Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation

Qingyu Tan[*1,2] **Ruidan He**[†1] **Lidong Bing**[1] **Hwee Tou Ng**[2]

[1]DAMO Academy, Alibaba Group

[2]Department of Computer Science, National University of Singapore

{qingyu.tan,ruidan.he,l.bing}@alibaba-inc.com

{qtan6,nght}@comp.nus.edu.sg

Code:https://github.com/tonytan48/KD-DocRE

2022.05.06 • ChongQing ACL2022

Introduction

Challenge

1.Aside from the imbalance of positive and negative examples, the distribution of relation types for the positive entity pairs is also highly imbalanced.

2. A large amount of training data is the training data of remote supervision, and there is a lot of remote supervision, which limits the training of the model.

Problem Formulation

document D

$\{e_i\}_{i=1}^{n}$ $\{m_j^i\}_{j=1}^{N_{e_i}}$

$(e_s, e_o)_{s,o\in\{1...n\},s\neq o}$

$\mathbf{R} \cup \{\mathbf{NR}\}$

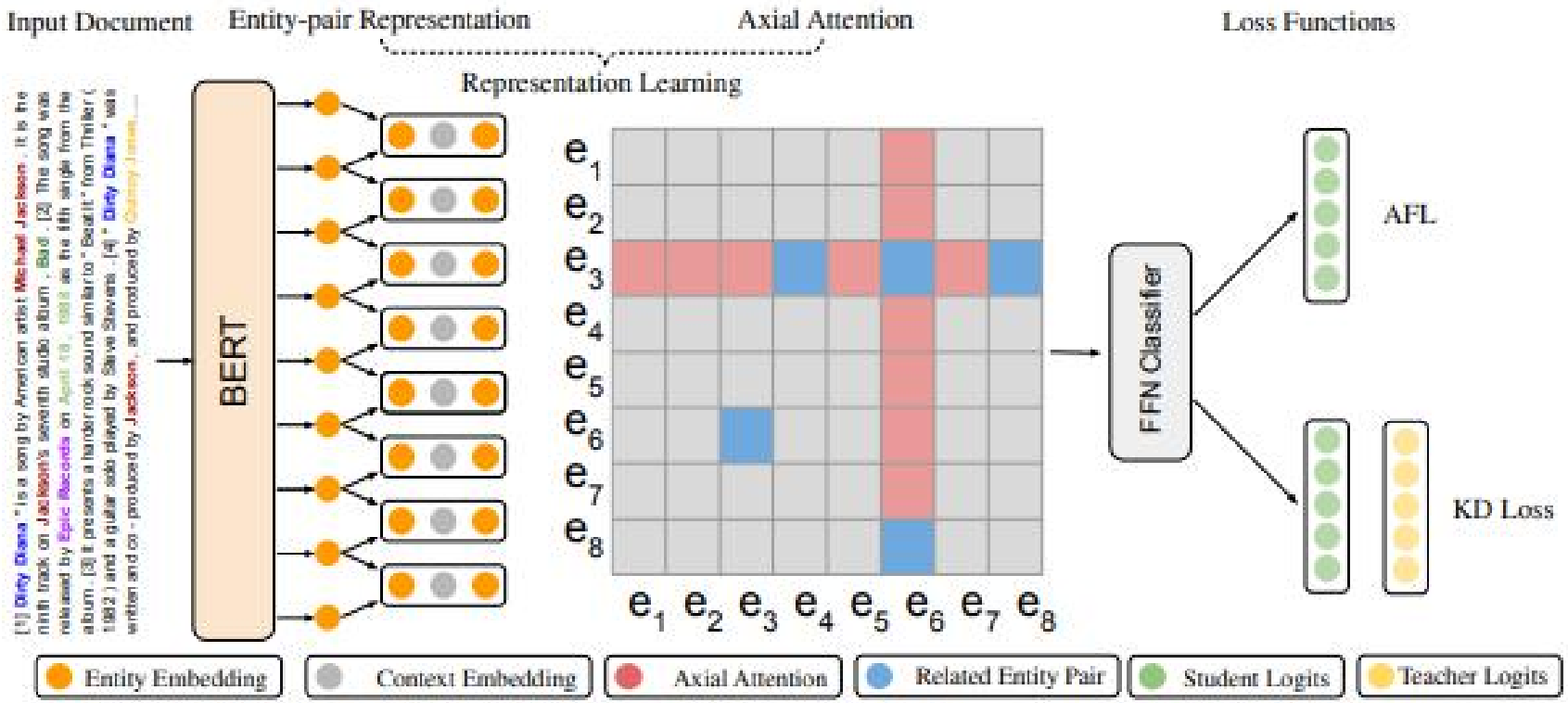


Figure 1: Model architecture of our DocRE system. We show the axial attention region for the entity pair (e_3, e_6).

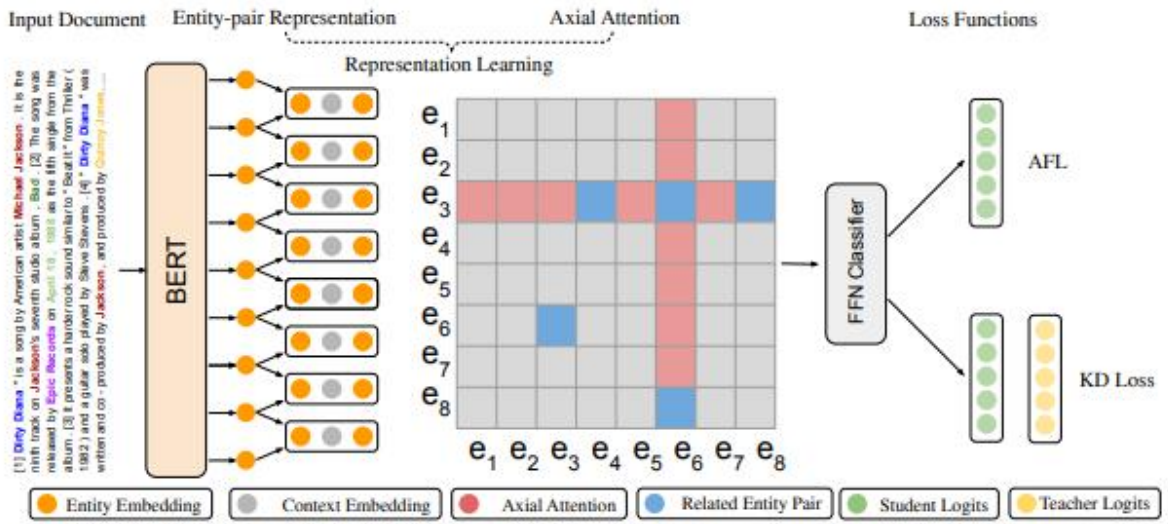


Figure 1: Model architecture of our DocRE system. We show the axial attention region for the entity pair (e_3, e_6).

Entity Representation

$$D = [x_t]_{t=1}^{l}$$

$$\mathbf{H} = PrLM([x_1, ..., x_l]) = [h_1, ..., h_l]) \quad (1)$$

$$h_{e_i} = log \sum_{j=1}^{N_{e_i}} exp(h_{m_j}) \quad (2)$$

Context-enhanced Entity Representation

$$q^{(s,o)} = \sum_{i=1}^{H} (A_{e_s}^{i} \circ A_{e_o}^{i}) \quad (3)$$

$$c^{(s,o)} = \mathbf{H}^{\top} q^{(s,o)} \quad (4)$$

$$z_s = tanh(\mathbf{W}_s h_{e_s} + \mathbf{W}_c c^{(s,o)}) \quad (5)$$

$$A_{e_i} = \sum_{j=1}^{N_{e_i}} (a_{m_j}), \text{ where } a_{m_j} \in \mathbb{R}^{H \times l}$$

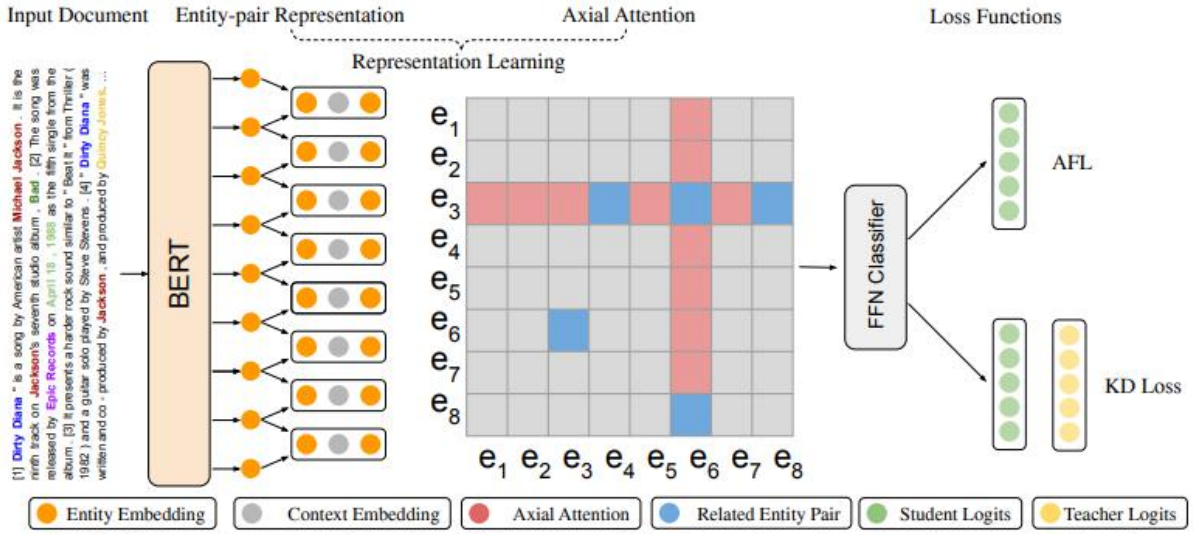


Figure 1: Model architecture of our DocRE system. We show the axial attention region for the entity pair (e_3, e_6).

Entity Pair Representation

$$z_s = [z_s^1, z_s^2, ..., z_s^k]$$

$$g_i^{(s,o)} = \sum_{j=1}^{k} (z_s^{j\intercal} W_{g_i}^j z_o^j) + b_i \tag{6}$$

$$g^{(s,o)} = [g_1^{(s,o)}, g_2^{(s,o)}, ..., g_d^{(s,o)}]$$

axial Attention-Enhanced Entity Pair Representation

$$r_w^{(s,o)} = r_h^{(s,o)} + \sum_{p \in 1..n} softmax_p(q_{(s,o)}^T k_{(s,p)}) v_{(s,p)}$$

$$r_h^{(s,o)} = g^{(s,o)} + \sum_{p \in 1...n} softmax_p(q_{(s,o)}^T k_{(p,o)}) v_{(p,o)} \tag{7}$$

where we denote query $q_{(i,j)} = W_Q g^{(i,j)}$, key $k_{(i,j)} = W_K g^{(i,j)}$, and value $v_{(i,j)} = W_V g^{(i,j)}$, which are all linear projections of the entity pair representation g at position (i, j). $W_Q \in \mathbb{R}^{d \times d}$, $W_K \in \mathbb{R}^{d \times d}$, and $W_V \in \mathbb{R}^{d \times d}$ are all learnable weight matrices. The output of the axial attention module is $r_w^{(s,o)} \in \mathbb{R}^d$. The $softmax_p$ function denotes a softmax function that applies to all possible $p = (i, j)$ positions. The formulation of this mechanism resembles Wang et al. (2020). However, our motivation is different, as Wang et al. (2020) aim to use this mechanism to reduce the computational complexity of semantic segmentation, whereas our motivation is to attend to the one-hop neighbors for the two-hop relation triplets.

Adaptive Focal Loss

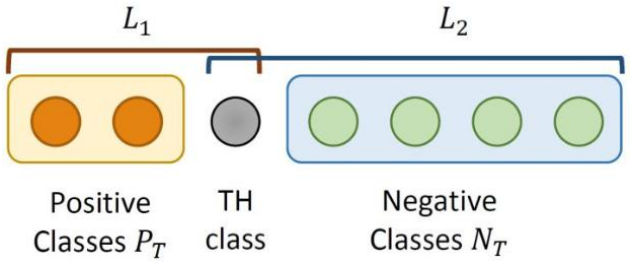
Adaptive Thresholding Loss

$$l^{(s,o)} = \mathbf{W}_l r_w^{(s,o)} + b_l \tag{8}$$

ATL introduced a special class *TH* as the adaptive threshold value for each example.

$\mathcal{P}_T$: positive class subset

$\mathcal{N}_T$. : negative class subset



$$P(r_i|e_s,e_o) = \frac{exp(l_{r_i}^{(s,o)})}{exp(l_{r_i}^{(s,o)}) + exp(l_{TH}^{(s,o)})} \tag{9}$$

$$P(r_{TH}|e_s,e_o) = \frac{exp(l_{rTH}^{(s,o)})}{\sum_{r_j \in \mathcal{N}_T \cup \{TH\}} exp(l_{r_j}^{(s,o)})} \tag{10}$$

$$\mathcal{L}_{RE} = \sum_{r_i \in \mathcal{P}_T} (1-P(r_i))^{\gamma} log(P(r_i)) + log(P(r_{TH})) \tag{11}$$

Knowledge Distillation

$$\mathcal{L}_{KD} = MSE(l_S^{(s,o)}, l_T^{(s,o)}) \qquad (12)$$

$$\mathcal{L} = \mathcal{L}_{KD} + \mathcal{L}_{RE} \qquad (13)$$

Experiment

Statistics	DocRED	HacRED
# distant docs	101,873	–
# training docs	3,053	6,231
# dev docs	1,000	1,500
# test docs	1,000	1,500
# relations	97	27
Avg # entities per doc	19.5	10.8
Avg # mentions per entity	1.4	1.2
Avg # relations per doc	12.5	7.4

Table 1: Dataset statistics of the DocRED and HacRED datasets.

	Dev		Test	
w/o Distant Supervision	Ign_F1	F1	Ign_F1	F1
Two-stage-B-b	56.67	58.83	56.47	58.69
ATLOP-B-b	59.22±0.15	61.09±0.16	59.31	61.30
SIRE-B-b	59.82	61.60	**60.18**	62.05
DocuNet-B-b	59.86±0.13	61.83±0.19	59.93	61.86
Ours-B-b	**60.08**±0.11	**62.03**±0.18	60.04	**62.08**
Coref-Rb-l	57.35	59.43	57.9	60.25
SSAN-Rb-l	59.40	61.42	60.25	62.08
GAIN-B-l	60.87	63.09	60.31	62.76
ATLOP-Rb-l	61.32±0.14	63.18±0.19	61.39	63.40
DocuNet-Rb-l	**62.23**±0.12	64.12±0.14	62.39	**64.55**
DocuNet-Rb-l*	61.56±0.14	63.58±0.17	61.79	63.73
Ours-Rb-l	62.16±0.10	**64.19**±0.16	**62.57**	64.28
with Distant Supervision	Ign_F1	F1	Ign_F1	F1
ATLOP-NA-Rb-l*	63.41±0.15	65.33±0.18	63.54	65.47
DocuNet-NA-Rb-l*	63.26±0.17	65.21±0.19	63.29	65.44
SSAN-NA-Rb-l	63.76	65.69	63.78	65.92
Ours-NA-B-b	62.18±0.12	64.17±0.16	61.77	64.12
Ours-KD-B-b	62.62±0.16	64.81±0.13	62.56	64.76
Ours-NA-Rb-l	63.38±0.11	65.64±0.17	63.63	65.71
Ours-KD-Rb-l	**65.27**±0.09	**67.12**±0.14	**65.24**	**67.28**

Table 2: Experimental results for the DocRED dataset. The reported metrics are F1 score and Ign_F1. We report the average of five random runs for the development set and the best checkpoint is used for the leaderboard submission for the test results. Results with * are obtained by our reproduction.

Experiments

	P	R	F1
GAIN*	73.38	**80.07**	76.09
ATLOP*	76.97	78.29	77.63
Ours	**78.53**	78.96	**78.75**

Table 3: Experimental results on HacRED dev set. Results with * are implemented by us. All experiments used XLM-R-base as the encoder.

Experiments

	Frequent F1	Long-tail F1	Overall F1
ATLOP-Rb-l	70.93	50.01	63.12
Ours-Rb-l	**71.26**	**51.97**	**64.19**
w/o **Axial**	70.86	50.77	63.56
w/o **AFL** *w* **ATL**	70.94	50.86	63.67
With Distant Supervision			
ATLOP-NA-Rb-l	73.26	52.39	65.33
Ours-KD-Rb-l	**74.15**	**56.51**	**67.12**
w/o **Axial**	73.52	54.96	66.36
w/o **AFL** *w* **ATL**	73.50	54.73	66.23

Table 4: Experiment results for frequent and long-tail type relations. Frequent types refer to the most popular 10 relation types, and long-tail relations refer to the rest of the 86 relations.

	P	R	Infer-F1
GAIN-B-b	38.71	59.45	46.89
Ours-Rb-l	**42.15**	**61.56**	**50.04**
w/o **Axial**	40.26	60.60	48.37

Table 5: Ablation study for the Infer-F1 relation triples on the development set of DocRED.

Experiments

Distant Adaptation	Ign_F1	F1
NA	52.29	54.67
KD$_{KL}$	53.89	56.97
KD$_{MSE}$	**55.28**	**57.74**
Continue-trained	Ign_F1	F1
NA	63.38	65.64
KD$_{KL}$	64.42	66.24
KD$_{MSE}$	**65.27**	**67.12**

Table 6: Development set performance of different knowledge adaptation methods for DocRED.

Experiments

		Ground Truth: $r \in \mathbf{R}$	Ground Truth: NR
Predictions	$r \in \mathbf{R}$	**C**: 8,273 (51.4%)	**MR**: 3,814 (23.7%)
		W: 242 (1.5%)	
	NR	**MS**: 3,761 (23.4%)	380,703

Table 7: Statistics of our error distribution. The final evaluation score is evaluated on $r \in \mathbf{R}$ triples, hence the correct predictions of **NR** are ignored when calculating the final scores.

	P	R	F1
Binary Labels	68.51	68.78	68.64
Original Labels	67.10	67.13	67.12

Table 8: Performance breakdown on the DocRED dev set.

Thanks