Exploring Self-distillation based Relational Reasoning Training for Document-Level Relation Extraction

Liang Zhang[1,2], Jinsong Su[1,2*], Zijun Min[1,2], Zhongjian Miao[1,2], Qingguo Hu[1,2]
Biao Fu[1,2], Xiaodong Shi[1,2], Yidong Chen[1,2*]
[1]School of Informatics, Xiamen University, China
[2]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
lzhang@stu.xmu.edu.cn, {jssu,ydchen}@xmu.edu.cn

Code:https://github.com/DeepLearnXMU/DocRE-SD

2023.11.2 • ChongQing

—— IJCAI 2023

gesis
Leibniz-Institut
für Sozialwissenschaften

Reported by Renhui Luo

Artificial

1.Introduction

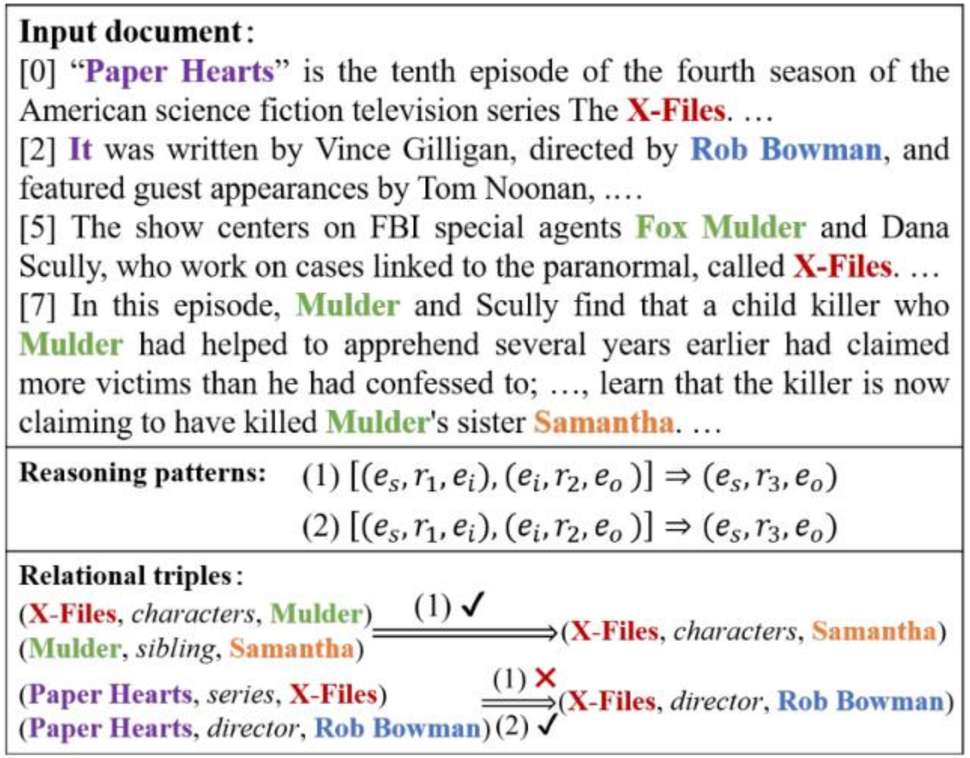
2.Overview

3.Methods

4.Experiments

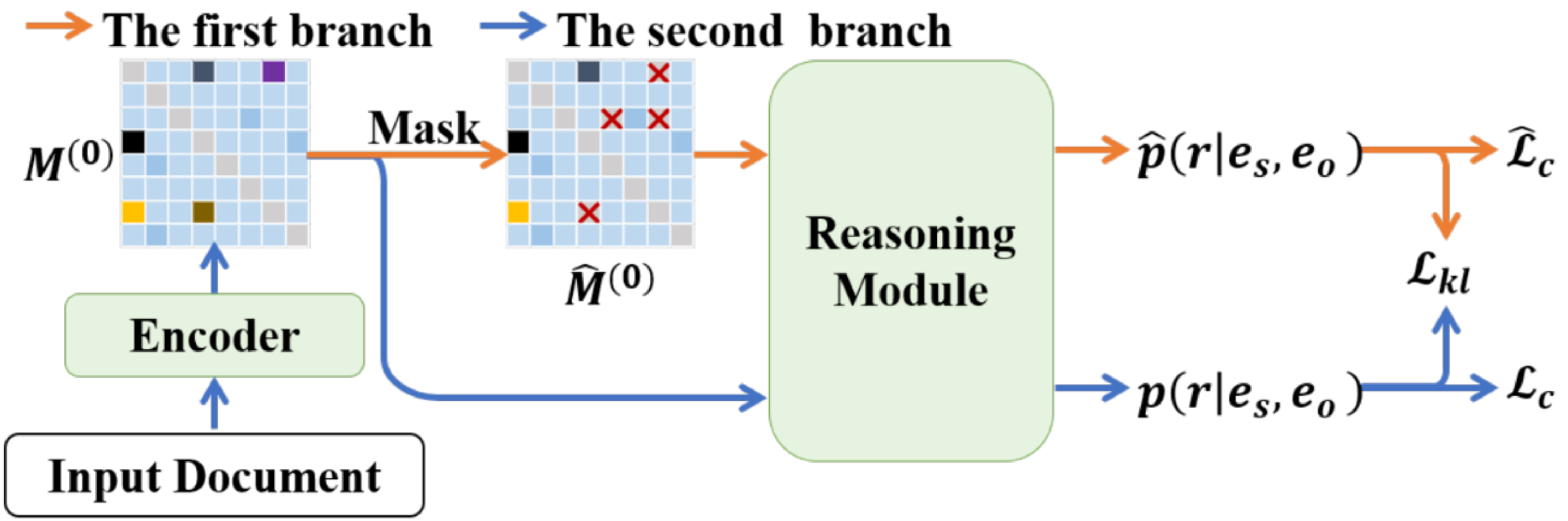
Leibniz-Institut
für Sozialwissenschaften

Introduction

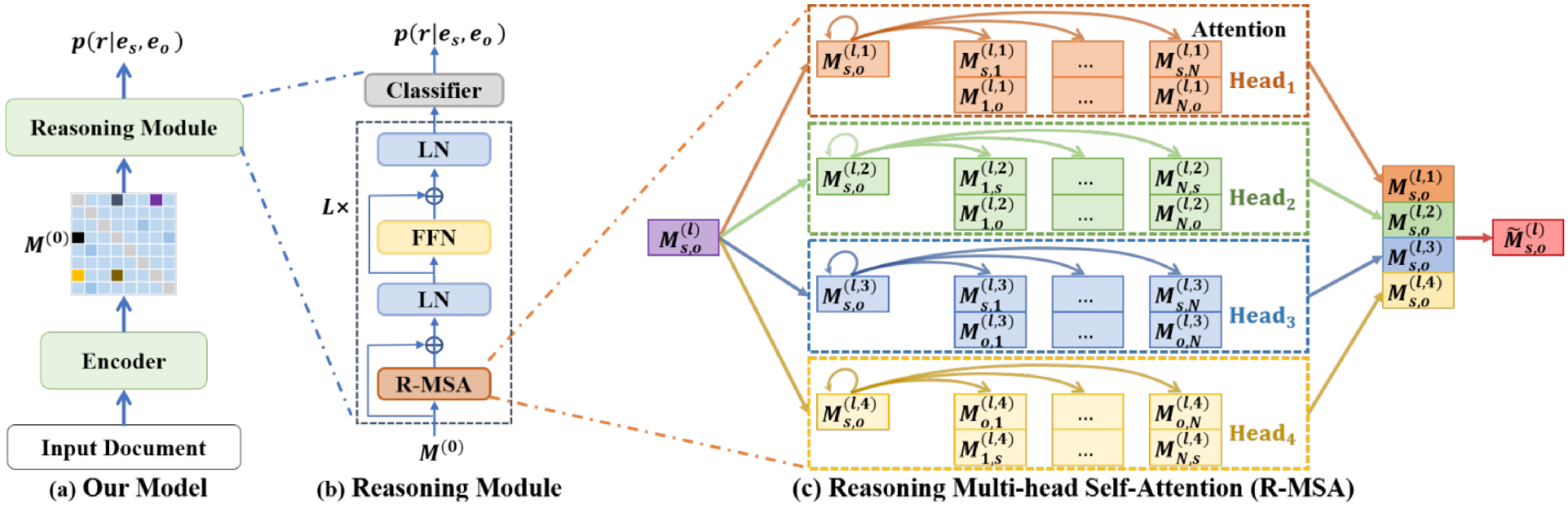


Reasoning Pattern	Example	Rate
(1) $[(e_s, r_1, e_i), (e_i, r_2, e_o)] \Rightarrow (e_s, r_3, e_o)$	[(Bob, father, Danny), (Danny, spouse, Anna)]⇒(Bob, mother, Anna)	24.83%
(2) $[(e_i, r_1, e_s), (e_i, r_2, e_o)] \Rightarrow (e_s, r_3, e_o)$	[(Bob, brother, Harry), (Bob, father, Danny)]⇒(Harry, father, Danny)	19.28%
(3) $[(e_s, r_1, e_i), (e_o, r_2, e_i)] \Rightarrow (e_s, r_3, e_o)$	[(Bob, father, Danny), (Harry, father, Danny)]⇒(Bob, brother, Harry)	24.69%
(4) $[(e_o, r_1, e_i), (e_i, r_2, e_s)] \Rightarrow (e_s, r_3, e_o)$	[(Bob, mother, Anna), (Anna, spouse, Danny)]⇒(Danny, child, Bob)	7.70%

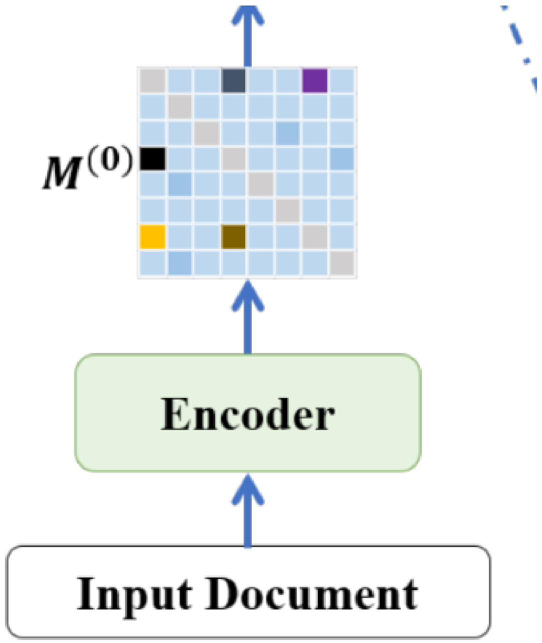
Overview


The first branch
The second branch
$M^{(0)}$
Mask
$\widehat{M}^{(0)}$
Encoder
Input Document
Reasoning Module
$\hat{p}(r|e_s, e_o)$
$\widehat{\mathcal{L}}_c$
$\mathcal{L}_{kl}$
$p(r|e_s, e_o)$
$\mathcal{L}_c$

Method


p(r|e_s, e_o)
Reasoning Module
M^(0)
Encoder
Input Document
(a) Our Model
p(r|e_s, e_o)
Classifier
LN
L×
FFN
LN
R-MSA
M^(0)
(b) Reasoning Module
Attention
Head_1
Head_2
Head_3
Head_4
(c) Reasoning Multi-head Self-Attention (R-MSA)

Method



$$\mathbf{H} = [h_1, h_2, ..., h_{|D|}]$$

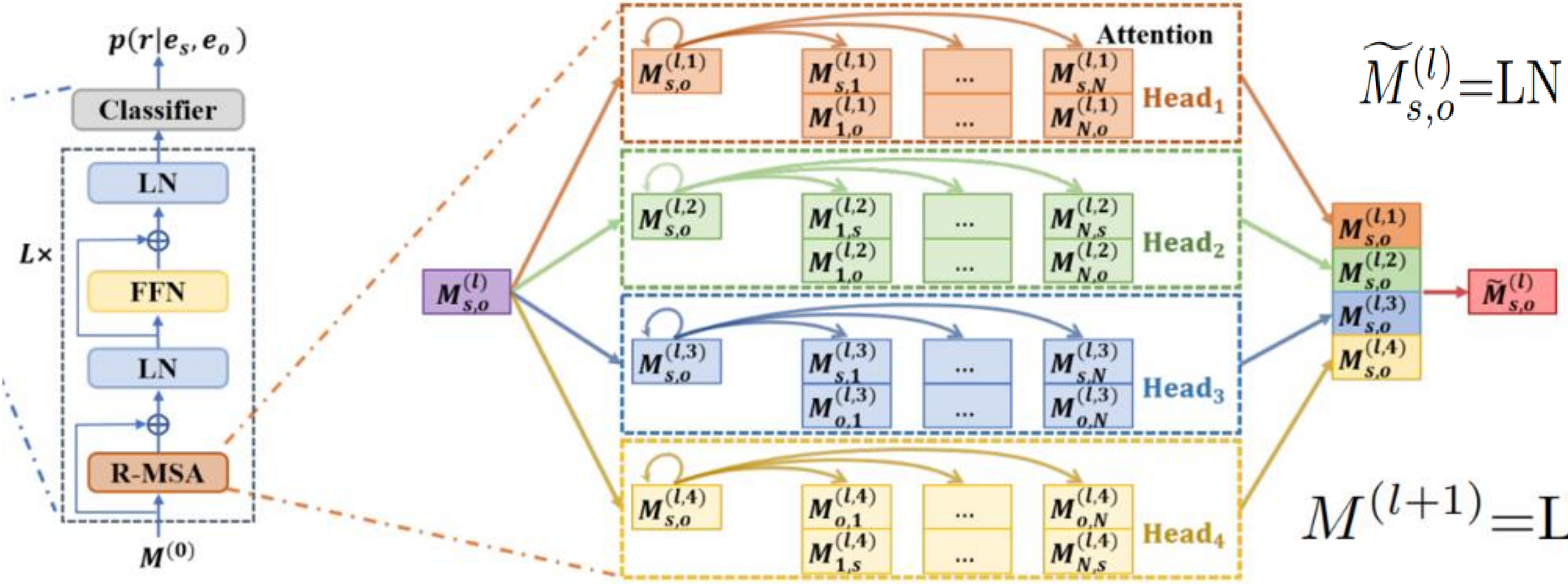
$$h(e_i) = \log \sum_{j=1}^{N_{e_i}} \exp(h(m_j^i))$$

$$\begin{aligned} F_{s,o} = \mathrm{FNN}\big(\big[&\tanh(W_s[h(e_s); c_{s,o}]); \\ &\tanh(W_o[h(e_o); c_{s,o}])\big]\big), \end{aligned} \quad (1)$$

$$c_{s,o} = \mathbf{H}^\top \frac{A_s \circ A_o}{\mathbf{1}^\top (A_s \circ A_o)}, \quad (2)$$

$$M^{(0)} = [F_{s,o}]_{N \times N}$$

Method



(b) Reasoning Module (c) Reasoning Multi-head Self-Attention (R-MSA)

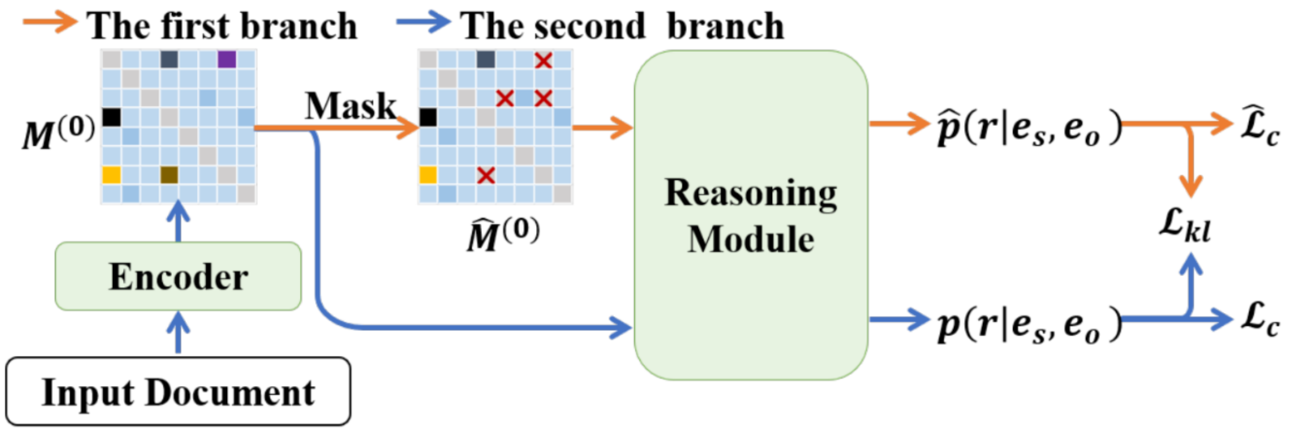
$$\widetilde{M}^{(l)}_{s,o} = \mathrm{LN}\left(M^{(l)} + W_O\left[M^{(l,1)}_{s,o}; \cdots ; M^{(l,4)}_{s,o}\right] + b_O\right), \quad (5)$$

$$M^{(l+1)} = \mathrm{LN}(\widetilde{M}^{(l)} + \mathrm{FNN}(\widetilde{M}^{(l)})), \quad (6)$$

$$F^{(l,1)}_i = W_d\left[M^{(l)}_{s,i}; M^{(l)}_{i,o}\right] + b_d, \quad i = \{1, 2, \cdots, N\}, \quad (3)$$

$$M^{(l,1)}_{s,o} = \mathrm{Attention}(Q, K, V),$$
$$\text{where } Q = M^{(l)}_{s,o}, \; K = V = \left[M^{(l)}_{s,o}; F^{(l,1)}_1; \cdots ; F^{(l,1)}_N\right]. \quad (4)$$

Method



$$p(r|e_s,e_o) = \sigma(W_c M_{s,o}^{(L)} + b_c), \tag{7}$$

$$\mathcal{L}_{kl} = \mathrm{KL}\big(p(r|e_s,e_o)||\widehat{p}(r|e_s,e_o)\big). \tag{8}$$

$$\mathcal{L} = \mathcal{L}_c + \widehat{\mathcal{L}}_c + \mathcal{L}_{kl}. \tag{9}$$

$$\mathcal{L}_c = -\Bigg(\sum_{r\in\mathcal{R}_{pos}} \log\Bigg(\frac{\exp(\mathrm{logit}_r)}{\sum_{r'\in\{\mathcal{R}_{pos},\mathrm{TH}\}}\exp(\mathrm{logit}_{r'})}\Bigg)\Bigg) - \log\Bigg(\frac{\exp(\mathrm{logit}_{\mathrm{TH}})}{\sum_{r'\in\{\mathcal{R}_{neg},\mathrm{TH}\}}\exp(\mathrm{logit}_{r'})}\Bigg). \tag{10}$$

$$\gamma_t = \min(\gamma_{max}, \frac{t}{T})$$

Experiments

Model	Dev					Test	
	IgnF_1	F_1	Intra-F_1	Inter-F_1	Infer-Ac	IgnF_1	F_1
GEDA-BERT (Li et al. 2020)†	54.52	56.16	–	–	–	53.71	55.74
LSR-BERT (Nan et al. 2020)†	52.43	59.00	65.26	52.05	–	56.97	59.05
GLRE-BERT (Wang et al. 2020)†	–	–	–	–	–	55.40	57.40
GAIN-BERT (Zeng et al. 2020)†	59.14	61.22	67.10	53.90	58.42*	59.00	61.24
HeterGSAN-BERT (Xu et al. 2021)†	58.13	60.18	–	–	–	57.12	59.45
SSAN-BERT (Xu et al. 2021)†	56.68	58.95	–	–	–	56.06	58.41
BERT-base (Wang et al. 2019)†	–	54.16	61.61	47.15	–	–	53.20
BERT-TS (Wang et al. 2019)†	–	54.42	61.80	47.28	–	–	53.92
HIN-BERT (Tang et al. 2020)†	54.29	56.31	–	–	–	53.70	55.60
CorefBERT (Ye et al. 2020)†	55.32	57.51	–	–	–	54.54	56.96
ATLOP-BERT (Zhou et al. 2021)†	59.22	61.09	–	–	58.29*	59.31	61.30
DocuNet-BERT (Zhang et al. 2021)†	59.86	61.83	–	–	–	59.93	61.86
SIRE-BERT (Zeng et al. 2021)†	59.82	61.60	68.07	54.01	–	60.18	62.05
KD-BERT (Tan et al. 2022)†	60.08	62.03	–	–	58.93*	60.04	62.08
Ours-BERT(SD→KD)	59.83	61.76	68.12	54.09	59.31	59.94	61.81
Ours-BERT(SD→R-Drop)	60.12	61.92	68.39	54.92	59.74	60.11	62.03
Ours-BERT	**60.85±0.10**	**62.81±0.13**	**68.67±0.11**	**56.09±0.21**	**61.08±0.18**	**60.91**	**62.85**

Table 2: Experimental results on the development and test sets of DocRED. We report the mean and standard deviation on the development set by conducting five experiments with different random seeds. Besides, we report the official test scores of the best checkpoint on the development set. † indicates original paper scores. Results with * are obtained by our reproduction. KD denotes the vanilla knowledge distillation and SD means our self-distillation training framework. SD→KD (SD→R-Drop) means to replace our SD with KD (R-Drop).

Experiments

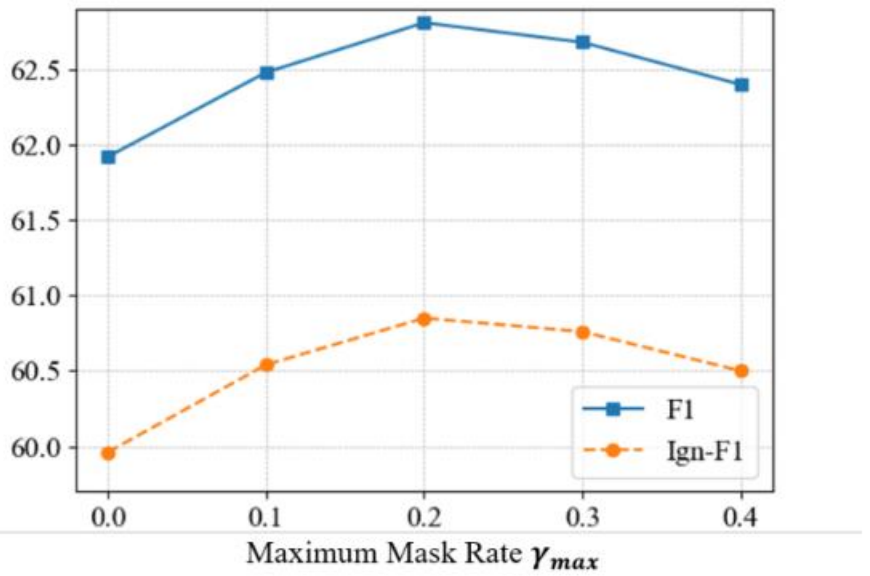


Figure 4: The performance of our model with different maximum mask rates γ_{max} on the development set of DocRED.

Experiments

Model	CDR	GDA
BRAN (Verga et al. 2018)	62.1	–
EoG (Christopoulou et al. 2019)	63.6	81.5
LSR (Nan et al. 2020)	64.8	82.2
DHG (Zhang et al. 2020)	65.9	83.1
GLRE (Wang et al. 2020)	68.5	–
SciBERT (Beltagy, Lo, and Cohan 2019)	65.1	82.5
ATLOP-SciBERT (Zhou et al. 2021)	69.4	83.9
DocuNet-SciBERT(Zhang et al. 2021)	76.3	85.3
Ours-SciBERT	**76.8**	**86.4**

Table 3: The F_1 scores on the CDR and GDA test sets.

Experiments

Model	IgnF_1	F_1
Ours-BERT	**60.85**	**62.81**
w/ R-MSA→MSA	57.45	59.39
w/ Only the first reasoning pattern	60.25	62.16
w/o The first branch	59.58	61.53
w/o The second branch	60.46	62.38
w/o Curriculum Learning	60.61	62.56

Table 4: Ablation study of our model on the development set of DocRED.

Experiments

Model	Infer-F_1	P	R
GAIN-GloVe	40.82	32.76	54.14
SIRE-GloVe	42.72	34.83	55.22
BERT-RE	39.62	34.12	47.23
GAIN-BERT	46.89	38.71	59.45
Ours-BERT	**50.11**	**42.99**	**60.05**
w/o The first branch	47.92	40.03	59.68
w/o Reasoning module	46.62	38.42	59.29

Table 5: Infer-F_1 scores on the development set of DocRED.

Experiments

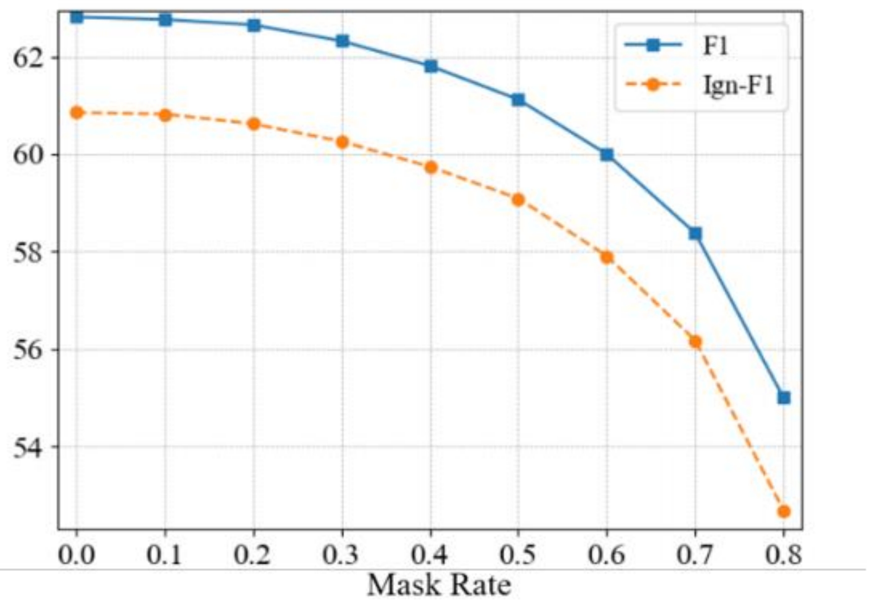


Figure 5: The performance of our model with different mask rates during testing on the development set of DocRED.

Thanks!