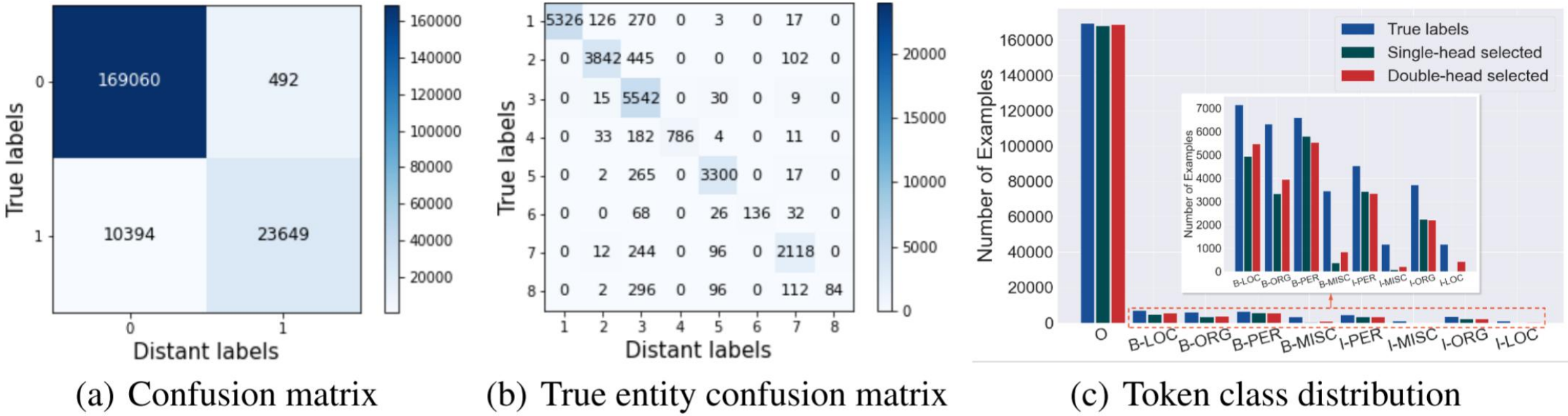
南京航空航天大學

Debiased and Denoised Entity Recognition from Distant Supervision

Haobo Wang1* , Yiwen Dong1* , Ruixuan Xiao1, Fei Huang2, Gang Chen1, Junbo Zhao1†
1Zhejiang University, Hangzhou, China
2Alibaba Group, Hangzhou, China

NeurIPS 2023

Introduction
(Dataset limitations & Methodological limitations)



(a) Confusion matrix (b) True entity confusion matrix (c) Token class distribution

(a) Confusion matrix of true labels and distant labels on the whole CoNLL03 dataset.
(b) The confusion matrix displays noise among true entity-type labels.
(c) The real token class distribution on the CoNLL03 dataset and tokens selected by a basic self-training framework with a single-head/double-head pathway. It can be shown that double-head selects more tokens than single-head on the minority entity classes such as MISC and LOC.

Self-training exists an **inherent confirmation bias** to assign erroneous pseudo-labels. Due to the existence of label noise, the DSNER task can face an amplified self-training bias, which, however, has never been touched on in prior studies.

Method (Notation and Preliminary)

南京航空航天大學

Named Entity Recognition.

$$\mathcal{D} = \{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^{M}$$

$$\boldsymbol{x} = [x_1, x_2, ..., x_n]$$

$$\boldsymbol{y} = [y_1, y_2, ..., y_n] \qquad y_i \in \mathcal{T} = \{0, 1, ..., K\}$$

1 ~ K denote entity types and 0 denotes non-entity

$$\mathcal{L}_{ce}(\boldsymbol{y}, f(\boldsymbol{x};\theta)) = -\frac{1}{n}\sum_{i=1}^{n} \log \text{softmax}(f_{i,y_i}(\boldsymbol{x};\theta)) \qquad f(\boldsymbol{x};\theta) = h \circ \phi(\boldsymbol{x})$$

predicted probability of x_i belonging to class y_i in sentence **x**

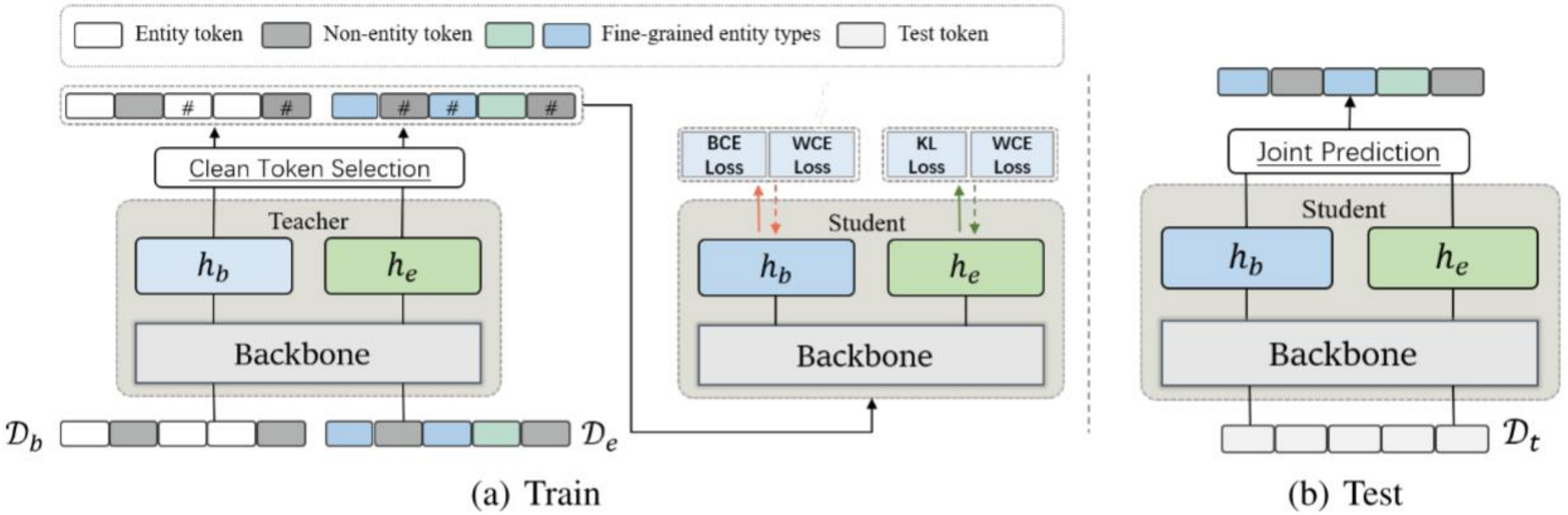
Distantly-Supervised NER.

separates the tokens in each sentence **x** into a set of potential clean tokens x^l and noisy tokens x^u

$$\mathcal{L}_{cls} = \mathcal{L}_L(\tilde{\boldsymbol{y}}^l, f(\boldsymbol{x}^l;\theta)) + \mathcal{L}_U(\hat{\boldsymbol{y}}^u, f(\boldsymbol{x}^u;\theta))$$

where L_L and L_U are classification losses on the two sets. $\hat{\boldsymbol{y}}^u$ is a pseudo-label generated via previous models.

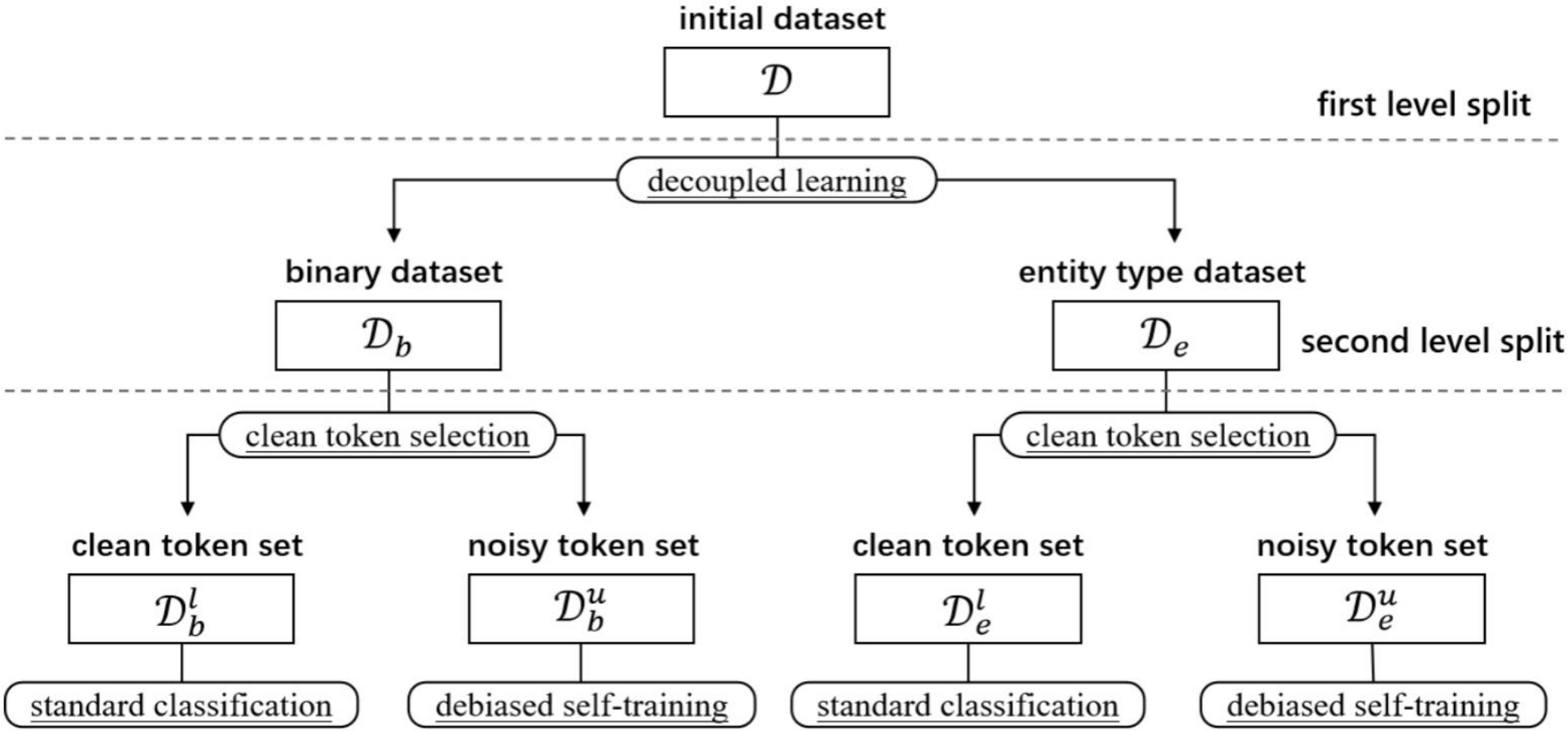
Method



(a): Illustration of decoupled learning paradigm and debiased self-training.
(b): Two classification heads are trained independently but make **joint** predictions when testing.
# denotes unselected noisy or invalid tokens.

$$\boldsymbol{p}_i = [1 - p_i^b, p_i^b * \boldsymbol{p}_i^e]$$

Method


initial dataset
$\mathcal{D}$
first level split
decoupled learning
binary dataset
$\mathcal{D}_b$
entity type dataset
$\mathcal{D}_e$
second level split
clean token selection
clean token selection
clean token set
$\mathcal{D}_b^l$
noisy token set
$\mathcal{D}_b^u$
clean token set
$\mathcal{D}_e^l$
noisy token set
$\mathcal{D}_e^u$
standard classification
debiased self-training
standard classification
debiased self-training

Method

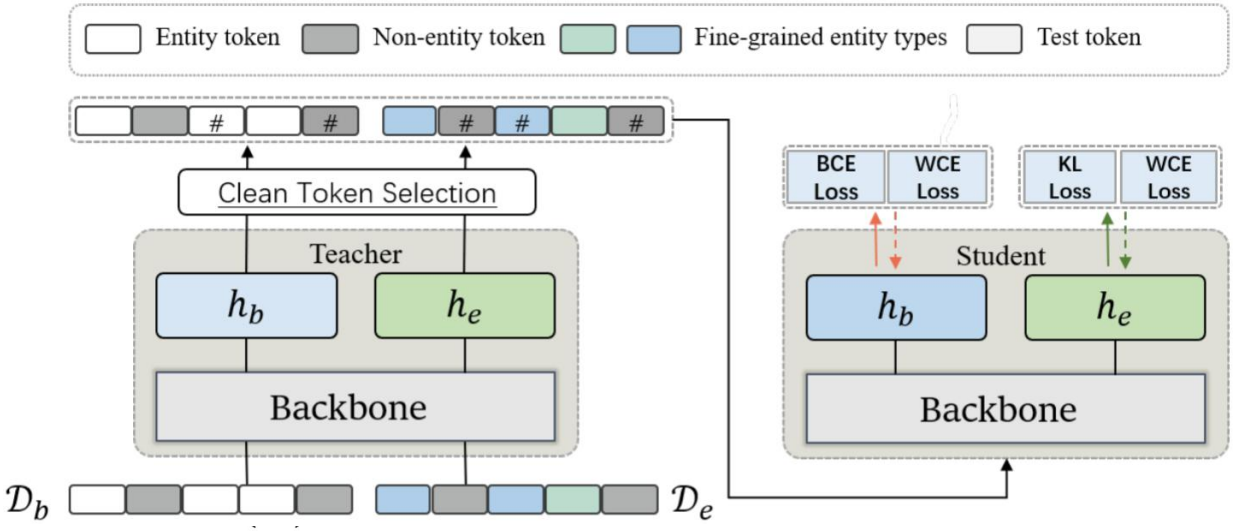
南京航空航天大學

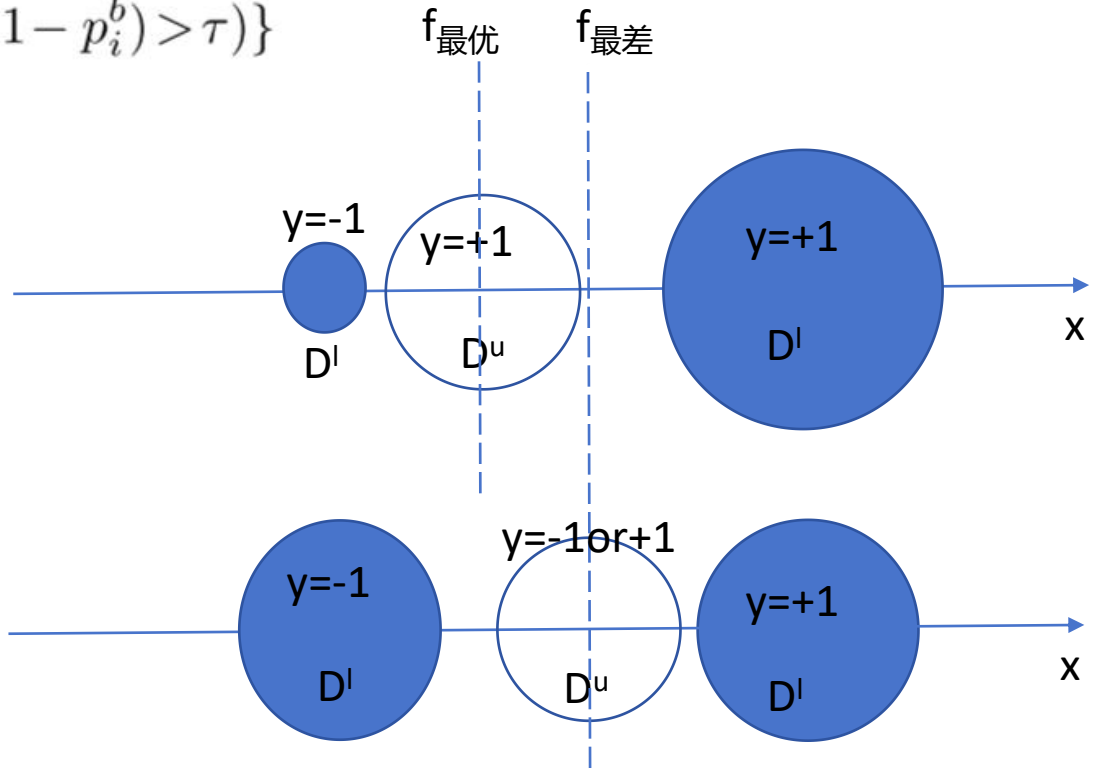
Clean Token Selection

In most of our experiments, we fix it to 0.95 without further tuning

$$\mathcal{D}_b^l = \{(x_i, \tilde{y}_i^b) | \mathbb{I}(\hat{y}_i^b = \tilde{y}_i^b) \wedge (\max(p_i^b, 1 - p_i^b) > \tau)\}$$

Debiased Self-Training





$$f^h = h \circ \phi$$

$$h_w = \arg\max_{h'} \mathcal{L}_U(\hat{\boldsymbol{y}}^u, f^{h'}(\boldsymbol{x}^u)) - \mathcal{L}_L(\tilde{\boldsymbol{y}}^l, f^{h'}(\boldsymbol{x}^l))$$ encoder ϕ is frozen at this step

$$\mathcal{L}_{wce}(\phi) = \mathcal{L}_U(\hat{\boldsymbol{y}}^u, f^{h_w}(\boldsymbol{x}^u)) - \mathcal{L}_L(\tilde{\boldsymbol{y}}^l, f^{h_w}(\boldsymbol{x}^l))$$ h_w is frozen and only update the encoder

Method

南京航空航天大學

Dual Co-guessing Mechanism

$f(\boldsymbol{x};\theta_1)$ $f(\boldsymbol{x};\theta_2)$ $\{(x_i,\hat{y}_i^{(1)})|\mathbb{I}(\hat{y}_i^{(1)}=\hat{y}_i^{(2)})\wedge(\max(\boldsymbol{p}_i^{(1)})>\tau)\wedge(\max(\boldsymbol{p}_i^{(2)})>\tau)\}$

Training loss

$$\mathcal{L}=\mathcal{L}_{b\_cls}+\mathcal{L}_{e\_cls}+w*\mathcal{L}_{e\_wce}$$

Post-hoc Entity Pathway Finetuning

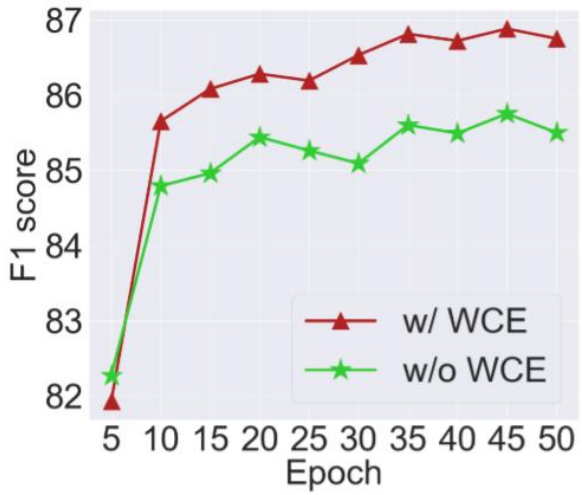
$$\mathcal{D}_e^l=\{(x_i,\hat{y}_i^e)|\mathbb{I}(\hat{y}_i^b=1)\wedge\mathbb{I}(\hat{y}_i^e=\tilde{y}_i^e)\}$$

Result

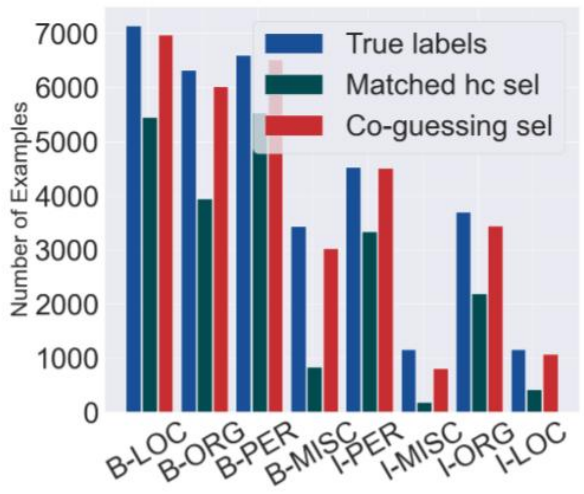
Table 1: Main results on five benchmark datasets measured by Precision (P), Recall (R), and F1 scores. We highlight the best overall performance for distant supervision in bold.

Methods	CoNLL03			OntoNotes5.0			Webpage			Wikigold			Twitter		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Fully-supervised methods															
BiLSTM-CC	91.35	91.06	91.21	85.99	86.36	86.17	50.07	54.76	52.34	55.40	54.30	54.90	60.01	46.16	52.18
RoBERTa-base	89.14	91.10	90.11	84.59	87.88	86.20	66.29	79.73	72.39	85.33	87.56	86.43	51.76	52.63	52.19
Distantly-supervised methods															
AutoNER	75.21	60.40	67.00	64.63	69.95	67.18	48.82	54.23	51.39	43.54	52.35	47.54	43.26	18.69	26.10
LRNT	79.91	61.87	69.74	67.36	68.02	67.69	46.70	48.83	47.74	45.60	46.84	46.21	46.94	15.98	23.84
Co-teaching+	86.04	68.74	76.42	66.63	69.32	67.95	61.65	55.41	58.36	55.23	49.26	52.08	51.67	42.66	46.73
JoCoR	83.65	69.69	76.04	66.74	68.74	67.73	62.14	58.78	60.42	51.48	51.23	51.35	49.40	45.59	47.42
NegSampling	80.17	77.72	78.93	64.59	**72.39**	68.26	70.16	58.78	63.97	49.49	55.35	52.26	50.25	44.95	47.45
BOND	82.05	80.92	81.48	67.14	69.61	68.35	67.37	64.19	65.74	53.44	68.58	60.07	53.16	43.76	48.01
SCDL	**87.96**	79.82	83.69	**67.49**	69.77	68.61	68.71	68.24	68.47	62.25	66.12	64.13	**59.87**	44.57	51.09
Ours*	86.23	87.28	86.75	66.38	72.08	69.11	71.52	**72.97**	72.24	60.77	68.10	64.23	56.44	**48.38**	52.10
Ours*(finetune)	86.41	**87.49**	**86.95**	66.63	71.92	**69.17**	**72.48**	**72.97**	**72.73**	**62.87**	**69.42**	**65.99**	57.65	47.80	**52.26**

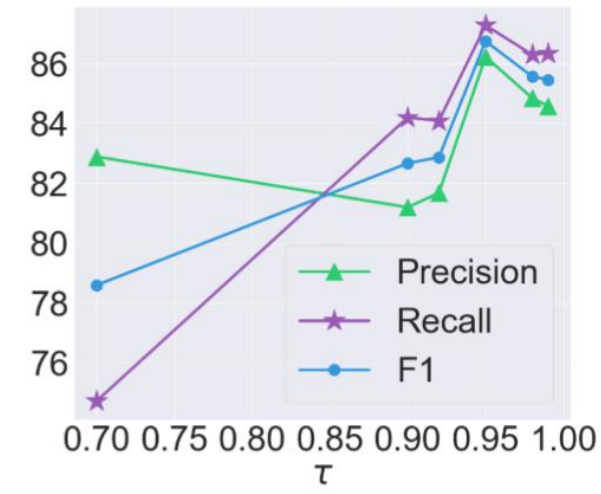
Ablation Study



(a) Analysis on WCE

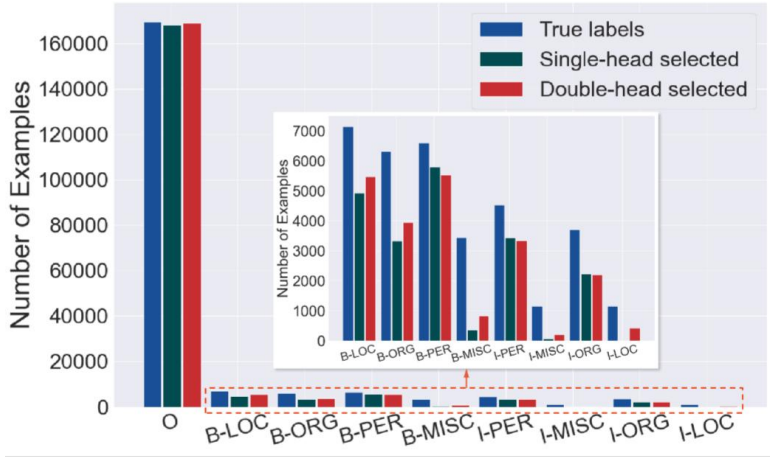


(b) Analysis on Co-Guessing



(c) Analysis on τ

Figure 4: (a) The F1 curves of DesERT with/without WCE loss. (b) The distribution of the true labels and selected labels with/without co-guessing. (c) The parameter study of different confidence thresholds τ on the CoNLL03 dataset.

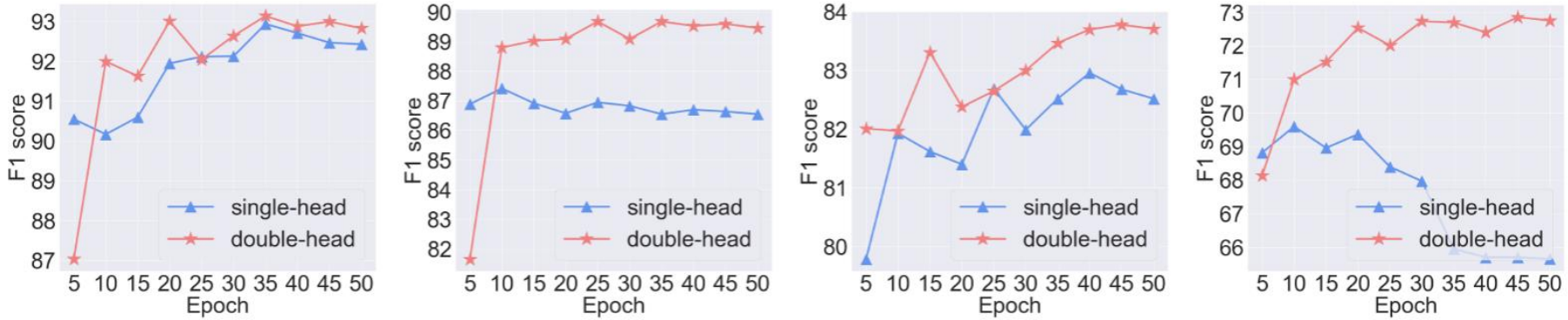


(c) Token class distribution

Further Analysis

南京航空航天大學

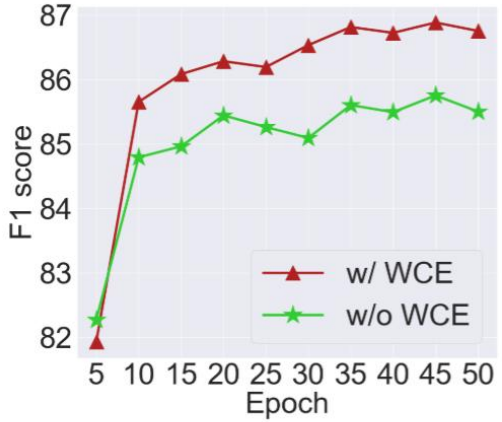
Efficacy of decoupled learning



(a) PER-F1 (b) LOC-F1 (c) ORG-F1 (d) MISC-F1

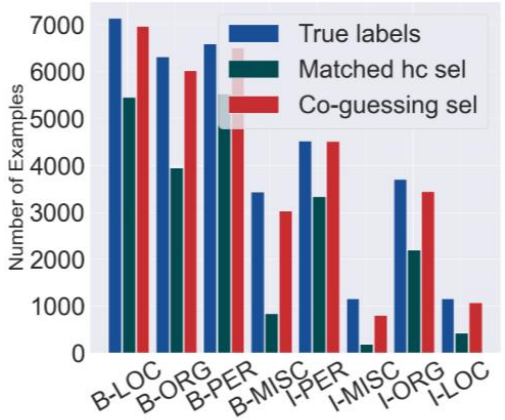
Figure 3: The F1-scores of DesERT with/without double-head pathway on four entity types, which have 11.1k/8.3k/10.0k/4.6k tokens respectively (from left to right).

Efficacy of debiased self-training



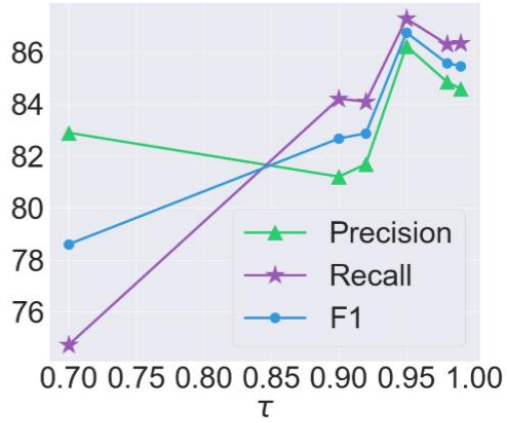
(a) Analysis on WCE

Efficacy of co-guessing



(b) Analysis on Co-Guessing

Efficacy of threshold parameter τ



(c) Analysis on τ

Distant Supervision from Large Language Models

Large language models (LLMs), including GPT-3 , ChatGPT, and GPT-43, have largely revolutionized the NLP landscape. Thanks to their emerging abilities like **in-context learning** (ICL) and **chain-of-thought**, LLMs demonstrate remarkable **zero-shot learning** performance in a wide range of downstream NLP tasks.
However, LLMs are still **legs behind** the fine-tuned small language models in many NLP applications including NER.

To deal with this problem, we extend the DSNER formulation and design a novel in-context learning algorithm that exploits self-generated text-tag pairs to generate distant labels. Moreover, we **modify** our original algorithms to **fully use hybrid labels** including ChatGPT-generated labels and original knowledge-base generated labels (KB labels).

Table 4: Performance of DesERT with different sources of distant labels on CoNLL03.

Supervision	**Unsupervised**		**ChatGPT Labels**		**KB Labels**		**Hybrid Labels**	
Model	ChatGPT	ChatGPT-A	SCDL	DesERT	SCDL	DesERT	SCDL*	DesERT*
Precision	68.95	79.11	68.39	81.91	**87.96**	86.23	83.87	87.24
Recall	64.16	63.13	72.74	77.38	79.82	87.28	85.50	**88.93**
F1	66.47	70.22	70.50	79.58	83.69	86.75	84.67	**88.08**

Conclusion

两种偏差：**高度结构化的噪声**和自训练框架引入的**固有偏差**

解决方法：**解耦学习**、**清洁令牌选择**、**去偏自训练**和**双共猜测机制**

实验结果：结果表明 DesERT 取得了最先进的性能。建立了基于 ChatGPT 的远程监督 NER **新基准**，DesERT 在该基准上同样表现出色。

Thanks