Unified Named Entity Recognition as Word-Word Relation Classification

Jingye Li,[1,*] Hao Fei,[1,*] Jiang Liu,[1] Shengqiong Wu,[1]
Meishan Zhang,[2] Chong Teng,[1] Donghong Ji,[1] Fei Li[1,†]
[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China
[2] Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen), China
{theodorelee, hao.fei, liujiang, whuwsq, tengchong, dhji, lifei_csnlp}@whu.edu.cn, mason.zms@gmail.com

Abstract

So far, named entity recognition (NER) has been involved with three major types, including flat, overlapped (aka. nested), and discontinuous NER, which have mostly been studied individually. Recently, a growing interest has been built for unified NER, tackling the above three jobs concurrently with one single model. Current best-performing methods mainly include span-based and sequence-to-sequence models, where unfortunately the former merely focus on boundary identification and the latter may suffer from exposure bias. In this work, we present a novel alternative by modeling the unified NER as word-word relation classification, namely W^{2}NER. The architecture resolves the kernel bottleneck of unified NER by effectively modeling the neighboring relations between entity words with `Next-Neighboring-Word` (NNW) and `Tail-Head-Word-*` (THW-$\star$) relations. Based on the W^{2}NER scheme we develop a neural framework, in which the unified NER is modeled as a 2D grid of word pairs. We then propose multi-granularity 2D convolutions for better refining the grid representations. Finally, a co-predictor is used to sufficiently reason the word-word relations. We perform extensive experiments on 14 widely-used benchmark datasets for flat, overlapped, and discontinuous NER (8 English and 6 Chinese datasets), where our model beats all the current top-performing baselines, pushing the state-of-the-art performances of unified NER.

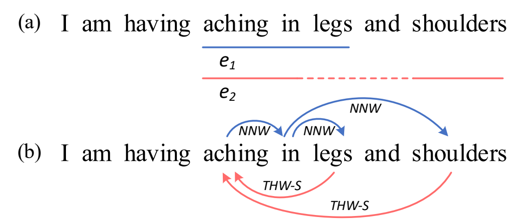

Figure 1: (a) An example to show three types of NER. e_1 is a flat entity overlapped with a discontinuous entity e_2 at the span "aching in". (b) We formalize three NER subtasks as word-word relation classification, where the `Next-Neighboring-Word` (NNW) relation indicates that a word pair are successively joint as a segment of an entity (e.g., aching→in), and the `Tail-Head-Word-*` (THW-$\star$) relation implies the edges where the tail words connect to the head words (e.g., legs→aching) as an entity with "$\star$" type (e.g., *Symptom*).

Introduction

Named entity recognition (NER) has long been a fundamental task in natural language processing (NLP) community, due to its wide variety of knowledge-based applications, e.g., relation extraction (Wei et al. 2020; Li et al. 2021b), entity linking (Le and Titov 2018; Hou et al. 2020), etc. Studies of NER have gradually evolved initially from the flat NER (Lample et al. 2016; Strubell et al. 2017), late to the overlapped NER (Yu et al. 2020; Shen et al. 2021), and recently to the discontinuous NER (Dai et al. 2020; Li et al. 2021a). Specifically, flat NER simply detects the mention spans and their semantic categories from text, while the problems in overlapped and discontinuous NER become more complicated, i.e., overlapped entities contain the same tokens,[1] and discontinuous entities entail non-adjacent spans, as illustrated in Figure 1.

Previous methods for multi-type NER can be roughly grouped into four major categories: 1) *sequence labeling*, 2) *hypergraph-based methods*, 3) *sequence-to-sequence methods* and 4) *span-based methods*. A majority of initial work formalizes NER as a sequence labeling problem (Lample et al. 2016; Zheng et al. 2019; Tang et al. 2018; Straková et al. 2019), assigning a tag to each token. However, it is difficult to design one tagging scheme for all NER subtasks. Then hypergraph-based models are proposed (Lu and Roth 2015; Wang and Lu 2018; Katiyar and Cardie 2018) to represent all entity spans, which however suffer from both the spurious structure and structural ambiguity issue during inference. Recently, Yan et al. (2021) propose a sequence-to-sequence (Seq2Seq) model to directly generate various entities, which unfortunately potentially suffers from the decoding efficiency problem and certain common shortages of Seq2Seq architecture, e.g., exposure bias. Span-based meth-

*Equal contribution
†Corresponding author

[1] Without losing generality, "nested" can be seen as a special case of "overlapped" (Zeng et al. 2018; Dai 2018; Fei et al. 2020).

ods (Luan et al. 2019; Li et al. 2021a) are another state-of-the-art (SoTA) approaches for unified NER, enumerating all possible spans and conduct span-level classification. Yet the span-based models can be subject to maximal span lengths and lead to considerable model complexity due to the enumerating nature. Thus, designing an effective unified NER system still remains challenging.

Most of the existing work has paid the major focus on how to accurately identify the entity boundary, i.e., the kernel problem of NER, especially for flat one (Straková et al. 2019; Fei et al. 2021). However, after carefully rethinking the common characteristics of all three types of NER, we find that the bottleneck of unified NER more lies in the modeling of the neighboring relations between entity words. Such adjacency correlations essentially describe the semantic connectivity between the partial text segments, which especially plays the key role for the overlapping and discontinuous ones. As exemplified in Figure 1(a), it could be effortless to detect the flat mention "aching in legs", since its constituent words all are naturally adjacent. But, to detect out the discontinuous entity "aching in shoulders", effectively capturing the semantic relations between the neighboring segments of "aching in" and "shoulders" is indispensable.

On the basis of the above observation, we in this paper investigate an alternative unified NER formalism with a novel word-word relation classification architecture, namely W^{2}NER. Our method resolves the unified NER by effectively modeling both the entity boundary identification as well as the neighboring relations between entity words. Specifically, W^{2}NER makes predictions for two types of relations, including the `Next-Neighboring-Word` (NNW) and the `Tail-Head-Word-`$\star$ (THW-$\star$), as illustrated in Figure 1(b). The NNW relation addresses entity word identification, indicating if two argument words are adjacent in an entity (e.g., aching$\rightarrow$in), while the THW-$\star$ relation accounts for entity boundary and type detection, revealing if two argument words are the tail and head boundaries respectively of "$\star$" entity (e.g., legs$\rightarrow$aching, *Symptom*).

Based on the W^{2}NER scheme, we further present a neural framework for unified NER (cf. Figure 3). First, BERT (Devlin et al. 2019) and BiLSTM (Lample et al. 2016) are used to provide contextualized word representations, based on which we construct a 2-dimensional (2D) grid for word pairs. Afterwards, we design multi-granularity 2D convolutions to refine the word-pair representations, effectively capturing the interactions between both the close and distant word pairs. A co-predictor finally reasons the word-word relations and produces all possible entity mentions, in which the biaffine and the multi-layer perceptron (MLP) classifiers are jointly employed for the complementary benefits.

We conduct extensive experiments on 14 datasets, ranging from 2 English and 4 Chinese datasets for flat NER, 3 English and 2 Chinese datasets for overlapped NER, 3 English datasets for discontinuous NER. Compared with 12 baselines for flat NER, 7 baselines for overlapped NER, 7 baselines for discontinuous NER, our model achieves the best performances on all the datasets, becoming the new SoTA method of unified NER. Our contributions include:

- We present an innovative method that casts unified NER as word-word relation classification, where both the relations between boundary-words and inside-words of entities are fully considered.
- We develop a neural framework for unified NER, in which we newly propose a multi-granularity 2D convolution method for sufficiently capturing the interactions between close and distant words.
- Our model pushes current SoTA performances of NER on total 14 datasets. Our code is available at https://github.com/ljynlp/W2NER.

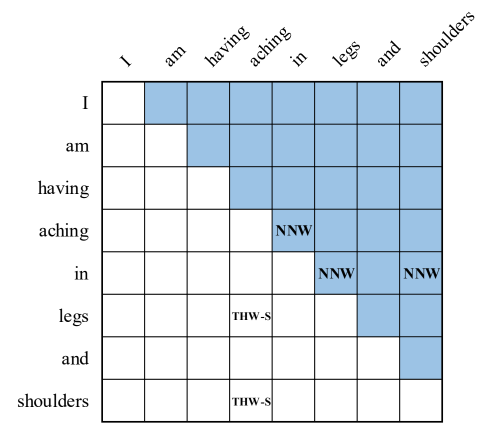

Figure 2: An example to show our relation classification method for NER. We leverage a word-pair grid to visualize the relations between each word pair. NNW denotes the `Next-Neighboring-Word` relation and THW-S denotes the `Tail-Head-Word` relation that exists in a "Symptom" entity. To avoid the sparsity of relation instances, NNW and THW relations are tagged in the upper and lower triangular regions.

NER as Word-Word Relation Classification

Flat, overlapped, discontinuous NER can be formalized as follows: given an input sentence consisting of N tokens or words $X = \{x_1, x_2, ..., x_N\}$, the task aims to extract the relations $\mathcal{R}$ between each token pairs (x_i, x_j), where $\mathcal{R}$ is predefined, including `NONE`, `Next-Neighboring-Word` (NNW), and `Tail-Head-Word-`$\star$ (THW-$\star$). These relations can be explained as below and we also give an example as demonstrated in Figure 2 for better understanding.

- `NONE`, indicating that the word pair does not have any relation defined in this paper.
- `Next-Neighboring-Word`: the NNW relation indicates that the word pair belongs to an entity mention, and the word in certain row of the grid has a successive word in certain column of the grid.
- `Tail-Head-Word-`$\star$: the THW relation indicates that the word in certain row of the grid is the tail of an entity mention, and the word in certain column of the grid is the head of an entity mention. "$\star$" indicates the entity type.

With such design, our framework is able to identify flat, overlapped and discontinuous entities simultaneously. As shown in Figure 2, it is effortless to decode out two

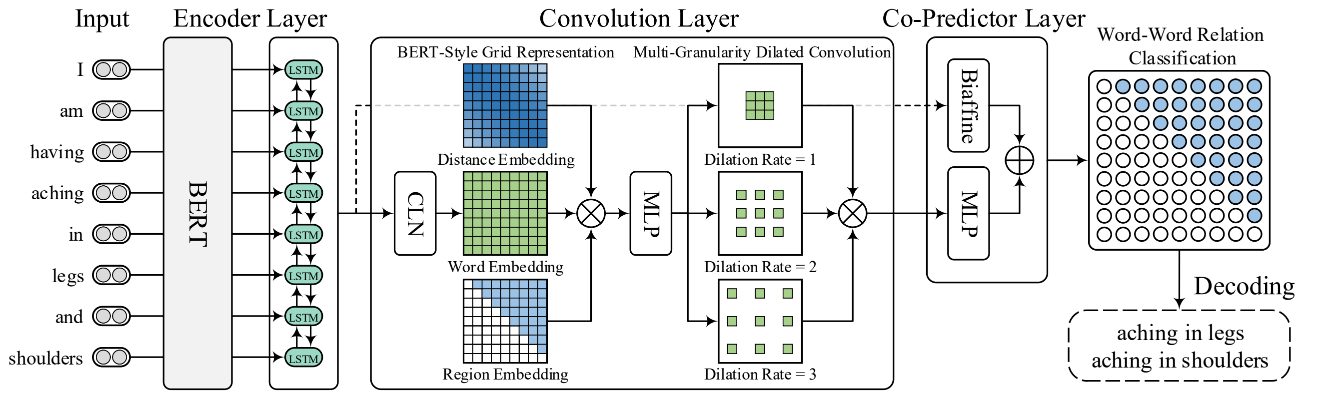

Figure 3: Overall NER architecture. CLN and MLP represent conditional layer normalization and multi-layer perceptron. $\oplus$ and $\otimes$ represent element-wise addition and concatenation operations.

entities "*aching in legs*" and "*aching in shoulders*" by NNW relations (aching→in), (in→legs), and (in→ shoulders), and THW relations (legs→aching, Symptom) and (shoulders→aching, Symptom). Moreover, NNW and THW relations imply other effects for NER. For example, NNW relations associate the segments of the same discontinuous entity (e.g., "aching in" and "shoulders"), and they are also beneficial for identifying entity words (neighbouring) and non-entity words (non-neighbouring). THW relations help identify the boundaries of entities, which plays an important role reported in recent NER studies (Zheng et al. 2019; Fei et al. 2021; Shen et al. 2021).

Unified NER Framework

The architecture of our framework is illustrated in Figure 3, which mainly consists of three components. First, the widely-used pretrained language model, BERT (Devlin et al. 2019), and bi-directional LSTM (Lample et al. 2016) are used as the encoder to yield contextualized word representations from input sentences. Then a convolution layer is used to build and refine the representation of the word-pair grid for later word-word relation classification. Afterward, a co-predictor layer (Li et al. 2021b) that contains a biaffine classifier and a multi-layer perceptron is leveraged for jointly reasoning the relations between all word pairs.

Encoder Layer

We leverage BERT (Devlin et al. 2019) as inputs for our model since it has been demonstrated to be one of the state-of-the-art models for representation learning in NER (Wang et al. 2021) and relation classification (Li et al. 2021b). Given an input sentence $X = \{x_1, x_2, ..., x_N\}$, we convert each token or word x_i into word pieces and then feed them into a pretrained BERT module. After the BERT calculation, each sentential word may involve vectorial representations of several pieces. Here we employ max pooling to produce word representations based on the word piece representations. To further enhance context modeling, we follow prior work (Wadden et al. 2019; Li et al. 2021a), adopting a bi-directional LSTM (Lample et al. 2016) to generate final word representations, i.e., $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N\} \in \mathbb{R}^{N \times d_h}$, where d_h denotes the dimension of a word representation.

Convolution Layer

We adopt convolution neural networks (CNNs) as the representation refiner, since CNNs are naturally suitable for 2-D convolution on the grid, and also show the very prominence on handling relation determination jobs (Zeng et al. 2014; Wang et al. 2016). Our convolution layer includes three modules, including a condition layer with normalization (Liu et al. 2021) for generating the representation of the word-pair grid, a BERT-style grid representation build-up to enrich the representation of the word-pair grid, and a multi-granularity dilated convolution for capturing the interactions between close and distant words.

Conditional Layer Normalization Since the goal of our framework is to predict the relations between word pairs, it is important to generate a high-quality representation of the word-pair grid, which can be regarded as a 3-dimensional matrix, $\mathbf{V} \in \mathbb{R}^{N \times N \times d_h}$, where $\mathbf{V}_{ij}$ denotes the representation of the word pair (x_i, x_j). Because both NNW and THW relations are directional, i.e., from a word x_i in certain row to a word x_j in certain column as shown in Figure 2 (e.g., aching→in and legs→aching), the representation $\mathbf{V}_{ij}$ of the word pair (x_i, x_j) can be considered as a combination of the representation $\mathbf{h}_i$ of x_i and $\mathbf{h}_j$ of x_j, where the combination should imply that x_j is conditioned on x_i. Inspired by Liu et al. (2021), we adopt the Conditional Layer Normalization (CLN) mechanism to calculate $\mathbf{V}_{ij}$:

$$\mathbf{V}_{ij} = \mathrm{CLN}(\mathbf{h}_i, \mathbf{h}_j) = \gamma_{ij} \odot (\frac{\mathbf{h}_j - \mu}{\sigma}) + \lambda_{ij}\,, \quad (1)$$

where $\mathbf{h}_i$ is the condition to generate the gain parameter $\gamma_{ij} = \mathbf{W}_\alpha \mathbf{h}_i + \mathbf{b}_\alpha$ and bias $\lambda_{ij} = \mathbf{W}_\beta \mathbf{h}_i + \mathbf{b}_\beta$ of layer normalization. μ and σ are the mean and standard deviation across the elements of $\mathbf{h}_j$, denoted as:

$$\mu = \frac{1}{d_h}\sum_{k=1}^{d_h} h_{jk}, \quad \sigma = \sqrt{\frac{1}{d_h}\sum_{k=1}^{d_h}(h_{jk} - \mu)^2}\,. \quad (2)$$

where h_{jk} denotes the k-th dimension of $\mathbf{h}_j$.

BERT-Style Grid Representation Build-Up As everyone knows, the inputs of BERT (Devlin et al. 2019) consist of three parts, namely token embeddings, position embeddings and segment embeddings, which model word, position and sentential information respectively. Motivated by BERT, we enrich the representation of the word-pair grid using a similar idea, where the tensor $\mathbf{V} \in \mathbb{R}^{N\times N\times d_h}$ represents word information, a tensor $\mathbf{E}^d \in \mathbb{R}^{N\times N\times d_{E_d}}$ represents the relative position information between each pair of words, and a tensor $\mathbf{E}^t \in \mathbb{R}^{N\times N\times d_{E_t}}$ represents the region information for distinguishing lower and upper triangle regions in the grid. We then concatenate three kinds of embeddings and adopt a multi-layer perceptron (MLP) to reduce their dimensions and mix these information to get the position-region-aware representation of the grid $\mathbf{C} \in \mathbb{R}^{N\times N\times d_c}$. The overall process can be formulated as:

$$\mathbf{C} = \mathrm{MLP}_1([\mathbf{V}; \mathbf{E}^d; \mathbf{E}^t]) \,. \tag{3}$$

Multi-Granularity Dilated Convolution Motivated by TextCNN (Kim 2014), we adopt multiple 2-dimensional dilated convolutions (DConv) with different dilation rates l (e.g., $l \in [1, 2, 3]$) to capture the interactions between the words with different distances, because our model is to predict the relations between these words. The calculation in one dilated convolution can be formulated as:

$$\mathbf{Q}^l = \sigma(\mathrm{DConv}_l(\mathbf{C})) \,, \tag{4}$$

where $\mathbf{Q}^l \in \mathbb{R}^{N\times N\times d_c}$ denotes the output of the dilation convolution with the dilation rate l, σ is the GELU activation function (Hendrycks and Gimpel 2016). After that, we can obtain the final word-pair grid representation $\mathbf{Q} = [\mathbf{Q}^1, \mathbf{Q}^2, \mathbf{Q}^3] \in \mathbb{R}^{N\times N\times 3d_c}$.

Co-Predictor Layer

After the convolution layer, we obtain the word-pair grid representations $\mathbf{Q}$, which are used to predict the relation between each pair of words using an MLP. However, prior work (Li et al. 2021b) has shown that MLP predictor can be enhanced by collaborating with a biaffine predictor for relation classification. We thus take these two predictors concurrently to calculate two separate relation distributions of word pair (x_i, x_j), and combine them as the final prediction.

Biaffine Predictor The input of the biaffine predictor is the output $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N\} \in \mathbb{R}^{N\times d_h}$ of the encoder layer, which can be considered as a residual connection (He et al. 2016) that is widely-used in current deep learning research. Given the word representations $\mathbf{H}$, we use two MLPs to calculate the subject (x_i) and object (x_j) word representations, $\mathbf{s}_i$ and $\mathbf{o}_j$ respectively. Then, a biaffine classifier (Dozat and Manning 2017) is used to compute the relation scores between a pair of subject and object words (x_i, x_j):

$$\mathbf{s}_i = \mathrm{MLP}_2(\mathbf{h}_i) \,, \tag{5}$$

$$\mathbf{o}_j = \mathrm{MLP}_3(\mathbf{h}_j) \,, \tag{6}$$

$$\mathbf{y}'_{ij} = {\mathbf{s}_i}^\top \mathbf{U} \mathbf{o}_j + \mathbf{W}[\mathbf{s}_i; \mathbf{o}_j] + \mathbf{b} \,, \tag{7}$$

where $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters, $\mathbf{s}_i$ and $\mathbf{o}_j$ denote the subject and object representations of the i-th and j-th word, respectively. Here $\mathbf{y}'_{ij} \in \mathbb{R}^{|\mathcal{R}|}$ is the scores of the relations pre-defined in $\mathcal{R}$.

MLP Predictor Based on the word-pair grid representation $\mathbf{Q}$, we adopt an MLP to calculate relations scores for word pairs (x_i, x_j) using $\mathbf{Q}_{ij}$:

$$\mathbf{y}''_{ij} = \mathrm{MLP}(\mathbf{Q}_{ij}) \,, \tag{8}$$

where $\mathbf{y}''_{ij} \in \mathbb{R}^{|\mathcal{R}|}$ is the scores of the relations pre-defined in $\mathcal{R}$. The final relation probabilities $\mathbf{y}_{ij}$ for the word pair (x_i, x_j) are calculated by combining the scores from the biaffine and MLP predictors:

$$\mathbf{y}_{ij} = \mathrm{Softmax}(\mathbf{y}'_{ij} + \mathbf{y}''_{ij}) \,. \tag{9}$$

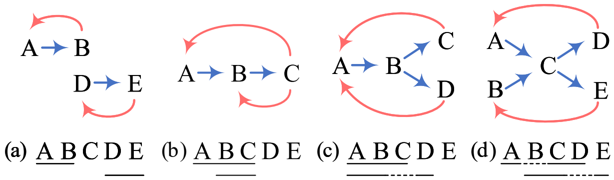

Figure 4: Four decoding cases for the word sequence "ABCDE". (a) "AB" and "DE" are flat entities. (b) The flat entity "BC" is nested in "ABC". (c) The entity "ABC" is overlapped with a discontinuous entity "ABD". (d) Two discontinuous entities "ACD" and "BCE" are overlapped. The blue and red arrows indicate NNW and THW relations.

Decoding

The predictions of our model are the words and their relations, which can be considered as a directional word graph. The decoding object is to find certain paths from one word to anther word in the graph using NNW relations. Each path corresponds to an entity mention. Besides the type and boundary identification for NER, THW relations can also be used as auxiliary information for disambiguation. Figure 4 illustrates four cases for decoding from easy to difficult.

- In the example (a), two paths "A→B" and "D→E" correspond to flat entities, and THW relations indicate their boundaries and types.
- In the example (b), if there is no THW relation, we can only find one path and thus "BC" is missing. In contrast, with the help of THW relations, it is easy to identify that "BC" is nested in "ABC", which demonstrates the necessity of THW relations.
- The case (c) shows how to identify discontinuous entities. Two paths "A→B→C" and "A→B→D" can be found, and the NNW relation contributes to connecting the discontinuous spans "AB" and "D".
- Considering a complex and rare case (d), it is impossible to decode correct entities "ACD" and "BCE" because we can find 4 paths in this ambiguous case using only NNW relations. In contrast, only using THW relations will recognize continuous entities (e.g., "ABCD") rather than correct discontinuous entities (e.g., "ACD"). Therefore, we can obtain correct answers by collaboratively using both relations.

Learning

For each sentence $X = \{x_1, x_2, ..., x_N\}$, our training target is to minimize the negative log-likelihood losses with

		CoNLL2003			OntoNotes 5.0		
		P	R	F1	P	R	F1
• **Sequence Labeling**	Lample et al. (2016)	-	-	90.94	-	-	-
	Strubell et al. (2017)	-	-	90.65	-	-	86.84
• **Span-based**	Yu et al. (2020) †	**92.91**	92.13	92.52	90.01	89.77	89.89
	Shen et al. (2021)	92.13	**93.73**	92.94	-	-	-
• **Hypergraph-based**	Wang and Lu (2018)	-	-	90.50	-	-	-
• **Seq2Seq**	Straková et al. (2019)	-	-	92.98	-	-	-
	Yan et al. (2021) †	92.56	93.56	93.05	89.62	90.92	90.27
	W^{2}NER (ours)	92.71	93.44	**93.07**	**90.03**	**90.97**	**90.50**

Table 1: Results for English flat NER datasets. "†" denotes our re-implementation via their code. We run our model for 5 times and report averaged values.[3]

	OntoNotes 4.0			MSRA			Resume			Weibo		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Zhang and Yang (2018)	76.35	71.56	73.88	93.57	92.79	93.18	94.81	94.11	94.46	53.04	62.25	58.79
Yan et al. (2019)	-	-	72.43	-	-	92.74	-	-	95.00	-	-	58.17
Gui et al. (2019)	76.40	72.60	74.45	94.50	92.93	93.71	95.37	94.84	95.11	57.14	66.67	59.92
Li et al. (2020b)	-	-	81.82	-	-	96.09	-	-	95.86	-	-	68.55
Ma et al. (2020)	**83.41**	82.21	82.81	95.75	95.10	95.42	96.08	96.13	96.11	**70.94**	67.02	70.50
W^{2}NER (ours)	82.31	**83.36**	**83.08**	**96.12**	**96.08**	**96.10**	**96.96**	**96.35**	**96.65**	70.84	**73.87**	**72.32**

Table 2: Results for Chinese flat NER datasets. All the baselines are sequence labeling methods or their variations.

		ACE2004			ACE2005			GENIA		
		P	R	F1	P	R	F1	P	R	F1
• **Sequence Labeling**	Ju et al. (2018)	-	-	-	74.20	70.30	72.20	78.50	71.30	74.70
• **Span-based**	Wang et al. (2020)	86.08	86.48	86.28	83.95	85.39	84.66	79.45	78.94	79.19
	Yu et al. (2020)	87.30	86.00	86.70	85.20	85.60	85.40	81.80	79.30	80.50
	Shen et al. (2021)	**87.44**	87.38	87.41	**86.09**	87.27	86.67	80.19	**80.89**	80.54
• **Hypergraph-based**	Wang and Lu (2018)	78.00	72.40	75.10	76.80	72.30	74.50	77.00	73.30	75.10
• **Seq2Seq**	Straková et al. (2019)	-	-	84.33	-	-	83.42	-	-	78.20
	Yan et al. (2021)	87.27	86.41	86.84	83.16	86.38	84.74	78.87	79.60	79.23
	W^{2}NER (ours)	87.33	**87.71**	**87.52**	85.03	**88.62**	**86.79**	**83.10**	79.76	**81.39**

Table 3: Results for English overlapped NER datasets.

regards to the corresponding gold labels, formalized as:

$$\mathcal{L} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{|\mathcal{R}|}\hat{\mathbf{y}}_{ij}^{r}\log\mathbf{y}_{ij}^{r}, \tag{10}$$

where N it the number of words in the sentence, $\hat{\mathbf{y}}_{ij}$ is the binary vector that denotes the gold relation labels for the word pair (x_i, x_j), and $\mathbf{y}_{ij}$ are the predicted probability vector. r indicates the r-th relation of the pre-defined relation set $\mathcal{R}$.

Experimental Settings

Datasets

To evaluate our framework for three NER subtasks, we conducted experiments on 14 datasets.

Flat NER Datasets We adopt CoNLL-2003 (Sang and Meulder 2003) and OntoNotes 5.0 (Pradhan et al. 2013b) in English, OntoNotes 4.0 (Weischedel et al. 2011), MSRA (Levow 2006), Weibo (Peng and Dredze 2015; He and Sun 2017), and Resume (Zhang and Yang 2018) in Chinese. We employ the same experimental settings in previous work (Lample et al. 2016; Yan et al. 2021; Ma et al. 2020; Li et al. 2020b).

Overlapped NER Datasets We conduct experiments on ACE 2004 (Doddington et al. 2004), ACE 2005 (Walker et al. 2011), GENIA (Kim et al. 2003). For GENIA, we follow Yan et al. (2021) to use five types of entities and split the train/dev/test as 8.1:0.9:1.0. For ACE 2004 and ACE 2005 in English, we use the same data split as Lu and Roth (2015); Yu et al. (2020). For ACE 2004 and ACE 2005 in Chinese, we split the train/dev/test as 8.0:1.0:1.0.

Discontinuous NER Datasets We experiment on three datasets for discontinuous NER, namely CADEC (Karimi et al. 2015), ShARe13 (Pradhan et al. 2013a) and ShARe14 (Mowery et al. 2014), all of which are derived from biomedical or clinical domain documents. We use the preprocessing scripts provided by Dai et al. (2020) for data splitting.

[3]The results in Table 2-6 are also the averaged values.

		CADEC			ShARe13			ShARe14		
		P	R	F1	P	R	F1	P	R	F1
• Sequence Labeling	Tang et al. (2018)	67.80	64.99	66.36	-	-	-	-	-	-
• Span-based	Li et al. (2021a)	-	-	69.90	-	-	82.50	-	-	-
• Hypergraph-based	Wang and Lu (2019)	72.10	48.40	58.00	83.80	60.40	70.30	79.10	70.70	74.70
• Seq2Seq	Yan et al. (2021)	70.08	71.21	70.64	82.09	77.42	79.69	77.20	83.75	80.34
	Fei et al. (2021)	**75.50**	71.80	72.40	**87.90**	77.20	80.30	-	-	-
• Others	Dai et al. (2020)	68.90	69.00	69.00	80.50	75.00	77.70	78.10	81.20	79.60
	Wang et al. (2021)	70.50	**72.50**	71.50	84.30	78.20	81.20	78.20	**84.70**	81.30
	W^{2}NER (ours)	74.09	72.35	**73.21**	85.57	**79.68**	**82.52**	**79.88**	83.71	**81.75**

Table 4: Results for discontinuous NER datasets.[4]

	ACE2004	ACE2005
Yu et al. (2020) ⋆	87.35	88.39
Shen et al. (2021) ⋆	87.47	88.21
W^{2}NER (ours)	**88.00**	**88.81**

Table 5: F1s for Chinese overlapped NER datasets. Models with "⋆" are adapted to target datasets using their code.

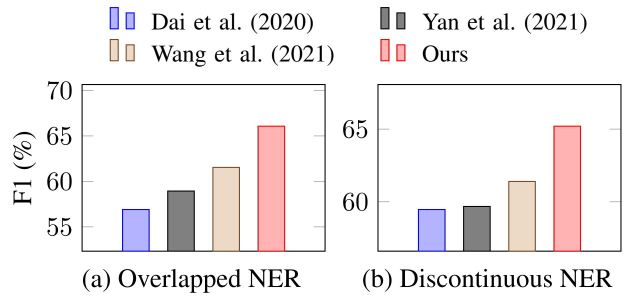

Figure 5: Results of overlapped (a) and discontinuous mentions (b) on ShARe14.

Around 10% of entities in these datasets are discontinuous.

Baselines

Tagging-based methods, which assign a tag to every token with different label schemes, such as BIO (Lample et al. 2016), BIOHD (Tang et al. 2018), and BIEOS (Li et al. 2020b; Ma et al. 2020). **Span-based methods**, which enumerate all possible spans and combine them into entities (Yu et al. 2020; Li et al. 2021a). **Hypergraph-based approaches**, which utilize hypergraphs to represent and infer entity mentions (Lu and Roth 2015; Wang and Lu 2018; Katiyar and Cardie 2018). **Seq2Seq methods**, which generate entity label sequences (Strubell et al. 2017), index or word sequences (Yan et al. 2021; Fei et al. 2021) at the decoder side. **Other methods**, which is different from the methods above, such as transition-based (Dai et al. 2020) and clique-based (Wang et al. 2021) approaches.

Experimental Results

Results for Flat NER

We evaluate our framework on six datasets. As shown in Table 1, Our model achieves the best performances with 93.07% F1 and 90.50% F1 on CoNLL 2003 and OntoNotes 5.0 datasets. Especially, our model outperforms another unified NER framework Yan et al. (2021) by 0.23% in terms of F1 on OntoNotes 5.0. The results in Chinese datasets are shown in Table 2, where baselines are all tagging-based methods. We find that our model outperforms the previous SoTA results by 0.27%, 0.01%, 0.54% and 1.82% on OntoNotes 4.0, MSRA, Resume and Weibo.

Results for Overlapped NER

Table 3 presents the results for three overlapped NER datasets in English. Our W^{2}NER model outperforms the previous works, including tagging-based (Ju et al. 2018), span-based (Wang et al. 2020; Yu et al. 2020; Shen et al. 2021), hypergraph-based (Wang and Lu 2018) and sequence-to-sequence (Straková et al. 2019; Yan et al. 2021) approaches, and achieves the SoTA performances on F1 scores, with 87.52%, 86.79% and 81.39% on ACE2004, ACE2005 and GENIA, respectively. For ACE2004 and ACE2005 corpora in Chinese, we reproduce the SoTA models proposed by Yu et al. (2020) and Shen et al. (2021), and list their results in Table 5. Our model can significantly outperform the two baselines by 0.53% and 0.42%.

[4]Note that discontinuous NER datasets include both flat and overlapped entities as well.

Results for Discontinuous NER

Table 4 presents the comparisons between our model and other baselines in three discontinuous NER datasets. As seen, our model outperforms previous best model (Fei et al. 2021; Wang et al. 2021) by 0.81%, 0.02%, and 0.45% in F1s in the CADEC, ShARe13 and ShARe14 datasets, respectively, leading to new SoTA results.

Since the above datasets also include flat entities, we further investigate the performances of our model on recognizing only overlapped or discontinuous entities, as shown in Figure 5. We can learn that the clique-based model (Wang et al. 2021) shows better performances than the Seq2Seq model (Yan et al. 2021) and transition-based method (Dai et al. 2020). Most importantly, our system achieves the best results against all other baselines for both overlapped and discontinuous NER.

	CoNLL2003	ACE2005	CADEC
Ours	**93.07**	**86.79**	**73.21**
- Region Emb.	92.80 (-0.27)	86.39 (-0.40)	72.56 (-0.65)
- Distance Emb.	92.89 (-0.18)	86.47 (-0.32)	72.66 (-0.55)
- All DConv	92.31 (-0.76)	86.07 (-0.72)	72.45 (-0.76)
- DConv(l=1)	93.05 (-0.02)	86.64 (-0.15)	73.12 (-0.09)
- DConv(l=2)	92.78 (-0.29)	86.58 (-0.21)	72.95 (-0.26)
- DConv(l=3)	92.82 (-0.25)	86.59 (-0.20)	73.10 (-0.11)
- Biaffine	93.02 (-0.05)	86.30 (-0.49)	72.71 (-0.50)
- MLP	91.87 (-1.20)	85.66 (-1.13)	68.04 (-5.17)
- NNW	92.65 (-0.42)	86.23 (-0.56)	69.01 (-4.20)

Table 6: Model ablation studies (F1s). DConv(l=1) denots the convolution with the dilation rate 1.

Model Ablation Studies

We ablate each part of our model on the CoNLL2003, ACE2005 and CADEC datasets, as shown in Table 6. First, without region and distance embeddings, we observe slight performance drops on the three datasets. By removing all convolutions, the performance also drops obviously, which verifies the usefulness of the multi-granularity dilated convolution. Furthermore, after removing convolutions with different dilation rate, the performance also decreases, especially for the convolution with the dilation rate 2.

Comparing the biaffine and MLP in the co-predictor layer, we find that although the MLP plays a leading role, the biaffine also brings about 0.5% gains at most. At last, when the NNW relation is removed, the F1s on all datasets drop, especially on the CADEC (4.2%). This is because the CADEC dataset also contains discontinuous entities and without the NNW relation, discontinuous spans will be incorrectly recognized as continuous ones, as shown in Figure 4(d). Therefore, the results of ablation studies on the NNW relation demonstrate its importance as we argued before.

Related Work on NER

Sequence Labeling Approaches NER is usually considered as a sequence labeling problem, to assign each token a tag from a pre-designed tagging scheme (e.g., *BIO*). Current mainstream work combine the CRF (Lafferty et al. 2001; Finkel et al. 2005) with neural architecture, such as CNN (Collobert et al. 2011; Strubell et al. 2017), bi-directional LSTM (Huang et al. 2015; Lample et al. 2016), and Transformer (Yan et al. 2019; Li et al. 2020b). However, these methods fail to directly solve neither overlapped nor discontinuous NER. Ju et al. (2018) propose a neural model for nested NER by dynamically stacking flat NER layers. Tang et al. (2018) extend the BIO label scheme to BIOHD to address the problem of discontinuous mention.

Span-based Approaches There have been several studies that cast NER as span-level classification, i.e., enumerating all possible spans, and determining if they are valid mentions and the types (Xu et al. 2017; Luan et al. 2019; Yamada et al. 2020). Yu et al. (2020) utilize biaffine attention (Dozat and Manning 2017) to measure the possibility as a mention of a text span. Li et al. (2020a) reformulate NER as a machine reading comprehension (MRC) task and extract entities as the answer spans. Shen et al. (2021) implement a two-stage identifier to generate span proposals through a filter and a regressor, and then classify them into the corresponding categories. Li et al. (2021a) convert the discontinuous NER to find complete subgraphs from a span-based entity fragment graph, and achieve competitive results. But, due to the exhaustively enumerating nature, those methods suffer from maximal span lengths and considerable model complexity, especially for long-span entities.

Hypergraph-based Approaches Lu and Roth (2015) first propose the hypergraph model for overlapped NER, by exponentially representing possible mentions. The method is then widely explored by follow-up work (Muis and Lu 2016; Katiyar and Cardie 2018; Wang and Lu 2018). For instance, Muis and Lu (2016) extend the method for discontinuous NER, and Wang and Lu (2018) utilize deep neural networks to enhance the hypergraph model.

Sequence-to-Sequence Approaches Gillick et al. (2016) first apply the Seq2Seq model for NER, taking as inputs the sentence, and outputting all the entity start positions, span lengths and labels. Straková et al. (2019) use the Seq2Seq architecture for overlapped NER with enhanced BILOU scheme. Fei et al. (2021) employ Seq2Seq with pointer network for discontinuous NER. The latest attempt in (Yan et al. 2021) tackles the unified NER via a Seq2Seq model with pointer network based-on BART (Lewis et al. 2020), generating a sequence of all possible entity start-end indexes and types. Seq2Seq architecture unfortunately suffers from the potential decoding efficiency problem as well as the exposure bias issue.

Differences between Our Approach and Previous Approaches Most of the existing NER work mainly consider more accurate entity boundary identification. In this work, we explore a different task modeling for unified NER, i.e., a formalism as word-word relation classification. Our method can effectively model the relations between both the boundary-words and inside-words of entities. Also, our method with 2D grid-tagging can substantially avoid the drawbacks in current best-performing baselines, e.g., span-based and sequence-to-sequence models.

Conclusion

In this paper, we propose a novel unified NER framework based on word-word relation classification to address unified NER concurrently. The relations between word pairs are predefined as next-neighboring-word relations and tail-head-word relations. We find that our framework is quite effective for various NER, which achieves SoTA performances for 14 widely-used benchmark datasets. Moreover, we propose a novel backbone model that consists of a BERT-BiLSTM encoder layer, a convolution layer for building and refining the representation of the word-pair grid, and a co-predictor layer for jointly reasoning relations. Through ablation studies, we find that our convolution-centric model performs well and several proposed modules such as the co-predictor and grid representation enrichment are also effective. Our framework and model are easy to follow, which will promote the development of NER research.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61772378, No. 62176187), the National Key Research and Development Program of China (No. 2017YFC1200500), the Research Foundation of Ministry of Education of China (No.18JZD015), the Youth Fund for Humanities and Social Science Research of Ministry of Education of China (No. 22YJCZH064). This work is also the research result of the independent scientific research project (humanities and social sciences) of Wuhan University, supported by the Fundamental Research Funds for the Central Universities.

References

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE): 2493–2537.

Dai, X. 2018. Recognizing Complex Entity Mentions: A Review and Future Directions. In *Proceedings of the ACL*, 37–44.

Dai, X.; Karimi, S.; Hachey, B.; and Paris, C. 2020. An Effective Transition-based Model for Discontinuous NER. In *Proceedings of the ACL*, 5860–5870.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the NAACL-HLT*, 4171–4186.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. 2(1): 837–840.

Dozat, T.; and Manning, C. D. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of the ICLR*.

Fei, H.; Li, F.; Li, B.; Liu, Y.; Ren, Y.; and Ji, D. 2021. Rethinking Boundaries: End-To-End Recognition of Discontinuous Mentions with Pointer Networks. In *Proceedings of the AAAI*, 12785–12793.

Fei, H.; Ren, Y.; and Ji, D. 2020. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6): 102311.

Finkel, J. R.; Grenager, T.; and Manning, C. D. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the ACL*, 363–370.

Gillick, D.; Brunk, C.; Vinyals, O.; and Subramanya, A. 2016. Multilingual Language Processing From Bytes. In *Proceedings of the NAACL*, 1296–1306.

Gui, T.; Zou, Y.; Zhang, Q.; Peng, M.; Fu, J.; Wei, Z.; and Huang, X. 2019. A Lexicon-Based Graph Neural Network for Chinese NER. In *Proceedings of the EMNLP-IJCNLP*, 1040–1050.

He, H.; and Sun, X. 2017. F-Score Driven Max Margin Neural Network for Named Entity Recognition in Chinese Social Media. In *Proceedings of the EACL*, 713–718.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the CVPR*, 770–778.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hou, F.; Wang, R.; He, J.; and Zhou, Y. 2020. Improving Entity Linking through Semantic Reinforced Entity Embeddings. In *Proceedings of the ACL*, 6843–6848.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ju, M.; Miwa, M.; and Ananiadou, S. 2018. A Neural Layered Model for Nested Named Entity Recognition. In *Proceedings of the NAACL*, 1446–1459.

Karimi, S.; Metke-Jimenez, A.; Kemp, M.; and Wang, C. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55: 73–81.

Katiyar, A.; and Cardie, C. 2018. Nested Named Entity Recognition Revisited. In *Proceedings of the NAACL*, 861–871.

Kim, J.-D.; Ohta, T.; Tateisi, Y.; and Tsujii, J. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1): i180–i182.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the EMNLP*, 1746–1751.

Lafferty, J.; McCallum, A.; and Pereira, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the NAACL*, 260–270.

Le, P.; and Titov, I. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In *Proceedings of the ACL*, 1595–1604.

Levow, G.-A. 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*, 108–117.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the ACL*, 7871–7880.

Li, F.; Lin, Z.; Zhang, M.; and Ji, D. 2021a. A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition. In *Proceedings of the ACL-IJCNLP*, 4814–4828.

Li, J.; Xu, K.; Li, F.; Fei, H.; Ren, Y.; and Ji, D. 2021b. MRN: A Locally and Globally Mention-Based Reasoning Network for Document-Level Relation Extraction. In *Proceedings of the ACL-IJCNLP findings*, 1359–1370.

Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; and Li, J. 2020a. A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the ACL*, 5849–5859.

Li, X.; Yan, H.; Qiu, X.; and Huang, X. 2020b. FLAT: Chinese NER Using Flat-Lattice Transformer. In *Proceedings of the ACL*, 6836–6842.

Liu, R.; Wei, J.; Jia, C.; and Vosoughi, S. 2021. Modulating Language Models with Emotions. In *Findings of the ACL-IJCNLP*, 4332–4339.

Lu, W.; and Roth, D. 2015. Joint Mention Extraction and Classification with Mention Hypergraphs. In *Proceedings of the EMNLP*, 857–867.

Luan, Y.; Wadden, D.; He, L.; Shah, A.; Ostendorf, M.; and Hajishirzi, H. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the NAACL*, 3036–3046.

Ma, R.; Peng, M.; Zhang, Q.; Wei, Z.; and Huang, X. 2020. Simplify the Usage of Lexicon in Chinese NER. In *Proceedings of the ACL*, 5951–5960.

Mowery, D. L.; Velupillai, S.; South, B. R.; Christensen, L.; Martinez, D.; Kelly, L.; Goeuriot, L.; Elhadad, N.; Pradhan, S.; Savova, G.; et al. 2014. Task 2: ShARe/CLEF eHealth evaluation lab 2014. In *Proceedings of CLEF 2014*.

Muis, A. O.; and Lu, W. 2016. Learning to Recognize Discontiguous Entities. In *Proceedings of the EMNLP*, 75–84.

Peng, N.; and Dredze, M. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the EMNLP*, 548–554.

Pradhan, S.; Elhadad, N.; South, B. R.; Martinez, D.; Christensen, L. M.; Vogel, A.; Suominen, H.; Chapman, W. W.; and Savova, G. K. 2013a. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *CLEF (Working Notes)*, 212–31.

Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013b. Towards robust linguistic analysis using ontonotes. In *Proceedings of the CoNLL*, 143–152.

Sang, E. F. T. K.; and Meulder, F. D. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the NAACL*, 142–147.

Shen, Y.; Ma, X.; Tan, Z.; Zhang, S.; Wang, W.; and Lu, W. 2021. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. In *Proceedings of the ACL-IJCNLP*, 2782–2794.

Straková, J.; Straka, M.; and Hajic, J. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the ACL*, 5326–5331.

Strubell, E.; Verga, P.; Belanger, D.; and McCallum, A. 2017. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *Proceedings of the EMNLP*, 2670–2680.

Tang, B.; Hu, J.; Wang, X.; and Chen, Q. 2018. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using LSTM-CRF. *Wireless Communications and Mobile Computing*, 2018.

Wadden, D.; Wennberg, U.; Luan, Y.; and Hajishirzi, H. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In *Proceedings of the EMNLP*, 5788–5793.

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2011. Ace 2005 multilingual training corpus. *LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium*.

Wang, B.; and Lu, W. 2018. Neural Segmental Hypergraphs for Overlapping Mention Recognition. In *Proceedings of the EMNLP*, 204–214.

Wang, B.; and Lu, W. 2019. Combining Spans into Entities: A Neural Two-Stage Approach for Recognizing Discontiguous Entities. In *Proceedings of the EMNLP-IJCNLP*, 6216–6224.

Wang, J.; Shou, L.; Chen, K.; and Chen, G. 2020. Pyramid: A Layered Model for Nested Named Entity Recognition. In *Proceedings of the ACL*, 5918–5928.

Wang, L.; Cao, Z.; De Melo, G.; and Liu, Z. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the ACL*, 1298–1307.

Wang, Y.; Yu, B.; Zhu, H.; Liu, T.; Yu, N.; and Sun, L. 2021. Discontinuous Named Entity Recognition as Maximal Clique Discovery. In *Proceedings of the ACL-IJCNLP*, 764–774.

Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; and Chang, Y. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the ACL*, 1476–1488.

Weischedel, R.; Pradhan, S.; Ramshaw, L.; Palmer, M.; Xue, N.; Marcus, M.; Taylor, A.; Greenberg, C.; Hovy, E.; Belvin, R.; et al. 2011. OntoNotes Release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Xu, M.; Jiang, H.; and Watcharawittayakul, S. 2017. A Local Detection Approach for Named Entity Recognition and Mention Detection. In *Proceedings of the ACL*, 1237–1247.

Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; and Matsumoto, Y. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the EMNLP*, 6442–6454.

Yan, H.; Deng, B.; Li, X.; and Qiu, X. 2019. TENER: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yan, H.; Gui, T.; Dai, J.; Guo, Q.; Zhang, Z.; and Qiu, X. 2021. A Unified Generative Framework for Various NER Subtasks. In *Proceedings of the ACL-IJCNLP*, 5808–5822.

Yu, J.; Bohnet, B.; and Poesio, M. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the ACL*, 6470–6476.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the COLING*, 2335–2344.

Zeng, X.; Zeng, D.; He, S.; Liu, K.; and Zhao, J. 2018. Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism. In *Proceedings of the ACL*, 506–514.

Zhang, Y.; and Yang, J. 2018. Chinese NER Using Lattice LSTM. In *Proceedings of the ACL*, 1554–1564.

Zheng, C.; Cai, Y.; Xu, J.; Leung, H.-f.; and Xu, G. 2019. A Boundary-aware Neural Model for Nested Named Entity Recognition. In *Proceedings of the EMNLP-IJCNLP*, 357–366.