Extracting Entities and Relations with Joint Minimum Risk Training (Supplementary)

Changzhi Sun[2], **Yuanbin Wu**[1, 2], **Man Lan**[1, 2], **Shiliang Sun**[2],
Wenting Wang[3], **Kuang-Chih Lee**[3], **and Kewen Wu**[3]

[1]Shanghai Key Laboratory of Multidimensional Information Processing
[2]Department of Computer Science and Technology, East China Normal University
[3]Alibaba Group
{changzhisun}@stu.ecnu.edu.cn
{ybwu,mlan,slsun}@cs.ecnu.edu.cn
{nentiao.wwt,kuang-chih.lee,kewen.wukw}@alibaba-inc.com

A The Gradient of the MRT Objective

We give a detailed derivation of the gradient of the MRT objective (Equation 6). The calculation is standard, and similar derivations have been given in (Xu et al., 2016; Shen et al., 2016). It is worth remarking that the non-decomposability of MRT's loss doesn't make the objective non-differentiable. In fact, given the sample set $\mathcal{Y}'(s)$, the following gradient can be computed by autograd tools.

$$\mathcal{L}_{\text{mrt}}(\boldsymbol{\theta}) = \sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} Q(\hat{\mathbf{y}}|s;\boldsymbol{\theta},\alpha)\Delta(\hat{\mathbf{y}},\mathbf{y}) = \frac{\sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}\Delta(\hat{\mathbf{y}},\mathbf{y})}{\sum_{\mathbf{y}^* \in \mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}} \triangleq \frac{G(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}.$$

The gradient of the enumerator $G(\boldsymbol{\theta})$ is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) &= \alpha \sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha-1} \nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})\Delta(\hat{\mathbf{y}},\mathbf{y}) \\ &= \alpha \sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha} \frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}\Delta(\hat{\mathbf{y}},\mathbf{y}).\end{aligned}$$

The gradient of the denominator $Z(\boldsymbol{\theta})$ is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) &= \alpha \sum_{\mathbf{y}^* \in \mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha-1} \nabla_{\boldsymbol{\theta}} P(\mathbf{y}^*|s;\boldsymbol{\theta}) \\ &= \alpha \sum_{\mathbf{y}^* \in \mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha} \frac{\nabla_{\boldsymbol{\theta}} P(\mathbf{y}^*|s;\boldsymbol{\theta})}{P(\mathbf{y}^*|s;\boldsymbol{\theta})}.\end{aligned}$$

Hence, we have

$$\begin{aligned}\frac{G'}{Z} &= \alpha \sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} \frac{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}}{\sum_{\mathbf{y}^* \in \mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}} \frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}\Delta(\hat{\mathbf{y}},\mathbf{y}) \\ \frac{Z'}{Z} &= \alpha \sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} \frac{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}}{\sum_{\mathbf{y}^* \in \mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}} \frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}.\end{aligned}$$

Following the standard result,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{mrt}}(s;\boldsymbol{\theta},\alpha) &= \frac{G' \cdot Z - G \cdot Z'}{Z \cdot Z} = \frac{G'}{Z} - \frac{G}{Z} \cdot \frac{Z'}{Z} = \frac{G'}{Z} - \mathcal{L}_{\text{mrt}} \cdot \frac{Z'}{Z} \\ &= \alpha \mathbf{E}_{\hat{\mathbf{y}} \sim Q(\hat{\mathbf{y}}|s;\boldsymbol{\theta},\alpha)} \left[\frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})} \left[\Delta(\hat{\mathbf{y}},\mathbf{y}) - \mathcal{L}_{\text{mrt}}(s;\boldsymbol{\theta},\alpha)\right] \right].\end{aligned}$$

B Default Hyper-parameter Settings

Table A lists the default model configurations.

Modules	Setting
Embeddings	· $\dim(\mathbf{w}_i) = 100$, initialized with Glove vectors (Pennington et al., 2014) · $\dim(\mathbf{c}_i) = \dim(\text{char embedding}) = 50$ the window sizes of CNN ($\boldsymbol{\theta}_c$) are 2 and 3
NN	· $\dim(\mathbf{h}_i) = 128$ · $\dim(\mathbf{f}_{e_1}) = \dim(\mathbf{f}_{e_2}) = \dim(\mathbf{f}_{\text{middle}}) = 50$ · the window sizes of CNNs ($\boldsymbol{\theta}_{e_1}, \boldsymbol{\theta}_{e_2}$) are 2 and 3 · training epochs: 1000 (ACE05), 50 (NYT) · Adadelta: gradient clipping with max norm 1 · batch size: 100 · dropout rate: 0.5
MRT	· without Γ: $\mu = 1.0, \alpha = 0.0001, K = 3$ · with Γ: $\mu = 1.0, \alpha = 1, K = 2$ · training epochs: 25 (ACE05), 10 (NYT)

Table A: Model configurations.

Relation Type	Model	P	R	F
ART (146)	M&B (2016)	36.3	55.2	43.8
	K&C (2017)	43.1	**61.1**	**50.5**
	NN	51.6	44.5	47.8
	MRT	**59.2**	41.8	49.0
PART-WHOLE (175)	M&B (2016)	56.0	**53.8**	54.8
	K&C (2017)	52.0	**53.8**	52.8
	NN	57.2	49.7	53.2
	MRT	**59.9**	52.0	**55.7**
PER-SOC (73)	M&B (2016)	67.1	67.1	67.1
	K&C (2017)	65.7	64.8	65.2
	NN	76.5	**71.2**	**73.8**
	MRT	**77.3**	69.9	73.4

Relation Type	Model	P	R	F
PHYS (278)	M&B (2016)	48.9	**51.3**	**50.0**
	K&C (2017)	38.8	42.6	40.6
	NN	45.8	48.9	47.3
	MRT	**50.0**	42.8	46.1
GEN-AFF (99)	M&B (2016)	41.4	**64.0**	**50.2**
	K&C (2017)	48.4	51.6	50.0
	NN	56.1	37.4	44.9
	MRT	**60.9**	39.4	47.9
ORG-AFF (354)	M&B (2016)	69.2	70.4	69.7
	K&C (2017)	70.6	70.0	70.3
	NN	72.1	**72.3**	72.2
	MRT	**78.0**	70.1	**73.8**

Table B: Results on different relation types. The numbers in the first column are counts of relations in the ACE05 test set.

C More Discussions on Experiments results

We list performances on each entity type in Table B. The results show that MRT is able to improve precision on different relation types. We suspect that by using the more general F_β score in $\Delta(\mathbf{y}', \mathbf{y})$, we could observe different behaviors on precision and recall.

$$F_\beta = (1 + \beta^2)\frac{PR}{\beta^2 P + R}.$$

Figure A illustrates performances of NN and MRT with respect to different distances between candidate entity pairs. We find that MRT only outperforms NN at distance 2. It implies that although the sentence-level F1 provides global information in the loss function, it is still not powerful enough to capture long distance dependencies. Thus, integrating MRT and joint decoding algorithms might be a promising direction.

In Table 2, we report performances of Δ_R and Δ_E under the default hyper-parameter setting in Table A, which is tuned with Δ_{E+R}. To make a fair comparison, here we give results of Δ_E and Δ_R with tuned hyper-parameters (Table C) and also their development set performances (Figure B). On Table C, relation extraction performances of Δ_E is improved with tuned parameters (comparing with Table 2), and weighting the entity model and the relation model with $\mu = 0.5$ could help Δ_R. As mentioned in the paper, although simple tuning could improve all settings of $\Delta(\hat{\mathbf{y}}, \mathbf{y})$, we find that it is hard to tell why such settings can obtain good performances, and it is valuable to study the interpretability of MRT settings in future work.

Finally, we note that since both development set and test set of ACE05 and NYT are small, the model selection process is somewhat fragile. For example, by cheating in the process of tuning α, μ, we find

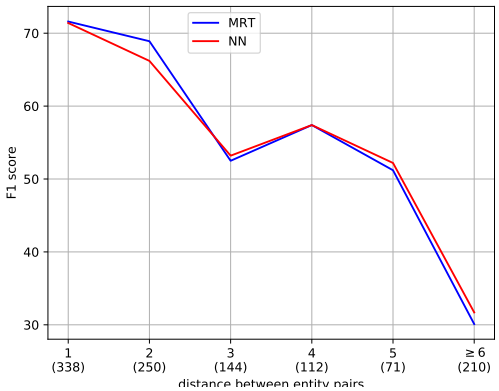
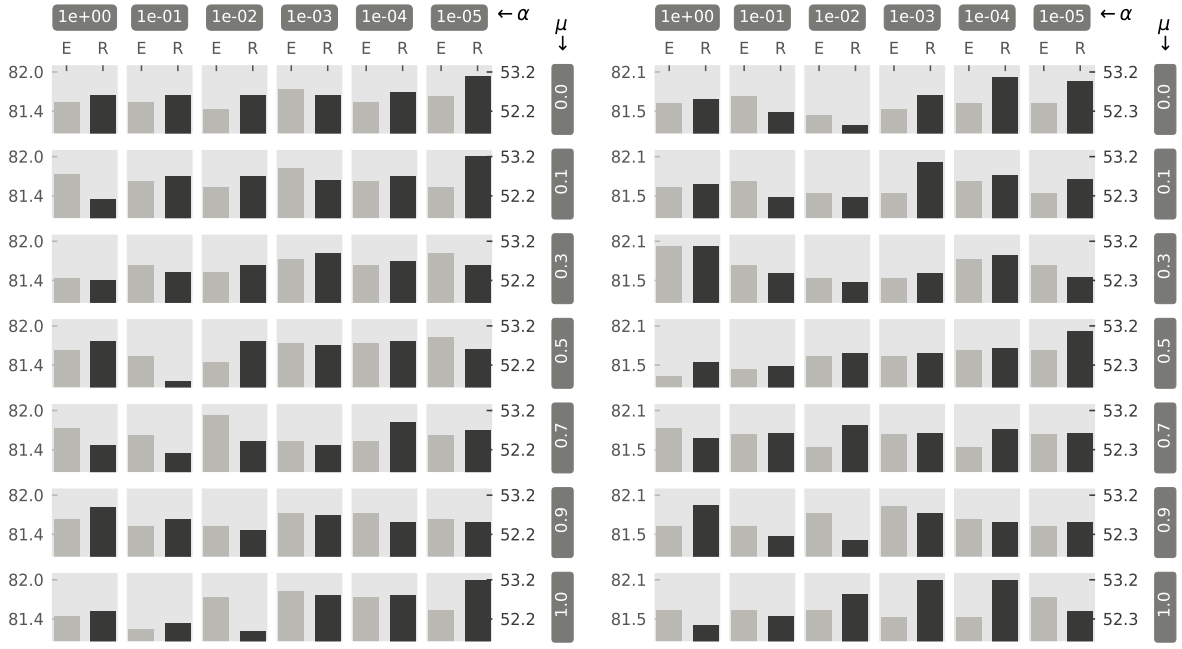


Figure A: F1 scores with respect to the distance between entity pairs. The numbers in parentheses are counts of relations in the ACE05 test set.

Settings	**F1 of Entity**	**F1 of Relation**	α	μ
Δ_{E+R}	83.6 +0.2	59.0 +1.2	1e−4	1.0
Δ_E	83.6 +0.2	58.6 +0.8	1e−5	1.0
Δ_R	83.4 +0.0	58.8 +1.0	1e−5	0.5

Table C: MRT with tuned Δ_E and Δ_R. The sample size K is fixed to 3 (default sampling).

that the testing time behavior and the validation time behavior could be quite different. Thus, how to strengthen the model selection strategy is also an important task for the two datasets.



(a) Δ_E (b) Δ_R

Figure B: Tuning α, μ with Δ_E and Δ_R in the development set. To break ties, we heuristically select models with a smaller α and a larger μ.

S1	first an update on a long running air safety investigation a year and a half near an [airline]$_{\text{VEH}}$ crashed near [new york] $_{\text{GPE,PART-WHOLE-1}}$'s [kennedy airport]$_{\text{FAC: PART-WHOLE-1}}$ there is controversy whether the disaster could have been averted .
NN	first an update on a long running air safety investigation a year and a half near an [airline]$_{\text{VEH}}$ crashed near [new york] $_{\text{GPE,PART-WHOLE-1}}$'s [kennedy airport]$_{\text{FAC: PART-WHOLE-1}}$ there is controversy whether the disaster could have been averted .
MRT	first an update on a long running air safety investigation a year and a half near an [airline]$_{\text{VEH}}$ crashed near [new york] $_{\text{GPE,PART-WHOLE-1}}$'s kennedy [airport]$_{\text{FAC: PART-WHOLE-1}}$ there is controversy whether the disaster could have been averted .

S2	the question , [i]$_{\text{PER}}$ 'm an [aol] $_{\text{ORG:ORG-AFF-1}}$ [shareholder] $_{\text{PER: ORG-AFF-1}}$ sitting at [home] $_{\text{FAC}}$, hearing this news , done this set off a few alarms ?
NN	the question , [i]$_{\text{PER}}$ 'm an aol [shareholder] $_{\text{PER:PHYS-1}}$ sitting at [home] $_{\text{FAC: PHYS-1}}$, hearing this news , done this set off a few alarms ?
NN	the question , [i]$_{\text{PER}}$ 'm an [aol] $_{\text{ORG:ORG-AFF-1}}$ [shareholder] $_{\text{PER:ORG-AFF-1,PHYS-1}}$ sitting at [home] $_{\text{FAC: PHYS-1}}$, hearing this news , done this set off a few alarms ?

S3	[our] $_{\text{ORG: ORG-AFF-1}}$ [founder] $_{\text{PER: ORG-AFF-1}}$ here at [cnn] $_{\text{ORG}}$, [ted turner]$_{\text{PER}}$, has sold more than half 0 [his]$_{\text{PER: ORG-AFF-2}}$ stake in [aol time warner] $_{\text{ORG: ORG-AFF-2}}$.
NN	[our] $_{\text{ORG}}$ [founder] $_{\text{PER}}$ here at [cnn] $_{\text{ORG}}$, [ted turner]$_{\text{PER}}$, has sold more than half 0 [his]$_{\text{PER: ORG-AFF-2}}$ stake in [aol] $_{\text{ORG: ORG-AFF-2}}$ time [warner] $_{\text{PER}}$.
MRT	[our] $_{\text{ORG: ORG-AFF-1}}$ [founder] $_{\text{PER: ORG-AFF-1}}$ here at [cnn] $_{\text{ORG}}$, [ted turner]$_{\text{PER}}$, has sold more than half 0 [his]$_{\text{PER: ORG-AFF-2}}$ stake in [aol] $_{\text{ORG: ORG-AFF-2}}$ time [warner] $_{\text{PER}}$.

S4	[john scottsdale] $_{\text{PER: PHYS-1}}$ is on the front lines in [iraq]$_{\text{GPE: PHYS-1}}$.
NN	[john scottsdale] $_{\text{PER: PHYS-1}}$ is on the front lines in [iraq]$_{\text{GPE: PHYS-1}}$.
MRT	[john scottsdale] $_{\text{PER}}$ is on the front lines in [iraq]$_{\text{GPE}}$.

D Error Analyses

In this section, we examine performances of the proposed MRT model on concrete examples. We will focus on comparing testing results of NN and MRT. In the following examples, we will use notations like "[entity span] $_{\text{ENT-TYPE[:REL-TYPE\_REL-ID]}}$". It means that an entity mention ("entity span") has an entity type ENT-TYPE, and (optionally) participates a REL-TYPE relation identified with REL-ID.

First, we show that MRT could help to discover those entities appearing in relations. For S1, NN identifies the entity "[kennedy airport] $_{\text{FAC}}$", but MRT misses "kennedy" of this entity. For S2, MRT identifies the entity "[aol] $_{\text{ORG}}$", but NN misses it. The two examples show that, although they both have some errors in entity detection, the MRT setting will bias the entity detector to find entities which may appear in relations.

Next, we give two examples where the candidate entity pairs have different distance. For S3, MRT identifies a ORG-AFF relation between "[our] $_{\text{ORG}}$" and "[founder] $_{\text{PER}}$", while NN does not find this relation even the entities are correct. For S4, MRT does not detect PHYS relation while NN correctly find it. It shows that, as Figure A, MRT is more powerful when the entities are close. Hence, how to improve its performances on distant entities might be an important future work.