

CARE: Co-Attention Network for Joint Entity and Relation Extraction

Wenjun Kong and Yamei Xia

Task Definition

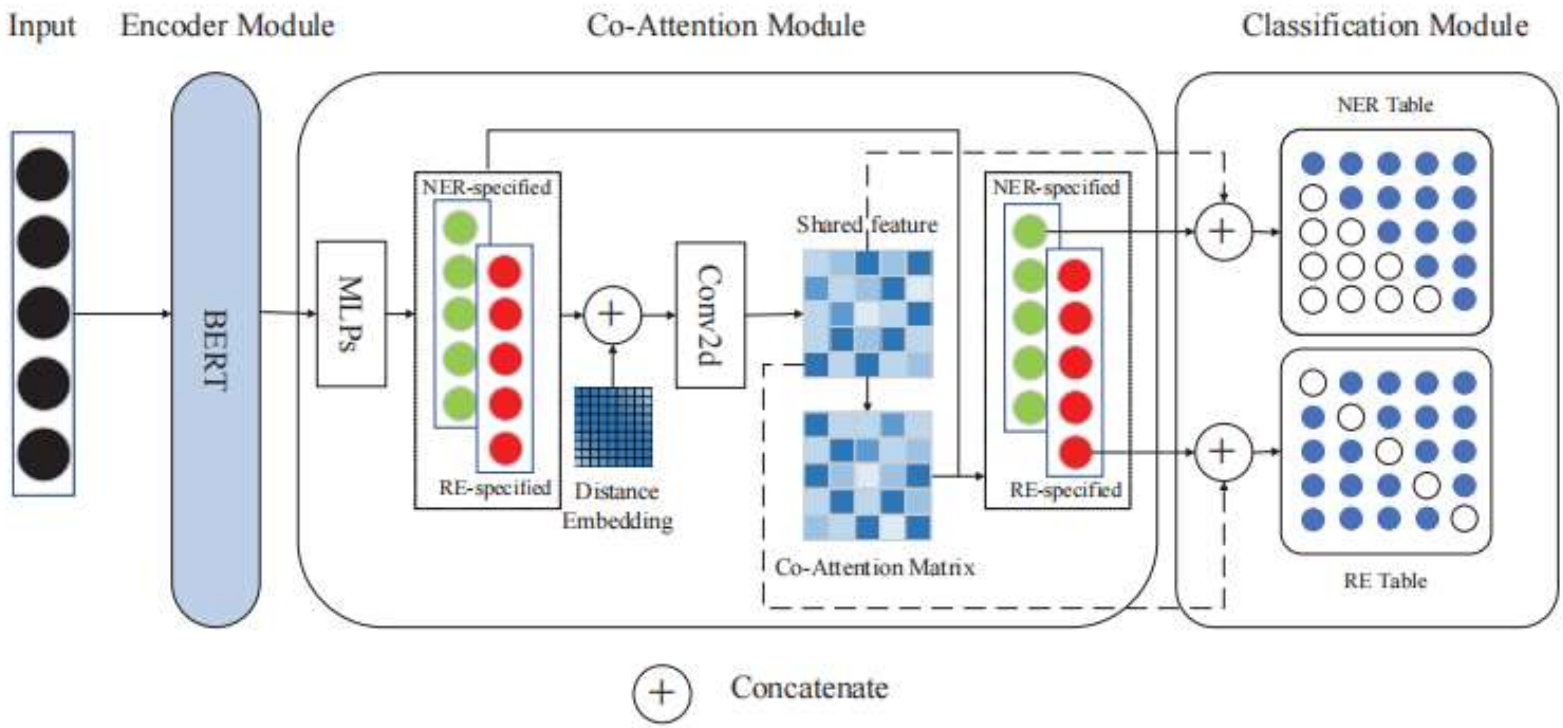
- NER: identify and classify named entities within a text
- RE: identify and classify relations between entities mentioned in text
- Joint NER and RE: combine both subtasks in a joint model

Text Furthermore, we propose the use of standard parser evaluation methods for automatically evaluating the summarization quality of sentence condensation systems .		
Entity		Entity Type
parser evaluation methods		Method
summarization quality		Metric
sentence condensation systems		Method
Head Entity	Relation	Tail Entity
parser evaluation methods	Evaluate For	summarization quality
summarization quality	Evaluate For	sentence condensation systems

Motivation

- Most joint models suffer from the issue of feature confusion, failing to effectively model the interaction between the two subtasks.
- Challenges
 - How to alleviate feature confusion within the joint model when learning features across the two subtasks?
 - How to effectively model the intricate interaction between the two subtasks within a joint framework?

Model



Model

- Parallel encoding

$$H = BERT(X)$$

$$H_{NER} = MLP(H)$$

$$H_{RE} = MLP(H)$$

- Shared presentation and Convolution layers

$$Q = [H_{NER}; H_{RE}; D]$$

$$H_{share} = Conv2d(Q)$$

- Attention aggregation

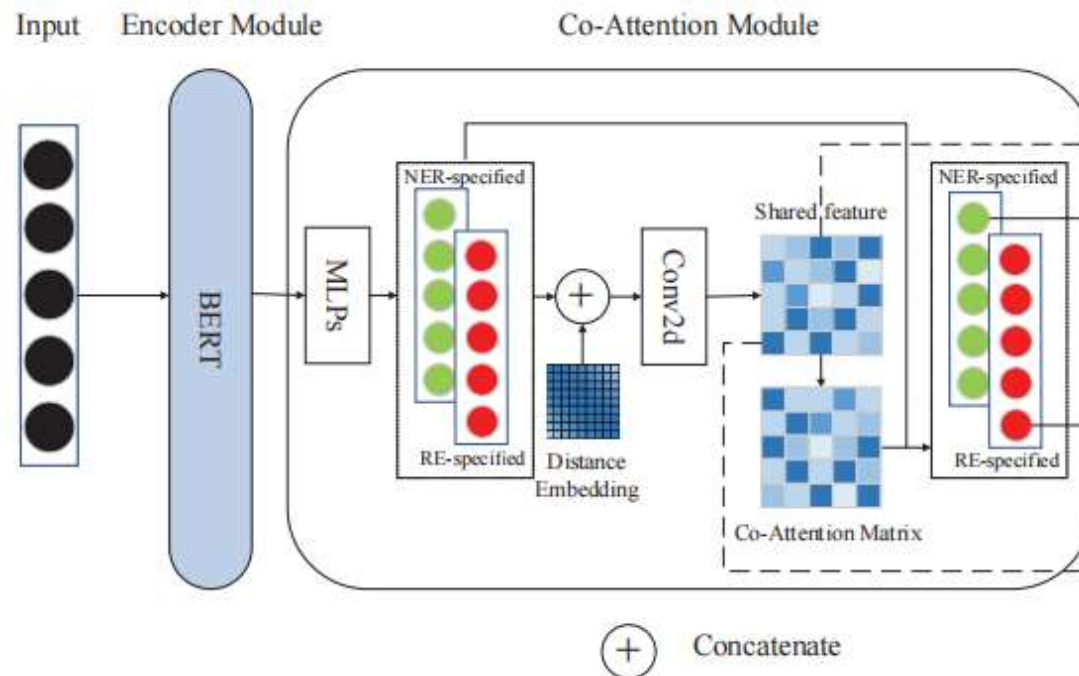
$$A = FFNN(H_{share})$$

$$\alpha = Softmax(A)$$

$$\beta = Softmax(A^T)$$

$$g_i^e = h_i^e + \sum_{j=1}^n \alpha_{ij} h_j^r$$

$$g_i^r = h_i^r + \sum_{j=1}^n \beta_{ij} h_j^e$$



Model

- Representations Concatenation

$$u_{ij}^e = [g_i^e; g_j^e; h_{ij}^s]$$

$$u_{ij}^r = [g_i^r; g_j^r; h_{ij}^s]$$

- Conditional probability prediction

$$P(e|x_i, x_j) = \text{sigmoid}(W_b \sigma(W_a u_{ij}^e + b_a) + b_b)$$

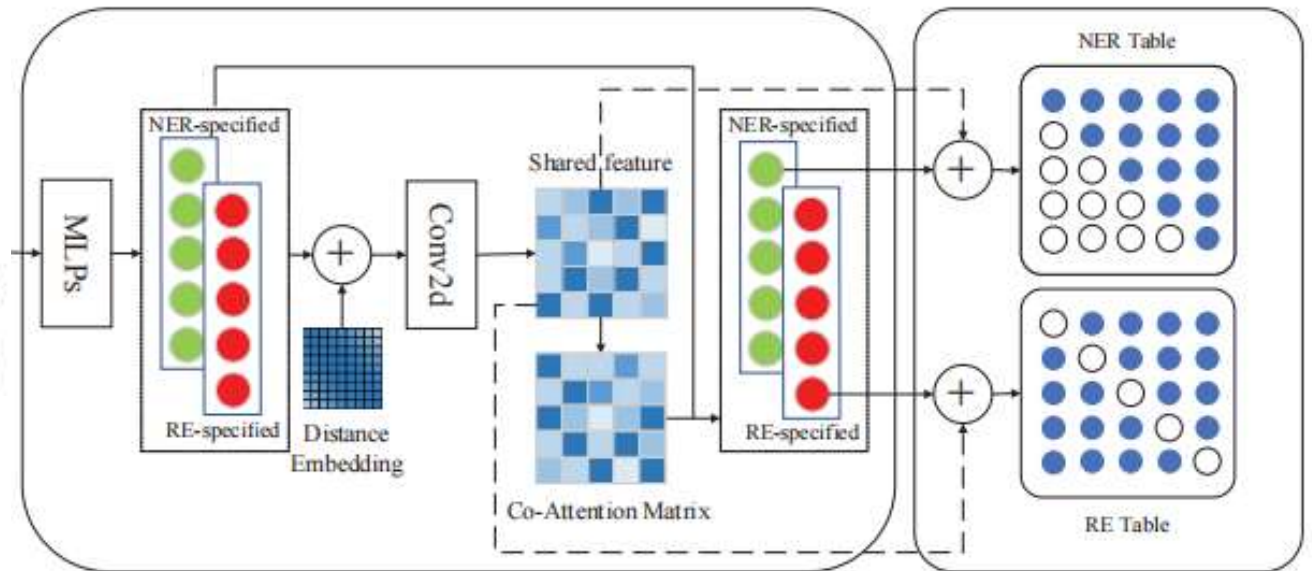
$$P(r|x_i, x_j) = \text{sigmoid}(W_d \sigma(W_c u_{ij}^r + b_c) + b_d)$$

- Joint training

$$\mathcal{L}_{NER} = - \sum_{i \leq j} \sum_{e \in E} \mathbb{I}(e = 1) \log P(e|x_i, x_j) \\ + \mathbb{I}(e = 0) \log(1 - P(e|x_i, x_j))$$

$$\mathcal{L}_{RE} = - \sum_{i \neq j} \sum_{r \in R} \mathbb{I}(r = 1) \log P(r|x_i, x_j) \\ + \mathbb{I}(r = 0) \log(1 - P(r|x_i, x_j))$$

$$\mathcal{L} = \mathcal{L}_{NER} + \mathcal{L}_{RE}$$



Experiments

Dataset	Method	PLM	NER	RE
NYT	CopyRL (Zeng et al., 2019)	-	-	72.1
	CasRel (Wei et al., 2020)	B	93.5	89.6
	TpLinker (Wang et al., 2020b)	B	-	91.9
	StereoRel (Tian et al., 2021)	B	-	92.2
	PFN (Yan et al., 2021)	B	95.8	92.4
	EmRel (Xu et al., 2022)	B	-	92.1
	CARE (Ours)	B	95.7	92.6
WebNLG	CopyRL (Zeng et al., 2019)	-	-	61.6
	CasRel (Wei et al., 2020)	B	95.5	91.8
	TpLinker (Wang et al., 2020b)	B	-	91.9
	StereoRel (Tian et al., 2021)	B	-	92.1
	PFN (Yan et al., 2021)	B	98.0	93.6
	EmRel (Xu et al., 2022)	B	-	92.9
	CARE (Ours)	B	98.1	93.9
SciERC	SPE (Wang et al., 2020a)	SciB	68.0	34.6
	UNIRE (Wang et al., 2021)	SciB	68.4	36.9
	PURE:single-sentence (Zhong and Chen, 2021)	SciB	66.6	35.6
	PURE:cross-sentence (Zhong and Chen, 2021)	SciB	68.9	36.8
	PFN (Yan et al., 2021)	SciB	66.8	38.4
	CARE (Ours)	SciB	69.9	40.9

Table 1: Performance on NYT, WebNLG and SciERC dataset. (PLM=pretrained language model, B=bert-base, SciB=scibert-base). The NER results of CasRel are reported from Yan et al. (2021).

Experiments

Setting	NER	Δ	RE	Δ
Default	69.9	-	40.9	-
- Distance embedding	69.3	0.6	40.4	0.5
- Shared representation	69.4	0.5	40.0	0.9
- 3 \times 3 convolution	69.1	0.8	40.2	0.7
- Co-attention	68.7	1.2	39.6	1.3

Table 2: Ablation study on the SciERC dataset.

Layers	NER	RE
N=1	69.5	40.3
N=2	69.6	40.6
N=3	69.9	40.9
N=4	68.2	40.5

Table 3: The impact of co-attention module depths on the SciERC dataset.

Conclusion

- We propose CARE, a **C**o-**A**ttention network for joint entity and **R**elation **E**xtraction, which can effectively exploit interactions between subtasks.
 - We adopt parallel encoding to separately learn task-specific representations for NER and RE, preventing feature confusion between the two subtasks.
 - Building upon parallel encoding, we introduce a co-attention mechanism to capture two-way interaction between NER and RE, effectively leveraging the information from one subtask to enhance the other.
- Extensive experiments on three benchmark datasets (NYT, WebNLG and SciERC) show that our model can achieve superior performance compared with existing methods, and ablation studies demonstrate the effectiveness of our method.