Towards Integration of Discriminability and Robustness For Document-Level Relation Extraction

Jia Guo[1,2], Stanley Kok[1], Lidong Bing[2]

[1]School of Computing, National University of Singapore

[2]DAMO Academy, Alibaba Group

Introduction

- ***Sentence-level relation extraction***
  - Single sentence input
  - Single entity mention
  - Single entity pair
  - Single label output

Input sentence:
General Director Placido Domingo will return to sing and conduct on the Washington stage.

Given entity pair: <Placido Domingo, Washington stage>
Output relation: "org:top_members/employees"

- ***Document-level relation extraction***

An example from a sentence-level RE dataset (TACRED)

  - Multiple sentence inputs
  - Multiple entity mentions
  - Multiple entity pairs
  - Multiple label outputs

Input document:
(1) The culture of Los Angeles[1] is rich with arts and ethnically diverse.
(2) The greater Los Angeles[1] metro area has several notable art museums including the Los Angeles County Museum of Art[2] (LACMA[2]), the J. Paul Getty Museum[3] on the Santa Monica[4] mountains overlooking the Pacific[5], the Museum of Contemporary Art[6] (MOCA[6]), the Hammer Museum[7] and the Norton Simon Museum[8].
(3) In the 1920s and 1930s ….

Given entity pair:
<J. Paul Getty Museum, Los Angeles>
Output relations:
"located in the administrative territorial entity",
"headquarters location".

An example from a document-level RE dataset (DocRED)

- ***Challenges:***
  - a. Inadequate in effectively distinguishing relations
  - b. Lack of sufficient learning for long-tailed relations
  - c. Vulnerable to annotation errors or missing annotations

- ***Our contributions:***
  - *A novel model based on Pairwise moving-threshold loss, Entropy Minimization, and Supervised Contrastive Learning (PEMSCL)*
    - a. Boost the discriminability of both probabilistic scores and internal embeddings
    - b. Adapt supervised contrastive learning for long-tailed relations
    - c. Improve model robustness by a novel negative label sampling strategy
  - *Validate the effectiveness of our model in various settings*

Negative Label Sampling Strategy

- Annotation error problem:
  - Some entity pairs labeled as NA class should have at least one relation label [Tan et al., EMNLP'22].
  - → The negative relations of NA examples may be wrong!
- We randomly sample a small set of negative relations for NA-labeled entity pairs. For $\mathcal{B}_{\mathcal{N}}, \mathcal{L}_1$ is modified as:

$$\mathcal{L}' = \sum_{(h,t)\in\mathcal{B}_{\mathcal{N}}} \sum_{r\in\mathcal{N}'_{h,t}} -\log P^{\eta}_{h,t}(r) + \frac{1}{\gamma_2}\sum_{r\in\mathcal{N}'_{h,t}} H_{h,t}(r)$$

- 1st loss function with negative label sampling :

$$\mathcal{L}^{\mathrm{NA}}_1 = \boxed{\mathcal{L}'} + \sum_{(h,t)\in\mathcal{B}_{\mathcal{P}}} \boxed{\mathcal{L}^{h,t}_{pmt} + \mathcal{L}^{h,t}_{em}}$$

Improve robustness

Improve discriminability

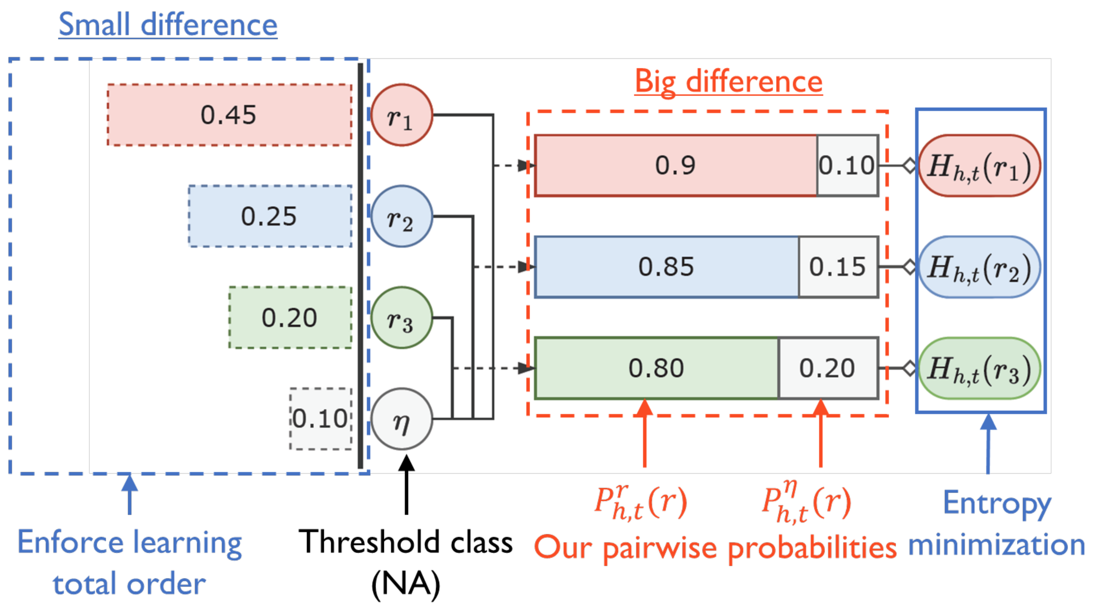
❖ **Our PEMCL with negative label sampling:**

$$\mathcal{L}^{\mathrm{NA}} = \mathcal{L}^{\mathrm{NA}}_1 + \boxed{\lambda\mathcal{L}_2}$$

Method

- ***Our PEMSCL model***
  - I. Pairwise moving-threshold loss with Entropy Minimization
  - II. Supervised Contrastive Learning for multi-labels and long-tailed relations
  - III. Negative label sampling strategy

- ***Problem definition:***



- Inputs:
  - A document: $D = \{w_l\}_{l=1}^{L}$ containing L words
  - A set of entities $\mathcal{E}_D = \{e_i\}_{i=1}^{|\mathcal{E}_D|}$
  - Each entity e_i is associated with a set of mentions $\mathcal{M}_{e_i} = \{m^i_j\}_{j=1}^{|\mathcal{M}_{e_i}|}$
- Outputs:
  - For each entity pair, $(e_h, e_t)_{h,t=1,\dots,|\mathcal{E}_D|, h\neq t}$, the model predicts a subset of pre-defined relations $\mathcal{R} = \{r_k\}_{k=1}^{|\mathcal{R}|}$
  - If an entity pair does not express any relation, it is labeled as NA

- ***Pairwise moving-threshold loss with entropy minimization***
- Split the predefined relation set into two mutually exclusive sets for (h,t): $\mathcal{R} = \mathcal{P}_{h,t} \cup \mathcal{N}_{h,t}$
  - Positive relations (i.e., the labels): $\mathcal{P}_{h,t}$
  - Negative relations: $\mathcal{N}_{h,t}$
- Compare each r to the threshold class (NA), define their pairwise probabilities:

$$P_{h,t}(C=r|C=\{r,\mathrm{NA}\}) = P^r_{h,t}(r) = \frac{\exp(f_r)}{\exp(f_r)+\exp(f_\eta)}$$

$$P_{h,t}(C=\mathrm{NA}|C=\{r,\mathrm{NA}\}) = P^\eta_{h,t}(r) = 1 - P^r_{h,t}(r) = \frac{\exp(f_\eta)}{\exp(f_r)+\exp(f_\eta)}$$

- Our pairwise moving-threshold loss:

$$\downarrow\mathcal{L}^{h,t}_{pmt} = -\log\Big(\prod_{r\in\mathcal{P}_{h,t}} P^r_{h,t}(r) \prod_{r\in\mathcal{N}_{h,t}} (1-P^r_{h,t}(r))\Big)$$

$$= -\sum_{r\in\mathcal{P}_{h,t}} \log P^r_{h,t}(r) - \sum_{r\in\mathcal{N}_{h,t}} \log P^\eta_{h,t}(r)$$

$$= \sum_{r\in\mathcal{P}_{h,t}} \log(1+\exp(f_\eta - f_r)) \rightarrow f_r > f_\eta$$

For positive relations

$$+ \sum_{r\in\mathcal{N}_{h,t}} \log(1+\exp(f_r - f_\eta)) \rightarrow f_\eta > f_r$$

For negative relations

- The definition of information entropy:

$$H_{h,t}(r) = -P^r_{h,t}(r)\log P^r_{h,t}(r) - P^\eta_{h,t}(r)\log P^\eta_{h,t}(r)$$

- The regularization of entropy minimization:

$$\mathcal{L}^{h,t}_{em} = \frac{1}{\gamma_1}\sum_{r\in\mathcal{P}_{h,t}} H_{h,t}(r) + \frac{1}{\gamma_2}\sum_{r\in\mathcal{N}_{h,t}} H_{h,t}(r)$$

❖ **1st Loss function:**

$$\mathcal{L}_1 = \sum_{(h,t)\in\mathcal{B}} \mathcal{L}^{h,t}_{pmt} + \mathcal{L}^{h,t}_{em}$$

- ***Supervised contrastive learning for multi-labels & long-tailed relations***
- "Pull" the embeddings of similar examples together, and "push" dissimilar examples apart:

$$\downarrow\mathcal{L}^{h,t}_{scl} = -\log\left\{\frac{1}{|\mathcal{S}_{h,t}|}\sum_{p\in\mathcal{S}_{h,t}} \frac{\exp(\boxed{\boldsymbol{x}_{h,t}\cdot\boldsymbol{x}_p}/\tau)\uparrow}{\sum_{d\in\mathcal{B},d\neq(h,t)}\exp(\boxed{\boldsymbol{x}_{h,t}\cdot\boldsymbol{x}_d}/\tau)\downarrow}\right\}$$

pull

push

- Handling long-tailed relations → for entity pairs with empty positive examples in a batch:

$$\mathcal{L}^{h,t}_{lt} = \log\sum_{d\in\mathcal{B},d\neq(h,t)} \exp(\boxed{\boldsymbol{x}_{h,t}\cdot\boldsymbol{x}_d}/\tau)\downarrow$$

push

❖ **2nd Loss function:**

$$\mathcal{L}_2 = \sum_{(h,t)\in\mathcal{B}_{\mathcal{P}}} \mathbb{I}_{\{|\mathcal{S}_{h,t}|\neq 0\}}\mathcal{L}^{h,t}_{scl} + \mathbb{I}_{\{|\mathcal{S}_{h,t}|=0\}}\mathcal{L}^{h,t}_{lt}$$

❖ **Final loss function:**

$$\mathcal{L} = \mathcal{L}_1 + \lambda\mathcal{L}_2$$

Experiments and Analysis

- Benchmarks:
  - DocRED [Yao et al., ACL'19] & Re-DocRED [Tan et al., EMNLP'22] (both $|\mathcal{R}| = 96$).
- Two new data regimes
  - OOG-DocRE / OGG-DocRE:
    - Original labels for the train set
    - Original labels / Gold labels for the dev set
    - Gold labels for the test set

Dataset	Train #Doc / #Example	Dev #Doc / #Example	Test #Doc / #Example
DocRED	3,053 / 1,198,650	1,000 / 396,790	1,000 / 392,158
Re-DocRED	3,053 / 1,193,092	500 / 193,232	500 / 198,670
Our new data regimes			
OOG-DocRE	3,053 / 1,198,650	500[3] / 195,682	500 / 198,670
OGG-DocRE	3,053 / 1,198,650	500 / 193,232	500 / 198,670

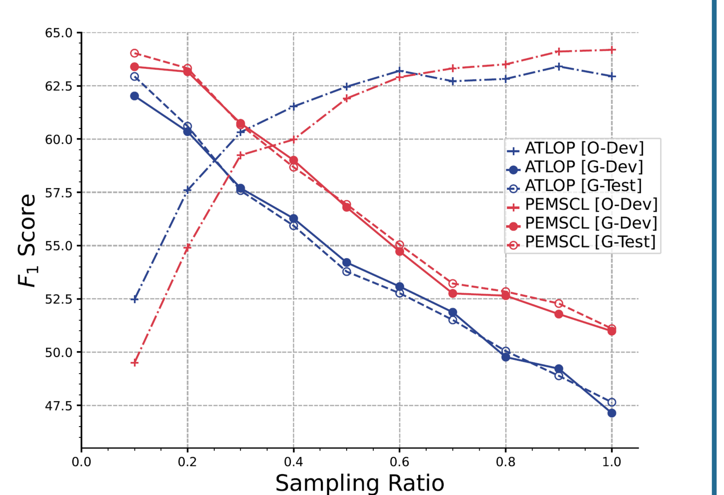
- Our PEMSCL outperforms previous strong baselines on the original DocRED dataset and its cleaned version, the Re-DocRED dataset.

Model	DocRED Dev Ign F_1	DocRED Dev F_1	DocRED Test Ign F_1	DocRED Test F_1
Implemented on DeBERTa$_{Large}$				
ATLOP (Zhou et al., 2021)	62.16±0.15	64.01±0.12	62.12	64.08
ATLOP + BCE (Zhou and Lee, 2022)	61.92±0.13	63.96±0.15	61.83	63.92
NCRL (Zhou and Lee, 2022)	<u>62.98±0.18</u>	<u>64.79±0.13</u>	<u>63.03</u>	<u>64.96</u>
PEMSCL (Ours)	**63.25±0.09**	**65.15±0.10**	**63.40**	**65.41**

Model	Re-DocRED Dev Ign F_1	Re-DocRED Dev F_1	Re-DocRED Test Ign F_1	Re-DocRED Test F_1
Implemented on RoBERTa$_{Large}$				
JEREX (Eberts and Ulges, 2021)	69.12	70.33	68.97	70.25
ATLOP + BCE* (Zhou and Lee, 2022)	75.86±0.13	75.25±0.11	75.91	75.36
ATLOP (Zhou et al., 2021)	76.88	77.63	76.94	77.73
DocuNet (Zhang et al., 2021)	77.53	78.16	77.27	77.92
KD-DocRE (Tan et al., 2022a)	77.92	78.65	77.63	78.35
NCRL* (Zhou and Lee, 2022)	<u>78.41±0.21</u>	<u>79.15±0.20</u>	<u>78.45</u>	<u>79.19</u>
PEMSCL (Ours)	**79.02±0.20**	**79.89±0.17**	**79.01**	**79.86**

- Our negative label sampling strategy is effective and robust in the noisy settings

	Orig-Dev Ign F_1	Orig-Dev F_1	Gold-Dev Ign F_1	Gold-Dev F_1	Gold-Test Ign F_1	Gold-Test F_1
On OOG-DocRE Regime						
ATLOP (Zhou et al., 2021)	60.94	62.95	46.99	47.14	47.52	47.65
NCRL (Zhou and Lee, 2022)	61.42	63.52	49.06	49.21	48.41	48.53
PEMSCL (Ours)	**62.05**	**64.19**	<u>50.82</u>	<u>50.99</u>	<u>50.92</u>	<u>51.10</u>
PEMSCL† (Ours)	46.07	49.51	**62.05**	**63.39**	**62.76**	**64.03**
On OGG-DocRE Regime						
ATLOP (Zhou et al., 2021)	-	-	48.23	48.54	48.50	48.77
NCRL (Zhou and Lee, 2022)	-	-	49.92	50.08	50.10	50.25
PEMSCL (Ours)	-	-	<u>50.43</u>	<u>50.62</u>	<u>51.09</u>	<u>51.25</u>
PEMSCL† (Ours)	-	-	**62.40**	**63.72**	**62.47**	**63.73**



- The logit difference between the relation and the threshold class in our PEMSCL is much lager than that of the ATLOP.
- Our PEMSCL can correctly predict relation that the ATLOP model fails to identify.