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,†]

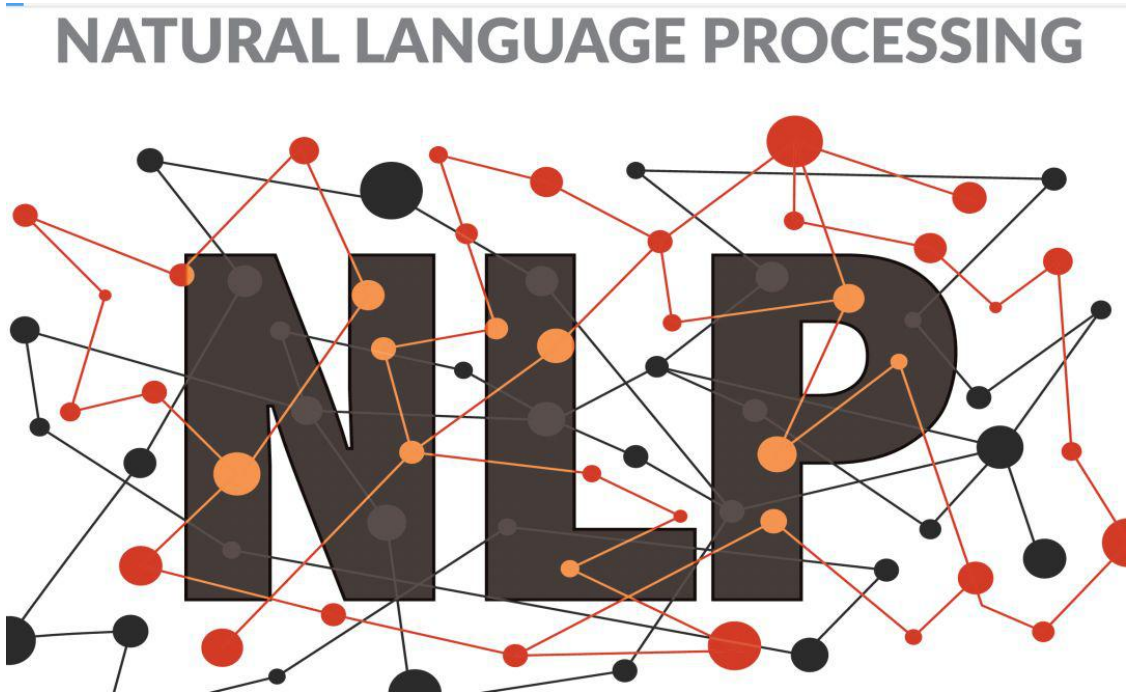
[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China
[2] Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen), China
{theodorelee, hao.fei, liujiang, whuwsq, tengchong, dhji, lifei_csnlp}@whu.edu.cn, mason.zms@gmail.com

2022.04.17 • ChongQing

2022_AAAI

Reported by Yidan Liu

Code:https://github.com/ljynlp/W2NER.git



1.Introduction

2.Method

3.Experiments

Introduction

- Flat
- overlapped (nested)
- discontinuous NER

Current best-performing methods

- Span-based
  merely focus on boundary identification
- Sequence-to-Sequence
  suffer from exposure bias.

In this work, we present a novel alternative by modeling the unified NER as word-word relation classification, namely W²NER.

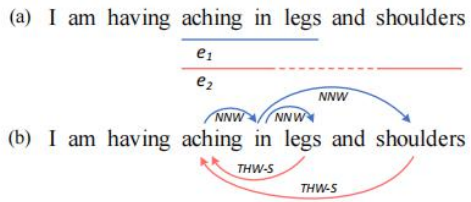


Figure 1: (a) An example to show three types of NER. e_1 is a flat entity overlapped with a discontinuous entity e_2 at the span "aching in". (b) We formalize three NER subtasks as word-word relation classification, where the `Next-Neighboring-Word` (NNW) relation indicates that a word pair are successively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

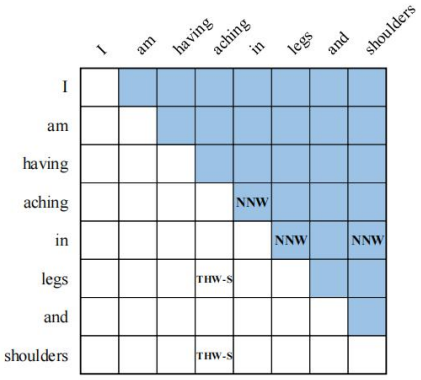


Figure 2: An example to show our relation classification method for NER. We leverage a word-pair grid to visualize the relations between each word pair. NNW denotes the `Next-Neighboring-Word` relation and THW-S denotes the `Tail-Head-Word` relation that exists in a "Symptom" entity. To avoid the sparsity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

Method

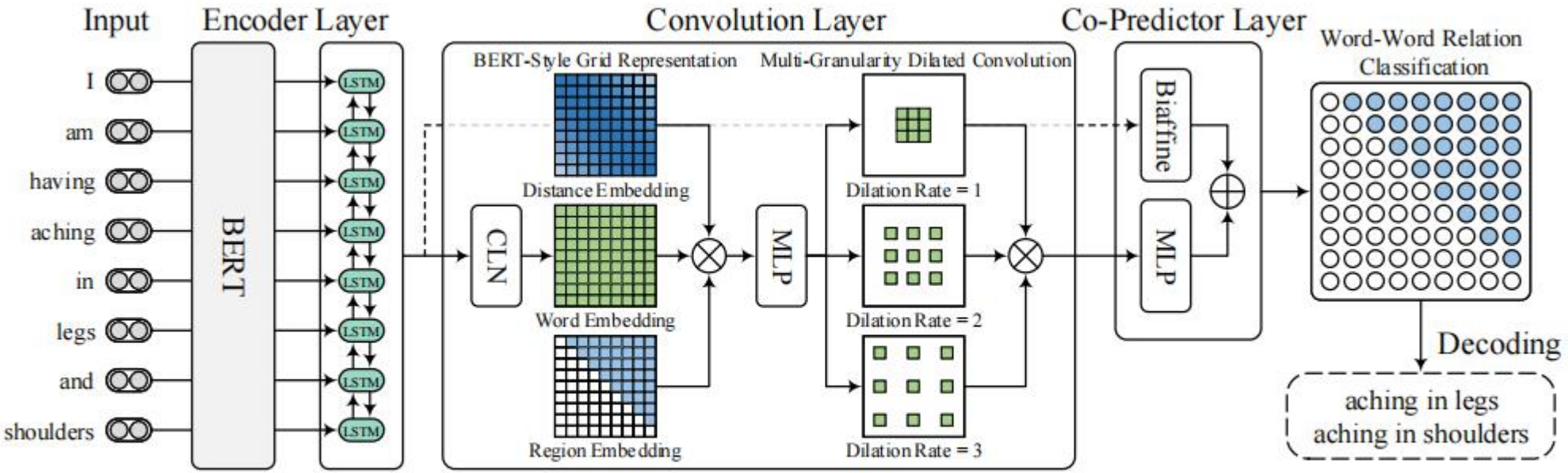
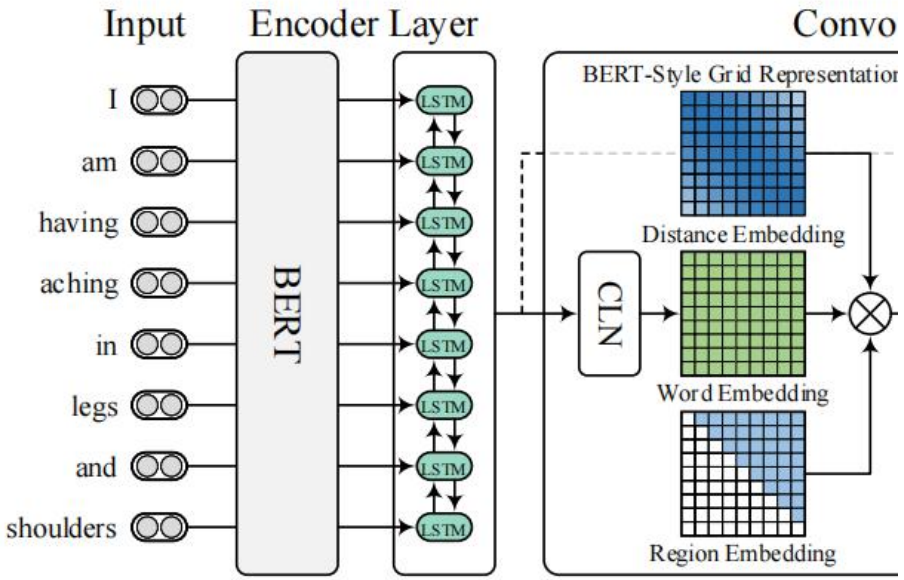


Figure 3: Overall NER architecture. CLN and MLP represent conditional layer normalization and multi-layer perceptron. $\oplus$ and $\otimes$ represent element-wise addition and concatenation operations.

Method

Encoder Layer



Given an input sentence $X = \{x_1, x_2, ..., x_N\}$

$$\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N\} \in \mathbb{R}^{N \times d_h}$$

where d_h denotes the dimension of a word representation.

Conditional Layer Normalization

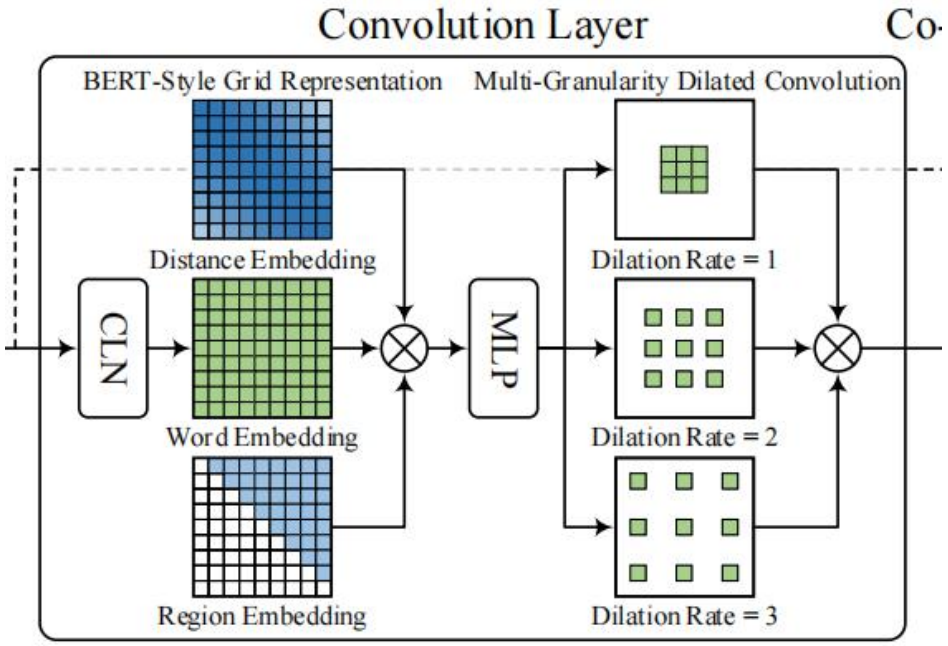
$$\mathbf{V}_{ij} = \mathrm{CLN}(\mathbf{h}_i, \mathbf{h}_j) = \gamma_{ij} \odot \left(\frac{\mathbf{h}_j - \mu}{\sigma}\right) + \lambda_{ij}, \quad (1)$$

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 $\mathbf{h}_j$.

Method



BERT-Style Grid Representation Build-Up

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

$\mathbf{V} \in \mathbb{R}^{N \times N \times d_h}$ $\mathbf{E}^d \in \mathbb{R}^{N \times N \times d_{E_d}}$ $\mathbf{E}^t \in \mathbb{R}^{N \times N \times d_{E_t}}$

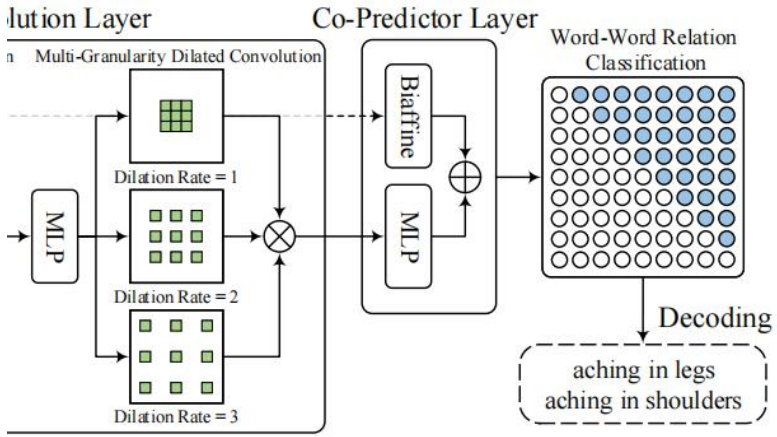
$\mathbf{C} \in \mathbb{R}^{N \times N \times d_c}$

Multi-Granularity Dilated Convolution

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

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 $\mathbf{Q} = [\mathbf{Q}^1, \mathbf{Q}^2, \mathbf{Q}^3] \in \mathbb{R}^{N \times N \times 3d_c}$.

Method



Biaffine Predictor

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

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

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

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

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

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

Decoding

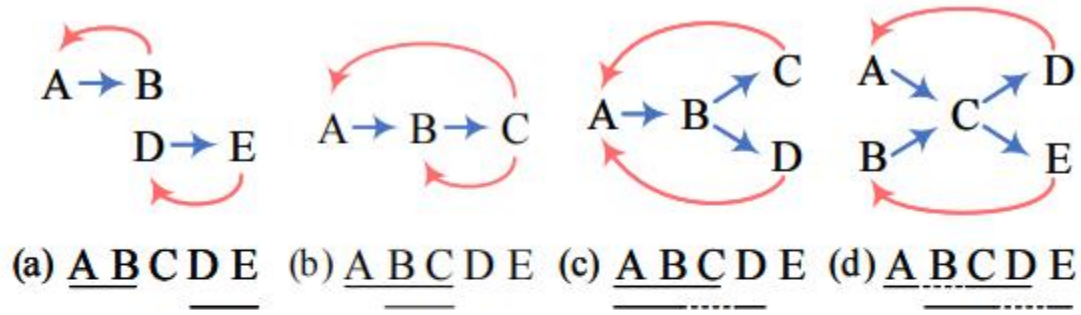


Figure 4: Four decoding cases for the word sequence "ABCDE". (a) "AB" and "DE" are flat entities. (b) The flat entity "BC" is nested in "ABC". (c) The entity "ABC" is overlapped with a discontinuous entity "ABD". (d) Two discontinuous entities "ACD" and "BCE" are overlapped. The blue and red arrows indicate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

Experiments

		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]

Experiments

	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.

Experiments

		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.

Experiments

		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]

Experiments

	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.

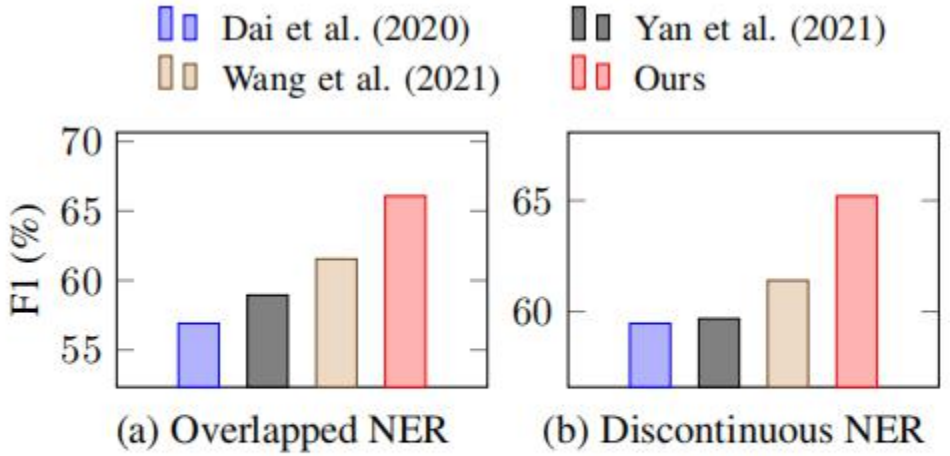


Figure 5: Results of overlapped (a) and discontinuous mentions (b) on ShARe14.

Experiments

	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.

Thank you!