

Deformable Stacked Structure for Named Entity Recognition

Shuyang Cao and Xipeng Qiu and Xuanjing Huang

Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{caosy14,xpqi,xjhuang}@fudan.edu.cn

Abstract

Neural architecture for named entity recognition has achieved great success in the field of natural language processing. Currently, the dominating architecture consists of a bi-directional recurrent neural network (RNN) as the encoder and a conditional random field (CRF) as the decoder. In this paper, we propose a deformable stacked structure for named entity recognition, in which the connections between two adjacent layers are dynamically established. We evaluate the deformable stacked structure by adapting it to different layers. Our model achieves the state-of-the-art performances on the OntoNotes dataset.

Introduction

Named entity recognition (NER) is a subtask of sequence labeling. It is similar to other sequence labeling tasks considering its working procedure that is assigning a certain label to each token of a sequence. But unlike part-of-speech (POS) tagging and other sequence labeling task evaluated on accuracy, the performance of an NER system is evaluated on the whole named entity using precision, recall, and f1 score. Thus, the output of an NER system at each position is not independent with each other and much related to its neighboring positions. The most common assumption of sequence labeling is the Markov property that the choice of label for a particular token is directly dependent only on the immediately adjacent labels; hence the labels of all the tokens in a sequence form a Markov chain. Therefore, the widely-used statistical models for sequence labeling involve hidden Markov model (HMM) (Rabiner and Juang 1986), maximum entropy Markov model (MEMM) (McCallum, Freitag, and Pereira 2000) and conditional random field (CRF) (Lafferty, McCallum, and Pereira 2001).

In recent years, neural network architectures for NER (Huang, Xu, and Yu 2015; Ma and Hovy 2016; Strubell et al. 2017) are proposed to reduce the efforts of feature engineering and the model complexity and have achieved great success. Currently, the dominative neural NER architecture consists of a bi-directional recurrent neural network (RNN) as the encoder and a conditional random field (CRF) as the decoder (Huang, Xu, and Yu 2015). The RNN encoder can effectively extract the context-aware features for each token, avoiding the cost of manually designing features.

Despite of their success, the network architecture of the encoder still need be manually designed for different tasks and lacks flexibility. Due to the inherent hierarchical structure of natural language, the crucial information for a token could appear in a changeable position. Taking the following NER instance for example, there are two entity mentions in the following sentence,

He bought 30 shares of Acme in 2006.

For the token “2006”, its self-information is enough to decide its entity type. But for the token “Acme”, the information from its neighbour “shares” could be more important. Since the position of the crucial information for each token is different, the current rigid network architecture heavily depends on the ability of RNNs to capture the context information.

In this paper, we propose a deformable stacked structure to flexibly choose the most informative features as input. Unlike the vanilla stacked structure, the connections between two adjacent layers are dynamically constructed. The input of each position in the upper layer is dynamically chosen from the lower layer, instead of a fixed position. Specifically, we introduce a dynamical offset to indicate the input’s position. The offsets are dynamically computed according to the current hidden states. To make the whole neural network end-to-end trainable, we further propose an approximate solution to use a continuous offset to softly select the inputs via a bilinear interpolation, instead of the exact discrete offset.

The contributions of the paper can be summarized as follows.

1. We propose a deformable stacked structure, whose stacking connections are dynamically determined, instead of in a pre-defined way. The deformable stacked way can effectively alleviate the pressure of RNNs for collecting the context information.
2. We also propose an approximate strategy to softly change the connections, which makes the whole neural network differentiable and end-to-end trainable.
3. Compared to the models with rigid network architecture, our model is more flexible and suitable for named entity recognition tasks and achieves the state-of-the-art performances for named entity recognition on OntoNotes dataset.

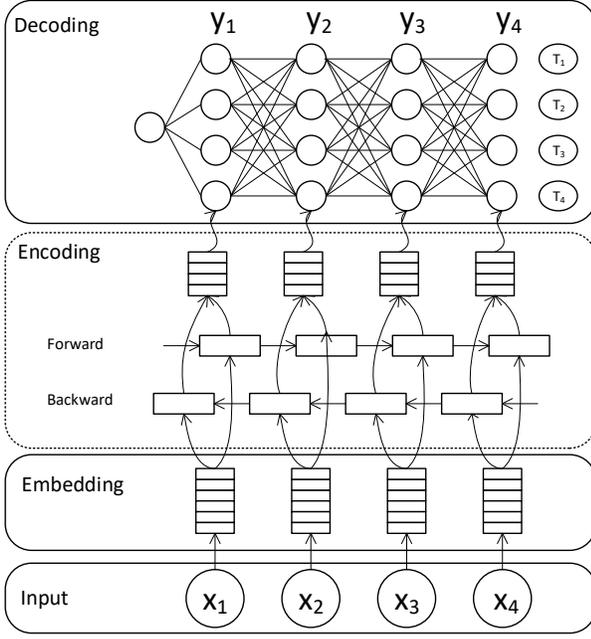


Figure 1: General neural architecture for named entity recognition.

General Neural Architecture for Named Entity Recognition

Given a sequence with n tokens $X = \{x_1, \dots, x_n\}$, the aim of named entity recognition is to figure out the ground truth of labels $Y^* = \{y_1^*, \dots, y_n^*\}$:

$$Y^* = \arg \max_{Y \in \mathcal{T}^n} p(Y|X), \quad (1)$$

where \mathcal{T} is the target set.

There are lots of prevalent methods to solve named entity recognition problem such as maximum entropy Markov model (MEMM), conditional random fields (CRF), etc. Recently, neural models are widely applied to named entity recognition for their ability to minimize the effort in feature engineering (Huang, Xu, and Yu 2015; Ma and Hovy 2016). Moreover, neural models also benefit from the distributed representations, which can enhance the generalization capabilities with the pre-trained word embeddings on the large-scale un-annotated corpus.

The general architecture of neural named entity recognition could be characterized by three components: (1) an embedding layer; (2) encoding layers consisting of several classical neural networks and (3) a decoding layer. The role of encoding layers is to extract features, which could be either convolution neural network or recurrent neural network. In this paper, we adopt the bidirectional long-short-term-memory (BiLSTM) neural networks followed by a CRF as decoding layer. Figure 1 illustrates the general architecture.

Embedding Layer

To represent discrete tokens as distributed vectors, the first step is usually to map them to distributed embedding vectors. Formally, we lookup embedding vector from embedding matrix for each token x_i as $e_{x_i} \in \mathbb{R}^{d_e}$, where d_e is a hyper-parameter indicating the size of embedding.

Encoding Layers

To incorporate information from both sides of sequence, we use bi-directional LSTM with forward and backward directions. Notably, the parameters of two LSTMs with different orientations are independent. The update of each hidden state can be written precisely as follows:

$$\mathbf{h}_i = \vec{\mathbf{h}}_i \oplus \overleftarrow{\mathbf{h}}_i, \quad (2)$$

$$= \text{BiLSTM}(e_{x_{1:n}}, i, \theta), \quad (3)$$

where $\vec{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ are the hidden states at position i of the forward and backward LSTMs respectively; \oplus is concatenation operation; θ denotes all the parameters in BiLSTM model.

Decoding Layer

After extracting features, we employ a conditional random fields (CRF) layer to inference tags. In CRF layer, $p(Y|X)$ in Eq (1) could be formalized as:

$$p(Y|X) = \frac{\Psi(Y|X)}{\sum_{Y' \in \mathcal{T}^n} \Psi(Y'|X)}. \quad (4)$$

Here, $\Psi(Y|X)$ is the potential function, and we only consider interactions between two successive labels (first order linear chain CRFs):

$$\Psi(Y|X) = \prod_{i=2}^n \psi(X, i, y_{i-1}, y_i), \quad (5)$$

$$\psi(\mathbf{x}, i, y', y) = \exp(s(X, i)_y + b_{y'y}), \quad (6)$$

where $b_{y'y} \in \mathbb{R}$ is transition parameter, indicating how possible a label y' will transfer to another label y . Score function $s(X, i) \in \mathbb{R}^{|\mathcal{T}|}$ assigns score for each label on tagging the i -th character:

$$s(X, i) = \mathbf{W}_s^\top \mathbf{h}_i + \mathbf{b}_s, \quad (7)$$

where \mathbf{h}_i is the hidden state of BiLSTM at position i ; $\mathbf{W}_s \in \mathbb{R}^{d_h \times |\mathcal{T}|}$ and $\mathbf{b}_s \in \mathbb{R}^{|\mathcal{T}|}$ are trainable parameters.

At test phase, the Viterbi algorithm is employed to decode the best target sequence in polynomial time complexity.

Deformable Stacked Structure

The critical factor of neural named entity recognition models is the encoding layer, whose role is to extract useful features to judge the label of each token. To better model the complex compositional features, we could increase the depth of neural network by stacking the recurrent encoding layers.

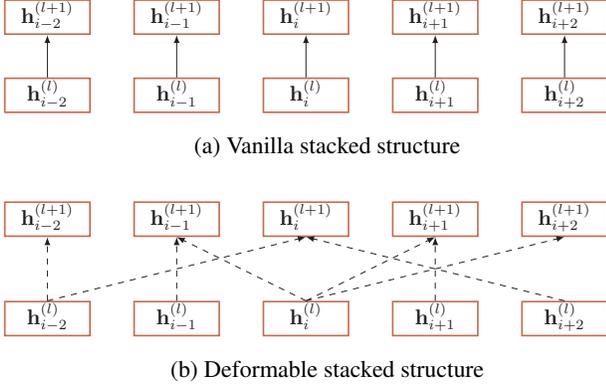


Figure 2: Two ways to stack layers. The dashed lines denote the connections are dynamically changed.

Vanilla Stacked Structure A conventional way to stack layers is to take the output of the lower layer as the input of upper layer at each position (Pascanu et al. 2013). For each position i at $(l + 1)$ -th layer, its input is taken from position i at l -th layer.

$$\mathbf{h}_i^{(l+1)} = f(\mathbf{h}_i^{(l)}, \theta^{(l+1)}) \quad (8)$$

where f is a non-linear function.

Despite being successful for NER, the vanilla stacked structure has a limitation of the fixed geometric structures. The stacking structures are manually designed and lack flexibility.

Deformable Stacked Structure Due to the inherently hierarchical structure of natural language, the crucial information for a token could appear in a changeable position. As usually seen in NER, while in some case the information of a token itself is enough to decide its entity type, in another case the information of its neighbors could be more important.

To flexibly capture the most informative features, we propose a deformable stacked structure to choose the input's position from the lower layers dynamically. For each position i at $(l + 1)$ -th layer, its input is taken from position $i + o$ at l -th layer.

$$\mathbf{h}_i^{(l+1)} = f(\mathbf{h}_{i+o}^{(l)}, \theta^{(l+1)}) \quad (9)$$

where o is a offset, $o \in \mathbb{Z}$.

The offsets o is predicted by an extra module based on the hidden states of the lower layers.

We can adapt deformable stacked structure between different layers, which we will further explain in the experiment section.

Figure 2 gives the comparison of two different stacked structure.

Differentiable Deformable Stacked Structure

In our proposed deformable stacked structure, the index of the lower layer is discrete, which results in a non-differentiable network. Although we can use reinforcement

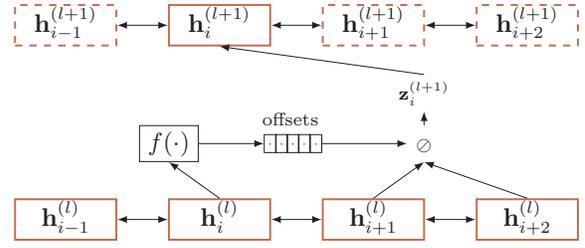


Figure 3: A differentiable implementation of deformable stacked structure. $f(\cdot)$ is defined by Eq (10), \oslash denotes the bilinear interpolation.

learning to learn the parameters, in this paper, we propose a differentiable variant to utilize the strengths of back-prorogation of the neural network.

Inspired by (Dai et al. 2017), we use bilinear interpolation to replace the exact discrete index. Figure 3 shows the architecture of deformable stack.

Offset Learning Instead of a discrete offset, we use a continuous offset to make the whole network to be differentiable. The offset is calculated by a simple function. The offset at the position i can be given as:

$$o_i^{(l)} = \mathbf{v}^T \mathbf{h}_i^{(l)}, \quad (10)$$

where $\mathbf{v} \in \mathbb{R}^d$ is the parameter vector, $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$ is the hidden states at position i of the (l) -th layer.

Deformable Input To bridge the continuous offset and discrete position, we apply bilinear interpolation to select the inputs from the lower layer softly.

For each position i at $(l + 1)$ -th layer, its input $\mathbf{z}_i^{(l+1)}$ can be given as:

$$\mathbf{z}_i^{(l+1)} = \sum_{j=1}^n g(i + o_i^{(l)}, j) \cdot \mathbf{h}_j^{(l)} \quad (11)$$

where $g(i + o_i^{(l)}, j)$ is a bilinear interpolation kernel and can be given as:

$$g(i + o_i^{(l)}, j) = \max(0, 1 - |i + o_i^{(l)} - j|), \quad (12)$$

thus $\{g(i + o_i^{(l)}, j)\}_{j=1}^n$ can be regarded as a mask vector which has two non-zero elements at most.

For example, assuming that the length of input sequence is 5, and the current position is 2, a continuous offset $o = 1.2$ gives the mask vector

$$\mathbf{g} = [0, 0, 0.8, 0.2, 0]. \quad (13)$$

By the above strategy, we can make the whole neural network to be differentiable, which can be end-to-end trained efficiently.

Multi-Offset Extension

To make the input of each recurrent layer more flexible, we can allow it to choose information from the multiple positions in the lower layer.

Therefore, a simple extension of our model is the multi-offset deformable structure, which allows the model to utilize the information at different positions jointly.

In multi-offset extension, we can predict k offsets at position i ,

$$\mathbf{o}_i^{(l)} = V\mathbf{h}_i^{(l)}, \quad (14)$$

where $\mathbf{o}_i^{(l)} \in \mathbb{R}^k$ is a vector consisting of k offsets, and $V \in \mathbb{R}^{k \times d}$ is the parameter matrix.

With k offsets, we can obtain k deformable inputs $\mathbf{z}_{i_1}^{(l+1)}, \mathbf{z}_{i_2}^{(l+1)}, \dots, \mathbf{z}_{i_k}^{(l+1)}$ via bilinear interpolation. Then we concatenate these inputs to get the final input of position i to the next layer.

$$\mathbf{z}_i^{(l+1)} = \mathbf{z}_{i_1}^{(l+1)} \oplus \mathbf{z}_{i_2}^{(l+1)} \oplus \dots \oplus \mathbf{z}_{i_k}^{(l+1)}, \quad (15)$$

where \oplus indicates the concatenation operation.

Wide-Window Extension

Another extension of our model is to involve neighbor hidden states when calculating the offsets.

In wide-window extension, we can predict k offsets at position i with window size w by a convolutional neural network,

$$\mathbf{o}_i^{(l)} = W\mathbf{h}_{i:i+d}^{(l)} \quad (16)$$

$$\mathbf{O}^{(l)} = \text{Conv}(\mathbf{H}, W, w), \quad (17)$$

where $\mathbf{O} \in \mathbb{R}^{n \times k}$ is a matrix consisting of k offsets in n positions. $W \in \mathbb{R}^{w \times k \times d}$ is the parameter matrix of the convolutional neural network.

Training

Given a trainset $(X^{(n)}, Y^{(n)})_{n=1}^N$, the objective is to minimize the cross entropy loss $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_n \log p(Y^{(n)} | X^{(n)}) + \lambda \|\theta\|^2, \quad (18)$$

where θ represents all the parameters, λ represents the regularizer factor.

We use stochastic gradient descent with a momentum of 0.9. The initial learning rate is set according to the dataset and task. To avoid overfitting, dropout is applied after each recurrent or convolutional layer.

Initialization We take advantage of pre-trained word embeddings such as Glove (Pennington, Socher, and Manning 2014) to transfer more knowledge from large unlabeled data. For the words that don't appear in Glove, we randomly initialize their embeddings from a normal distribution with mean 0 and standard deviation $\sqrt{\frac{1}{dim}}$ following (Ma and Hovy 2016), dim is the dimension of word embedding.

Dataset	Train	Dev	Test
CoNLL-2003	204,567	51,578	46,666
OntoNotes 5.0 (CoNLL-2012)	1,088,503	147,724	152,728

Table 1: Number of tokens in different dataset.

The network weights are initialized with Xavier normalization (Glorot and Bengio 2010) to maintain the variance of activations throughout the forward and backward passes. Biases are uniformly set to zero when the network is constructed.

Character Embedding Following (Ma and Hovy 2016), we also apply convolutional neural network (CNN) to extract character-level features of words. The character-level feature of words helps the model better handle the OOV (out of vocabulary) problems. For each character, we randomly initialize its embedding from a normal distribution with mean 0 and standard deviation 1.

Experiment

We consider three different kinds of deformable stacked structure in our paper.

1. Deformable stacked structure between BiLSTM layers.
2. Deformable stacked structure between the encoder layer (BiLSTM in our paper) and decoder layer (CRF in our paper).
3. Both 1 and 2.

Datasets

We evaluate our model on two datasets: CoNLL-2003 NER dataset (Tjong Kim Sang and De Meulder 2003) dataset and OntoNotes 5.0 (Pradhan et al. 2013) dataset. We adapt structure 2 on CoNLL-2003 dataset and structure 1, 2, 3 on OntoNotes 5.0 dataset.

For both datasets, we perform the experiments on the English portion.

For the OntoNotes 5.0 dataset, we split the data according to the CoNLL-2012 shared task following (Chiu and Nichols 2016a). The Pivot Text portion is excluded because it lacks gold annotations for named entities.

The original tags are converted to tags of "BIOES" (begin, inside, outside, end, singleton) tagging.

The CoNLL-2003 dataset contains only 4 types of named entities: PERSON, LOCATION, ORGANIZATION, AND MISCELLANEOUS, while the OntoNotes 5.0 dataset contains 18 types of named entities, including works of art, dates, cardinal numbers, languages, and events. With BIOES tagging, we have 18 labels for CoNLL-2003 dataset and 74 labels for OntoNotes 5.0 dataset (including padding label).

For the digits in both of the datasets, we replace them with digit 0. The details of the two datasets are shown in Table 1.

	CoNLL 2003	OntoNotes 5.0
Offset CNN window size	3	3
Number of offsets k	3	3
Char-CNN filters	30	30
Char-CNN window size	3	3
Size of LSTM state	256	200
LSTM layers	1	2
Learning rate	0.008	0.005
Dropout	0.5	0.5
Batch size	10	8

Table 2: Hyper-parameters.

Model	P	R	F
(Finkel and Manning 2009)*	84.04	80.86	82.42
(Ratinov and Roth 2009)	82.00	84.95	83.45
(Durrett and Klein 2014)	85.22	82.89	84.04
(Chiu and Nichols 2016b)	86.16	86.65	86.40
(Strubell et al. 2017)	-	-	86.84
(Strubell et al. 2017)**	-	-	86.99
BiLSTM-CNN-CRF	87.03	86.97	87.00
Deformable stacked structure	88.02	88.01	88.01

Table 3: Performances on OntoNotes 5.0 dataset. * denotes the result from (Pradhan et al. 2013). ** denotes the baseline from their paper. (Finkel and Manning 2009): joint parsing and NER model. (Ratinov and Roth 2009): using many resources, such as Wikipedia, non-local features. (Durrett and Klein 2014): combining coreference resolution, entity linking, and NER into a single CRF model with cross-task interaction factors. (Chiu and Nichols 2016b): BiLSTM-CRF network with with many composite features. (Strubell et al. 2017): iterated dilated CNN and CRF.

Hyper-parameters

Hyper-parameters of our models are shown in Table 2.

For the word embedding, we use Glove pre-trained embedding of 100 dimensions on 6 billion words.

Result

The results of our model on OntoNotes 5.0 and CoNLL-2003 dataset datasets are shown in Table 3 and Table 4.

On both of the two datasets, our model outperforms the vanilla stacked BiLSTM-CNN-CRF baseline. We also compare our model with existing models on the two datasets. On OntoNotes dataset, our model also achieves the state-of-the-art result.

Deformable stacked structure utilizes the feature of neighbor words. Thus, it should make better predictions when OOV occurs. We conduct experiments on models without character embedding. The results are shown in Table 5.

Comparison of different deformable stacked structure

On OntoNotes dataset, we adopt three different deformable

Model	P	R	F
(Collobert et al. 2011)	-	-	86.96
(Luo et al. 2015)	-	-	91.20
(Lample et al. 2016)	-	-	90.33
(Ma and Hovy 2016)	91.35	91.06	91.21
(Strubell et al. 2017)	-	-	90.54
Bi-LSTM-CNN-CRF	91.03	91.11	91.07
Deformable stacked structure	91.01	91.24	91.12

Table 4: Performances on CoNLL-2003 dataset. (Collobert et al. 2011): a earlier neural model. (Luo et al. 2015): joint NER/entity linking model. (Lample et al. 2016): BiLSTM-CRF with RNN-based char information. (Ma and Hovy 2016): BiLSTM-CRF with CNN-based char information. (Strubell et al. 2017): iterated dilated CNN and CRF.

stacked structures. We evaluate the performance of the three different structure with varying numbers of offsets.

As we can see from Table 6, the deformable stacked structure between LSTM layers has greater improvement compared with the deformable stacked structure between the encoder layer and the decoder layer. These two structures differ on the function of their next layers, and the difference in improvement also comes from that. For structure 1, the deformable stacked structure reconstructs the input of the next LSTM layer. Thus, it plays a role of feature augmentation for the next LSTM layer to better extract related information. And better extraction of feature also helps with the decoding procedure. For structure 2, the deformable stacked structure reconstructs the input of the CRF layer and plays a role of feature selection for decoding.

Taking information from multiple positions helps as increasing the number of offsets shows improvement.

We can also draw a significant character of the deformable stacked structure from Table 6. Compared with the baseline model, the deformable stacked structure has great improvement in recall. It indicates that the deformable stacked structure can discover more named entity in the sentence.

Offsets Visualization

To give an intuitive impression of how the offset dynamically changing, we show the kernel density estimation of offset values of different deformable stacked structures on the OntoNotes test set in Figure 4.

For the multiple offsets setting, we give the kernel density estimation of each offset. From Figure 4, we can see that these offsets have the normal distribution in general. For multiple offsets of deformable stacked structure between LSTM layers, we observe that at least one of the offsets has a distribution similar to a normal distribution with mean 0 and the means of the rest distribution are slightly shifted from 0. For multiple offsets of deformable stacked structure between the encoder layer and the decoder layer, the distributions shift from 0 significantly.

Model	CoNLL-2003			OntoNotes 5.0		
	P	R	F	P	R	F
BiLSTM-CRF	87.97	86.49	87.22	85.26	85.87	85.56
Deformable stacked structure	87.91	86.79	87.33	85.79	86.14	85.96

Table 5: Performances of models without character embedding on test set.

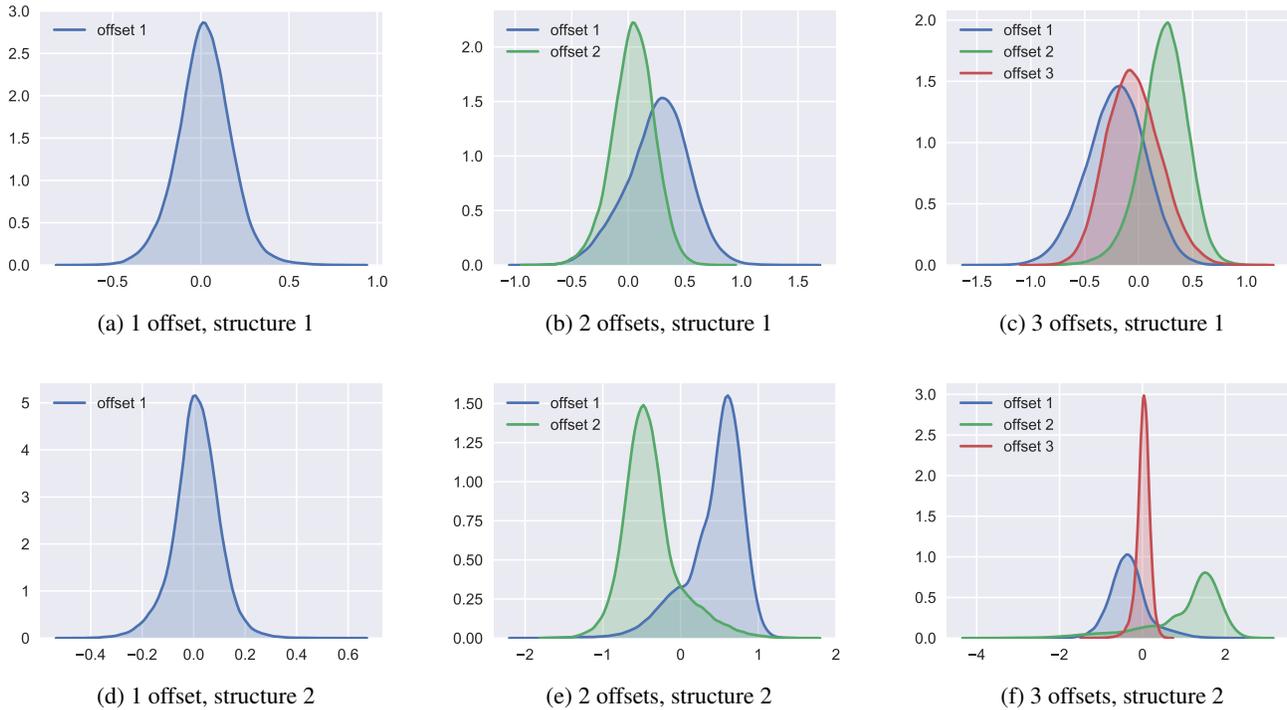


Figure 4: Kernel density estimation of offsets on test set. Y-axis represents the kernel density, and X-axis represents different values of offset. (a)(d): offset number $k = 1$. (b)(e): offset number $k = 2$. (c)(f): offset number $k = 3$. Structure 1: deformable stacked structure between LSTM layers. Structure 2: deformable stacked structure between LSTM layers and CRF. Structure 3: both structure 1 and structure 2.

Case Study

We also visualize the offsets of a real case of structure 3 with offset number $k = 3$ in Figure 5. For simplicity, we only display the part of sentences that have different predictions from the vanilla stacked model.

With deformable stacking, our model can better recognize the borders of named entities. We can see that the positions that our model predicts correctly while the vanilla stacked structure predicts wrongly have similar offsets between the CRF layer and the second LSTM layer. These offsets provide features from the following positions and let the model know that the named entity doesn't break at the word "Faith". Correctly predicting the borders of named entities shows great performance improvement when using BIOES tagging.

Related Work

There are mainly two lines of work related to ours.

One is the neural architecture for named entity recognition. Recently, several different neural network architectures have been proposed and successfully applied to NER. Among these neural architectures, BiLSTM+CRF (Huang, Xu, and Yu 2015) has become a fundamental architecture, which consists of a bi-directional LSTM as the encoder and a conditional random field (CRF) as the decoder. Some work (Chiu and Nichols 2016a; Ma and Hovy 2016; Chen, Qiu, and Huang 2017) also introduced a CNN layer before BiLSTM layer to model character-level information and achieved better performances. Besides BiLSTM, there is also some work to adopt CNN as encoder to capture the context information. (Strubell et al. 2017) use a dilated convolutional neural networks to efficiently aggregate broad context information.

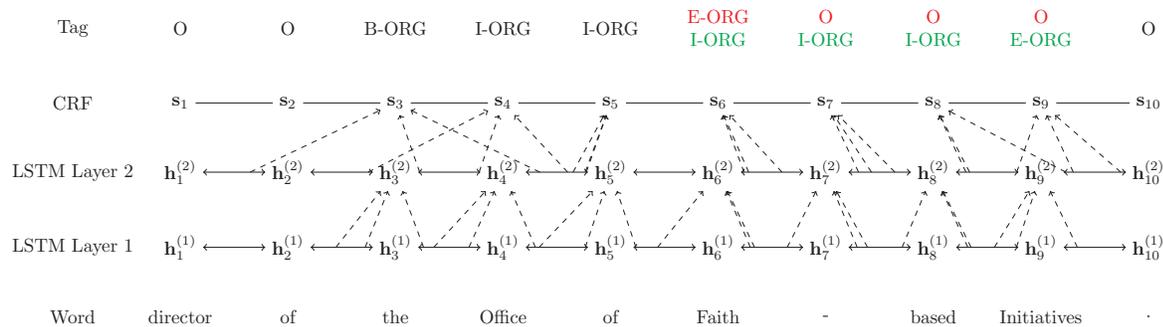


Figure 5: A real case on OntoNotes. The dashed lines indicate the dynamic offsets in our model. The red tags denote wrong predictions by the vanilla stacked structure, and the green tags denote the correct predictions by our model. The black tags denote the correct predictions by both models.

Structure	# of offsets	P	R	F
baseline	-	87.03	86.97	87.00
1	1	87.14	88.66	87.81
1	2	86.96	88.68	87.81
1	3	87.34	88.51	87.83
2	1	87.20	87.86	87.53
2	2	86.92	88.18	87.55
2	3	86.99	88.27	87.63
3	1	87.24	88.56	87.90
3	2	87.15	88.81	87.97
3	3	87.64	88.38	88.01

Table 6: Effects of different structure setting on OntoNotes 5.0 dataset. Structure 1: deformable stacked structure between LSTM layers. Structure 2: deformable stacked structure between LSTM layers and CRF. Structure 3: both structure 1 and structure 2.

Compared to these models, our model can effectively increase the input width of stacked layers and help aggregate more broad context.

Another is neural architecture search (Pham et al. 2018; Zoph and Le 2016; Zhang, Huang, and Zhao 2018). Neural architecture search aims to automatically design the architecture of neural networks for a specific task. The current methods mainly adopt reinforcement learning to maximize the expected accuracy of the generated architectures on a validation set.

Our model can be regarded as a “lightweight” architecture search model, and changes the connections between the adjacent stacked layers. Moreover, we use an approximate strategy to change the connections softly.

Conclusion

We present deformable stacked structure, in which connections between two adjacent layers are dynamically generated. Three different deformable stacked structures are designed and evaluated. Moreover, we also propose an ap-

proximate strategy to softly change the connections, which makes the whole neural network differentiable and end-to-end trainable. Our model achieves the state-of-the-art performances on the OntoNotes dataset.

There are several potential directions for future work. First, we hope to extend this work to build more flexible neural architecture. We have already established deformable connections between the most of layers of the encoder, but some layers are still vanilla stacked such as the embedding layer. Another exciting direction is to apply our model to other NLP tasks, such as parsing. Since our model does not require any task-specific knowledge, it might be effortless to apply it to these tasks.

References

- Chen, X.; Qiu, X.; and Huang, X. 2017. A feature-enriched neural model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 3960–3966.
- Chiu, J., and Nichols, E. 2016a. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association of Computational Linguistics* 4:357–370.
- Chiu, J., and Nichols, E. 2016b. Sequential labeling with bidirectional LSTM-CNNs. In *Proc. International Conf. of Japanese Association for NLP*, 937–940.
- Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* 12:2493–2537.
- Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 764–773.
- Durrett, G., and Klein, D. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association of Computational Linguistics* 2:477–490.
- Finkel, J. R., and Manning, C. D. 2009. Joint parsing and named entity recognition. In *Proceedings of Human*

- Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 326–334. Association for Computational Linguistics.
- Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, 249–256.
- Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.
- Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270. Association for Computational Linguistics.
- Luo, G.; Huang, X.; Lin, C.-Y.; and Nie, Z. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 879–888. Association for Computational Linguistics.
- Ma, X., and Hovy, E. 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: Long Papers)*, 1064–1074. Association for Computational Linguistics.
- McCallum, A.; Freitag, D.; and Pereira, F. C. N. 2000. Maximum markov models for information extraction and segmentation. In *Proceedings of the 17th International Conference on Machine Learning*, 591–598.
- Pascanu, R.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*.
- Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)* 12:1532–1543.
- Pham, H.; Guan, M. Y.; Zoph, B.; Le, Q. V.; and Dean, J. 2018. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*.
- Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 143–152. Association for Computational Linguistics.
- Rabiner, L., and Juang, B. 1986. An introduction to hidden markov models. *IEEE ASSP Magazine* 3(1):4–16.
- Ratinov, L., and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, 147–155. Association for Computational Linguistics.
- Strubell, E.; Verga, P.; Belanger, D.; and McCallum, A. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2670–2680. Association for Computational Linguistics.
- Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Daelemans, W., and Osborne, M., eds., *Proceedings of CoNLL-2003*, 142–147. Edmonton, Canada.
- Zhang, T.; Huang, M.; and Zhao, L. 2018. Learning structured representation for text classification via reinforcement learning.
- Zoph, B., and Le, Q. V. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.