machine comprehension, and related NLP tasks (Vaswani et al., 2017; Seo et al., 2016; Tan et al., 2018a). In language understanding, Shen et al. (2018) exploit self-attention to learn long range dependencies. Rei et al. (2016) proposed a model employing an attention mechanism to combine the character-based representation with the word embedding instead of simply concatenating them. This method allows the model to dynamically decide which source of information to use for each word, and therefore outperforming the concatenation method used in previous work. More recently, Tan et al. (2018b) and Cao et al. (2018) employ self-attention to directly capture the global dependencies of the inputs for NER tasks and demonstrate the effectiveness of self-attention in Chinese NER.

4.3 Chinese NER

Multiple previous efforts have tried to address the Chinese language challenge of not having explicit word boundaries. Traditional models depended on hand-crafted features and CRFs-based models (He and Wang, 2008; Mao et al., 2008) and character-based LSTM-CRF models have been applied to Chinese NER to utilize both character- and radical-level representations (Dong et al., 2016). Peng and Dredze (2015) applied character positional embeddings and proposed a jointly trained model for embeddings and NER. To better integrate word boundary information into Chinese NER model, Peng and Dredze (2016) co-trained NER and word segmentation to improve performance in both tasks. He and Sun (2017b) unified cross-domain learning and semi-supervised learning to obtain information from out-of-domain corpora and in-domain unannotated text. Instead of performing word segmentation first, Recently, Zhang and Yang (2018) proposed constructing a word-character lattice by matching words in texts with a lexicon to avoid segmentation errors. Cao et al. (2018) use an adversarial network to jointly train Chinese NER task and Chinese Word Segmentation tasks to extract task-shared word boundary information. Also, Yang et al. (2018c) leverage character-level BiLSTM to extract higher-level features from crowd-annotations.

5 Conclusion

In this paper, we propose CAN-NER, a Convolutional Attention Network model to improve Chinese NER performance and preclude word embedding and additional lexicon dependencies; thus making the model more efficient and robust. In our model, we implement local-attention CNN and BiGRU with the global self-attention structure to capture word-level features and context information with char-level features. Extensive experiments show that our model outperforms the state-of-art systems on the different domain datasets.

Acknowledgements

We'd like to thank our colleague Börje Karlsson for his contribution and support in this work, as well as thank our colleagues Haoyan Liu, Zijia Lin, and the anonymous reviewers for their valuable feedback.

References

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 52–62.

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.

Hui Chen, Zijia Lin, Guiguang Ding, Jianguang Lou, Yusen Zhang, and Börje F. Karlsson. 2019. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. In *Proceedings of AAAI 2019*.

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.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association of Computational Linguistics*, 4(1):357–370.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based lstm-crf with radical-level features for chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*, pages 239–250. Springer.

Pavlina Fragkou. 2017. Applying named entity recognition and co-reference resolution for segmenting english texts. *Progress in Artificial Intelligence*, 6(4):325–346.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690.

Hangfeng He and Xu Sun. 2017a. F-score driven max margin neural network for named entity recognition in chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 713–718.

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

Jingzhou He and Houfeng Wang. 2008. Chinese named entity recognition and word segmentation based on character. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Shen Huang, Xu Sun, and Houfeng Wang. 2017. Addressing domain adaptation for chinese word segmentation with global recurrent structure. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 1, pages 184–193.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

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.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model.

Yanan Lu, Yue Zhang, and Dong-Hong Ji. 2016. Multi-prototype chinese character embedding. In *LREC*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1064–1074.

Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao, and Haila Wang. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1105–1116.

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 *The 54th Annual Meeting of the Association for Computational Linguistics*, page 149.

Nanyun Peng and Mark Dredze. 2017. Supplementary results for named entity recognition on chinese social media with an updated dataset. Technical report.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1756–1765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237.

Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016. Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318.

Cıcero dos Santos, Victor Guimaraes, RJ Niterói, and Rio de Janeiro. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018a. Deep semantic role labeling with self-attention. In *AAAI Conference on Artificial Intelligence*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018b. Deep semantic role labeling with self-attention. In *AAAI Conference on Artificial Intelligence*.

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 5998–6008.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 1, pages 163–172.

Mengqiu Wang, Wanxiang Che, and Christopher D Manning. 2013. Effective bilingual constraints for semi-supervised learning of named entity recognizers. In *AAAI Conference on Artificial Intelligence*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Fan Yang, Jianhu Zhang, Gongshen Liu, Jie Zhou, Cheng Zhou, and Huanrong Sun. 2018a. Five-stroke based cnn-birnn-crf network for chinese named entity recognition. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 184–195. Springer.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018b. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889.

Jie Yang, Zhiyang Teng, Meishan Zhang, and Yue Zhang. 2016. Combining discrete and neural features for sequence labeling. *In CICLing*.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 839–849.

YaoSheng Yang, Meishan Zhang, Wenliang Chen, Wei Zhang, Haofen Wang, and Min Zhang. 2018c. Adversarial learning for chinese ner from crowd annotations. In *AAAI Conference on Artificial Intelligence*.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Suxiang Zhang, Ying Qin, Juan Wen, and Xiaojie Wang. 2006. Word segmentation and named entity recognition for sighan bakeoff3. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 158–161.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1554–1564.

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.