CIKM2023

Task Relation Distillation and Prototypical Pseudo Label for Incremental Named Entity Recognition

Duzhen Zhang, Hongliu Li, Wei Cong, Rongtao Xu, Jiahua Dong and Xiuyi Chen

Reporter: Duzhen Zhang

Baidu Inc, Beijing, China

Sunday, October 22, 2023

Background

- Named Entity Recognition (NER) aims to annotate each token in a sentence with predefined sets of entity types or the non-entity type.

- The traditional NER paradigm annotates tokens with a fixed set of entity types, and the NER model learns this in one go.

- In a more realistic scenario, NER models need to continuously identify newly emerging entity types without the need for retraining from scratch. This is known as Incremental Named Entity Recognition (INER).

- For instance, Siri voice assistant is often required to extract new entity types (such as genres, actors) to comprehend new user intents (e.g., retrieving movie information).

INER Task Definition

- INER aims to gradually train a model through a series of steps, denoted as t=1,...,T, learning an expanding set of entity types.

- At each step, there exists a corresponding training set D_t, containing several pairs (X^t, Y^t), where X^t represents the input token sequence and Y^t represents the corresponding label sequence.

- Y^t contains labels only from the current entity type set E^t, while all other labels (possible old entity types $E^{1:t-1}$or future entity types $E^{t+1:T}$) are masked as non-entity type e_o.

- Learning objective: In the t-th step (t>1), given the old model M_{t-1} and the current training set D_t, train a new model M_t that can identify all entity types up to that point, denoted as $E^{1:t}$.

Challenges

- Common issues in incremental learning: catastrophic forgetting.

- Specific issue in INER: semantic drift of non-entity types.

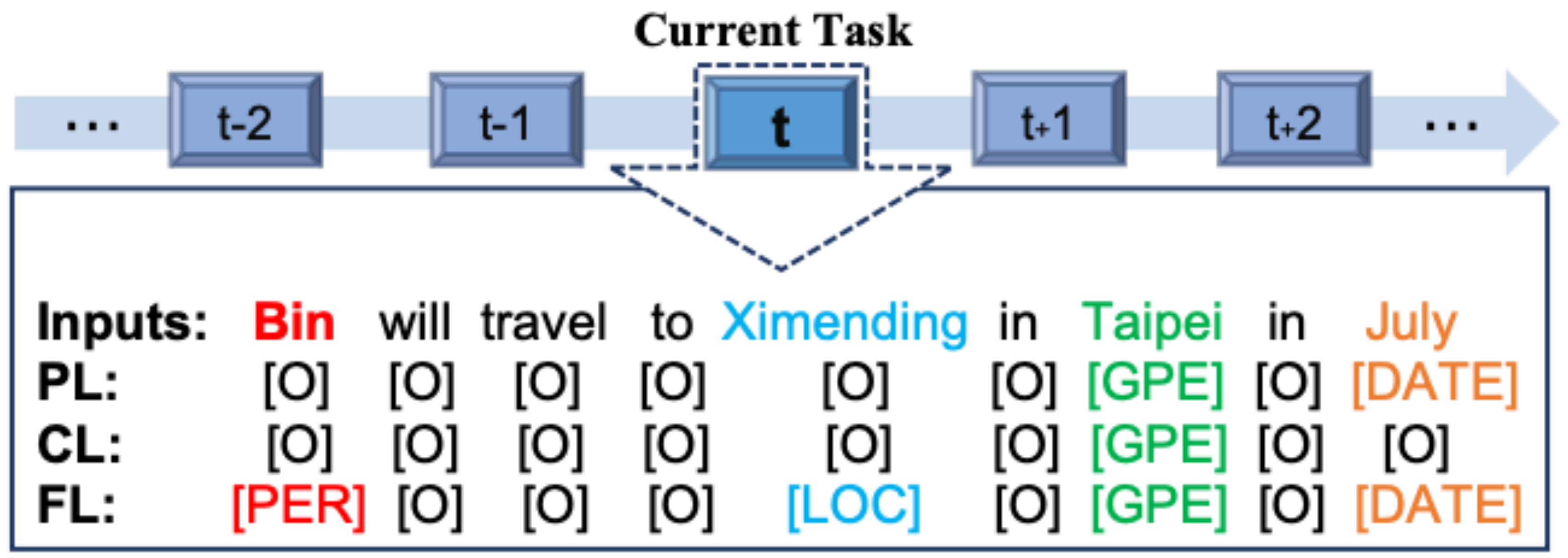

Figure 1: A simplified INER example, where PL, CL, and FL denote Predicted Labels of the current model, Current ground-truth Labels, and Full ground-truth Labels, respectively. Old entity types (*e.g.*, [PER] (Person), [DATE] (Date)) and future entity type (*e.g.*, [LOC] (Location)) are labeled as non-entity type ([O]) in the current task *t* where [GPE] (Countries, Cities, or States) is the current entity type being learned, leading to background shift (the third row CL). Furthermore, the NER model incrementally learns new entity types without accessing previous samples, suffering from catastrophic forgetting of old entity types (e.g., the model forgets old entity types [PER]) (the second row PL).

Existing Work

Existing INER methods typically use knowledge distillation to retain the predicted logits, preventing significant changes in model weights.

- ExtendNER AAAI2021[1]:

Distills the predicted logits of the old model to encourage the new model to produce results similar to those generated by the old model.

- L&R ACL2022[2]:

Adopts a two-stage learn-and-review (L&R) framework for INER.

The learning stage is similar to ExtendNER, while the review stage synthesizes samples of old entity types to augment the current dataset.

- CFNER EMNLP2022[3] (SOTA):

Combines ExtendNER with a causal inference framework. Distills causal effects from non-entity types.

[1] Monaikul N, Castellucci G, Filice S, et al. Continual learning for named entity recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13570-13577.
[2] Xia Y, Wang Q, Lyu Y, et al. Learn and review: Enhancing continual named entity recognition via reviewing synthetic samples[C]//Findings of the Association for Computational Linguistics: ACL 2022. 2022: 2291-2300.
[3] Zheng J, Liang Z, Chen H, et al. Distilling Causal Effect from Miscellaneous Other-Class for Continual Named Entity Recognition[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 3602-3615.

Existing Work's Shortcomings:

- The designed logits distillation did not adequately consider the trade-off between stability and plasticity.

- Only general forgetting issues were considered, without addressing the specific INER problem, such as the semantic drift of non-entity types.

Our Contributions

- We propose a task relation distillation scheme to consider task relationships in different incremental learning tasks, mitigating the catastrophic forgetting problem by constituting a suitable trade-off between stability and plasticity.

- We introduce a prototypical pseudo label strategy to utilize the old entity type information contained in the non-entity type, better tackling the semantic shift problem by correcting the prediction error of the old model and producing high-quality pseudo labels.

- We conduct extensive experiments on ten INER settings of three benchmark datasets (i.e., CoNLL2003, I2B2, and OntoNotes5). The results demonstrate that our RDP achieves significant improvements over the previous State-Of-The-Art (SOTA) method CFNER, with an average gain of 6.08% in Micro F1 scores and 7.71% in Macro F1 scores.

Method Overview

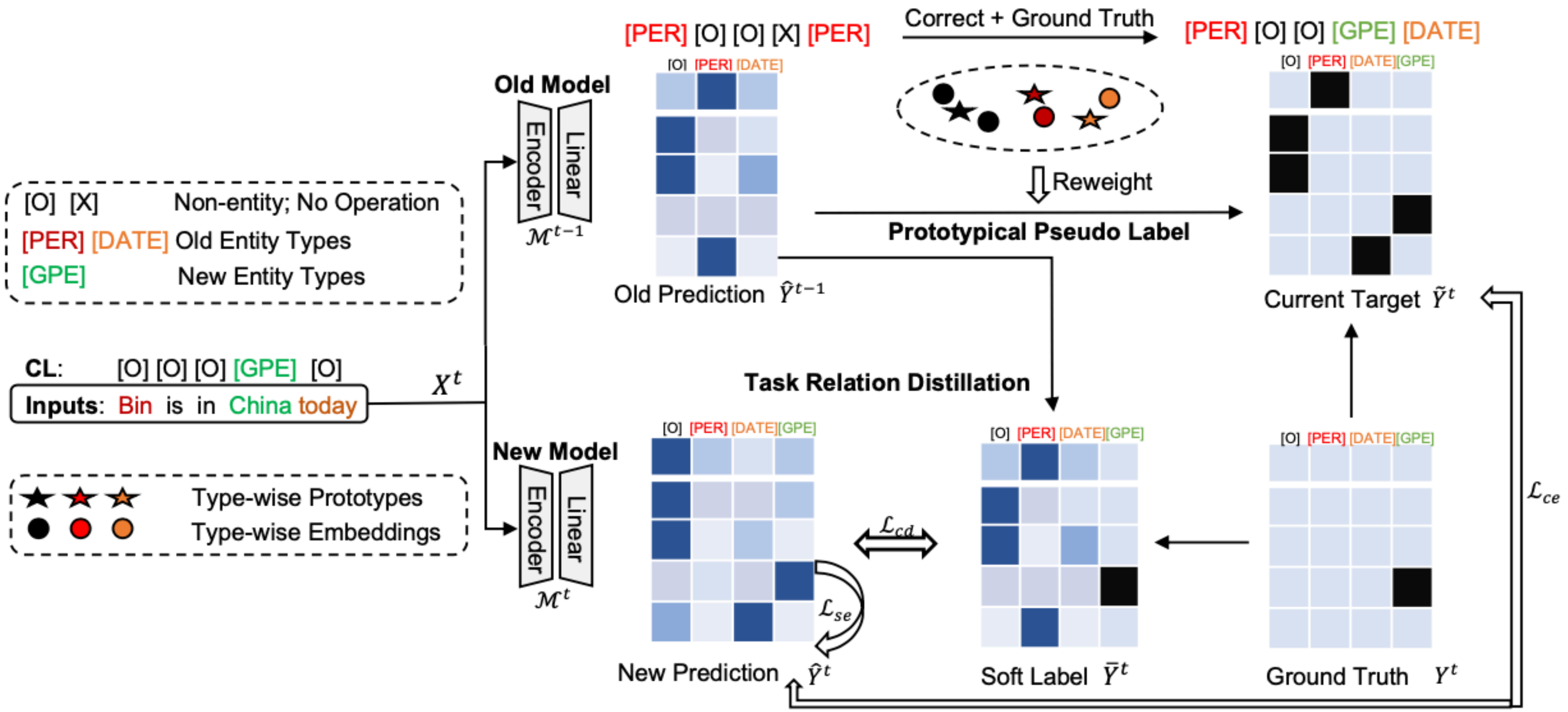

Figure 2: The overall framework of our RDP, demonstrated by a simplified INER example. CL denotes the current ground-truth labels. For a current input token sequence X^t, the soft label $\overline{Y}^t$ is calculated by combining the old prediction $\widehat{Y}^{t-1}$ with the current ont-hot ground truth Y^t. The current target $\widetilde{Y}^t$ is obtained by the prototypical pseudo label strategy. Then, we update the new model $\mathcal{M}^t$ with the task relation distillation loss (*e.g.*, $\mathcal{L}_{\text{cd}}$ and $\mathcal{L}_{\text{se}}$) and pseudo label based cross entropy loss (*i.e.*, $\mathcal{L}_{\text{ce}}$).

Method Details

- We propose an effective method called task Relation Distillation and Prototypical pseudo label (RDP) for INER.
- Firstly, we introduce a task relation distillation scheme that considers task relationships to mitigate catastrophic forgetting. This scheme comprises two components: an inter-task relation distillation loss and an intra-task self-entropy loss, striking a balance between stability and plasticity.
    - The inter-task relation distillation loss transfers knowledge from soft labels to the current model's output probabilities. These soft labels are constructed by combining the one-hot ground truth and the output probabilities of the old model, which helps capture the inter-task semantic relations between old tasks and between old and new tasks by smoothing the one-hot ground truth.
    - Moreover, the intra-task self-entropy loss enhances the confidence of the current predictions by minimizing self-entropy.
- Secondly, we develop a prototypical pseudo label strategy to explicitly retrieve old entity types within the current non-entity type for classification, effectively overcoming the semantic shift.
    - To correct mistaken labels predicted by the old model and produce high-quality pseudo labels, it exploits the distances between token embeddings and type-wise prototypes to reweight the output probabilities of the old model.

Method Details

- Task relation distillation scheme
  - The inter-task relation distillation loss

$$\mathcal{L}_{\text{cd}}(\Theta^t) = -\frac{1}{|X^t|}\sum_{i=1}^{|X^t|} \overline{Y}^t(i) \log \widehat{Y}^t(i),$$

  - The intra-task self-entropy loss

$$\mathcal{L}_{\text{se}}(\Theta^t) = -\frac{1}{|X^t|}\sum_{i=1}^{|X^t|} \widehat{Y}^t(i) \log \widehat{Y}^t(i),$$

- Prototypical pseudo label strategy

$$\widetilde{Y}^t(i,e) = \begin{cases} 1 \text{ if } Y^t(i,e_o) = 0 \text{ and } e = \operatorname*{argmax}_{e' \in \mathcal{E}^t} Y^t(i,e') \\ 1 \text{ if } Y^t(i,e_o) = 1 \text{ and } e = \operatorname*{argmax}_{e' \in e_o \cup \mathcal{E}^{1:t-1}} \omega^t(i,e')\widehat{Y}^{t-1}(i,e') \\ 0 \text{ otherwise} \end{cases}$$

$$\omega^t(i,e) = \frac{\exp(-||F^{t-1}(X^t(i)) - \eta^e||/\tau)}{\sum_e \exp(-||F^{t-1}(X^t(i)) - \eta^e||/\tau)},$$

$$\eta^e = \frac{\sum \left[F^{t-1}(X^t(i)) * \mathbb{I}\left(e == \operatorname*{argmax}_{e' \in e_o \cup \mathcal{E}^{1:t-1}} \widehat{Y}^{t-1}(i,e')\right)\right]}{\sum \mathbb{I}\left(e == \operatorname*{argmax}_{e' \in e_o \cup \mathcal{E}^{1:t-1}} \widehat{Y}^{t-1}(i,e')\right)},$$

Experimental Setting

- Datasets

Table 3: The statistics for each dataset.

	# Entity Type	# Sample	Entity Type Sequence (Alphabetical Order)
CoNLL2003	4	21k	LOCATION, MISC, ORGANISATION, PERSON
I2B2	16	141k	AGE, CITY, COUNTRY, DATE, DOCTOR, HOSPITAL, IDNUM, MEDICALRECORD, ORGANIZATION, PATIENT, PHONE, PROFESSION, STATE, STREET, USERNAME, ZIP
OntoNotes5	18	77k	CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK_OF_ART

- Partition the training set into disjoint slides, where each slide corresponds to a different incremental learning step.
- In each slide, retain labels only for the entity types to be learned, while masking the other labels as non-entity types.

- INER Settings

• CoNLL2003	FG-1-PG-1	FG-2-PG-1		
• I2B2	FG-1-PG-1	FG-2-PG-2	FG-8-PG-1	FG-8-PG-2
• OntoNotes5	FG-1-PG-1	FG-2-PG-2	FG-8-PG-1	FG-8-PG-2

Experimental Setting

- Evaluation Metrics
  - Consideration was given to the issue of imbalanced entity types in NER, utilizing Micro F1 and Macro F1 scores to assess model performance.
  - A line plot for step-wise performance comparison was created.
  - The final performance is the average result across all steps, including the first step.

Result Analysis

Main Results

- As depicted in the upper part of Table, our RDP achieves improvements over the previous SOTA baseline CFNER ranging from 4.87% to 17.71% in Micro-F1, and 1.77% to 25.69% in Macro-F1, under four INER settings (FG-1-PG-1, FG-2-PG-2, FG-8-PG-1, and FG-8-PG-2) of the I2B2 dataset.

- Similarly, in the lower part of Table, our RDP achieves improvements over the previous SOTA baseline CFNER ranging from 0.97% to 9.34% in Micro-F1, and 1.77% to 11.34% in Macro-F1, under four INER settings (FG-1-PG-1, FG-2-PG-2, FG-8-PG-1, and FG-8-PG-2) of the OntoNotes5 dataset.

Table 4: Comparisons with baselines on the I2B2 [36] and OntoNotes5 [15] datasets. The red denotes the highest result, and the blue denotes the second highest result. The marker † refers to significant test *p-value*<0.05 comparing with CFNER [55]. * represents results from our re-implementation. Other baseline results are directly cited from CFNER [55].

Dataset	Baseline	FG-1-PG-1		FG-2-PG-2		FG-8-PG-1		FG-8-PG-2	
		Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
I2B2 [36]	Only Finetuning	17.43±0.54	13.81±1.14	28.57±0.26	21.43±0.41	20.83±1.78	18.11±1.66	23.60±0.15	23.54±0.38
	PODNet [8]	12.31±0.35	17.14±1.03	34.67±2.65	24.62±1.76	39.26±1.38	27.23±0.93	36.22±12.9	26.08±7.42
	LUCIR [14]	43.86±2.43	31.31±1.62	64.32±0.76	43.53±0.59	57.86±0.87	33.04±0.39	68.54±0.27	46.94±0.63
	Self-Training [41]	31.98±2.12	14.76±1.31	55.44±4.78	33.38±3.13	49.51±1.35	23.77±1.01	48.94±6.78	29.00±3.04
	ExtendNER* [35]	41.65±10.11	23.11±2.70	67.60±1.15	42.58±1.59	45.14±2.91	27.41±0.88	56.48±2.41	38.88±1.38
	ExtendNER [35]	42.85±2.86	24.05±1.35	57.01±4.14	35.29±3.38	43.95±2.01	23.12±1.79	52.25±5.36	30.93±2.77
	CFNER* [55]	**64.79±0.26**	**37.79±0.65**	**72.58±0.59**	**51.71±0.84**	56.66±3.22	36.84±1.35	**69.12±0.94**	**51.61±0.87**
	CFNER [55]	62.73±3.62	36.26±2.24	71.98±0.50	49.09±1.38	**59.79±1.70**	**37.30±1.15**	69.07±0.89	51.09±1.05
	RDP (Ours)	**71.39±1.01**†	**44.00±2.31**†	**77.45±0.55**†	**53.48±0.66**†	**77.50±1.26**†	**62.99±0.36**†	**80.08±0.40**†	**63.72±0.71**†
	Imp.	**⇑6.60**	**⇑6.21**	**⇑4.87**	**⇑1.77**	**⇑17.71**	**⇑25.69**	**⇑10.96**	**⇑12.11**
OntoNotes5 [15]	Only Finetuning	15.27±0.26	10.85±1.11	25.85±0.11	20.55±0.24	17.63±0.57	12.23±1.08	29.81±0.12	20.05±0.16
	PODNet [8]	9.06±0.56	8.36±0.57	34.67±1.08	24.62±0.85	29.00±0.86	20.54±0.91	37.38±0.26	25.85±0.29
	LUCIR [14]	28.18±1.15	21.11±0.84	64.32±1.79	43.53±1.11	66.46±0.46	46.29±0.38	76.17±0.09	55.58±0.55
	Self-Training [41]	50.71±0.79	33.24±1.06	68.93±1.67	50.63±1.66	73.59±0.66	49.41±0.77	77.07±0.62	53.32±0.63
	ExtendNER* [35]	51.36±0.77	33.38±0.98	63.03±9.39	47.64±5.15	73.65±0.19	50.55±0.56	77.86±0.10	55.21±0.51
	ExtendNER [35]	50.53±0.86	32.84±0.84	67.61±1.53	49.26±1.49	73.12±0.93	49.55±0.90	76.85±0.77	54.37±0.57
	CFNER* [55]	58.44±0.71	41.75±1.51	72.10±0.31	55.02±0.35	78.25±0.33	**58.64±0.42**	80.09±0.37	**61.06±0.37**
	CFNER [55]	**58.94±0.57**	**42.22±1.10**	**72.59±0.48**	**55.96±0.69**	**78.92±0.58**	57.51±1.32	**80.68±0.25**	60.52±0.84
	RDP (Ours)	**68.28±1.09**†	**53.56±0.39**†	**74.38±0.26**†	**57.73±0.54**†	**79.89±0.20**†	**63.20±0.58**†	**83.30±0.30**†	**66.92±1.26**†
	Imp.	**⇑9.34**	**⇑11.34**	**⇑1.79**	**⇑1.77**	**⇑0.97**	**⇑4.56**	**⇑2.62**	**⇑5.86**

Result Analysis

Main Results

- As illustrated in Figure, our RDP outperforms the INER baselines in task-wise Micro-F1 comparisons across the eight settings of the I2B2 and OntoNotes5 datasets.

- These results quantitatively confirm the superiority and effectiveness of our RDP compared to competitive baselines, showcasing its ability to learn a robust INER model and indicating improved resilience to catastrophic forgetting and background shift problems.

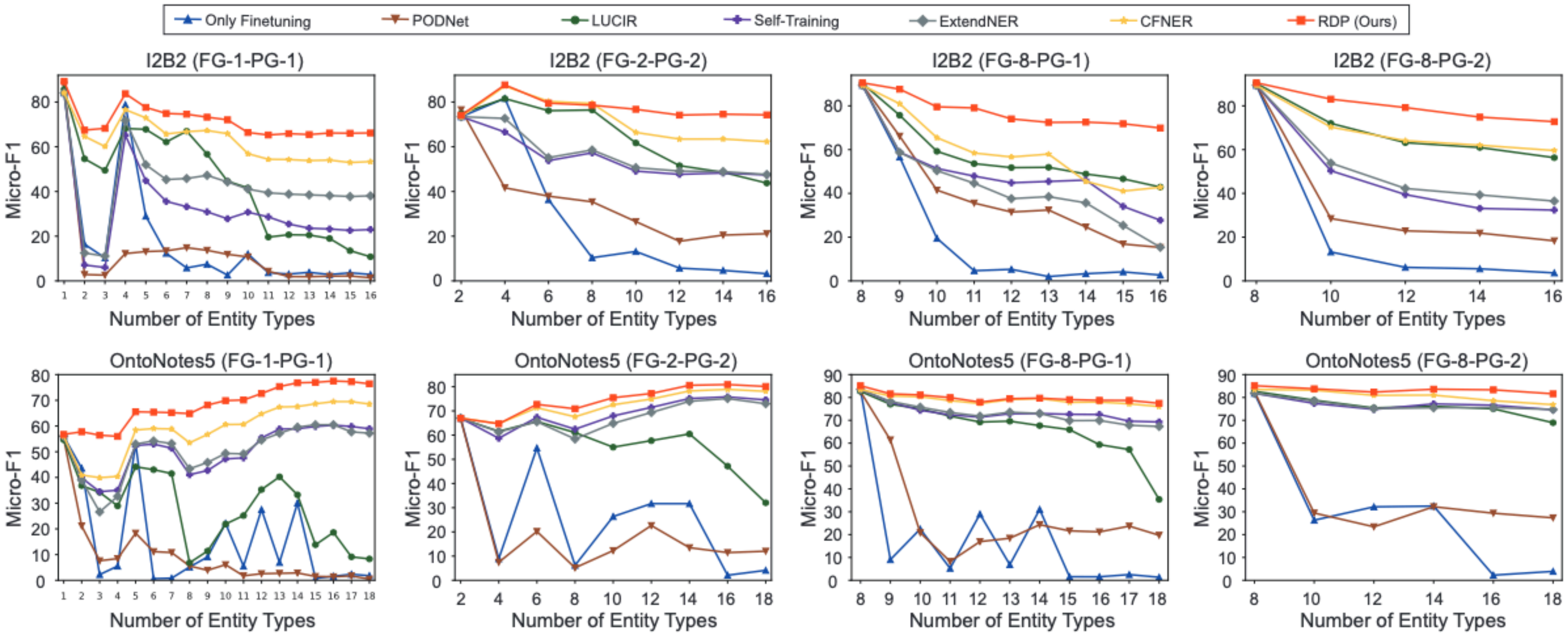

Figure 3: Comparison of the task-wise Micro-F1 on I2B2 [36] and OntoNotes5 [15]. Results of baselines are from CFNER [55].

Result Analysis

Ablation Study

- We conducted ablation studies to analyze the effects of critical components in our RDP, as presented in Table.

Table 5: The ablation study of our RDP under the FG-1-PG-1 setting of the I2B2 [36] and OntoNotes5 [15] datasets. Compared with our RDP, all ablation variants significantly degrade INER performance, verifying the importance of all components to address INER collaboratively.

Method	I2B2		OntoNotes5	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
RDP (Ours)	**71.39±1.01**	**44.00±2.31**	**68.28±1.09**	**53.56±0.39**
w/o $\mathcal{L}_{cd}$	64.97±0.55	38.76±1.01	63.56±0.37	47.49±1.36
w/o $\mathcal{L}_{se}$	67.59±1.42	41.32±2.66	65.47±0.43	50.27±0.59
w/o PPL	64.17±1.19	39.86±2.03	64.09±0.57	46.09±0.80
w/o PL	48.93±0.69	31.66±0.64	56.64±0.45	39.54±1.05

Result Analysis

Case Study

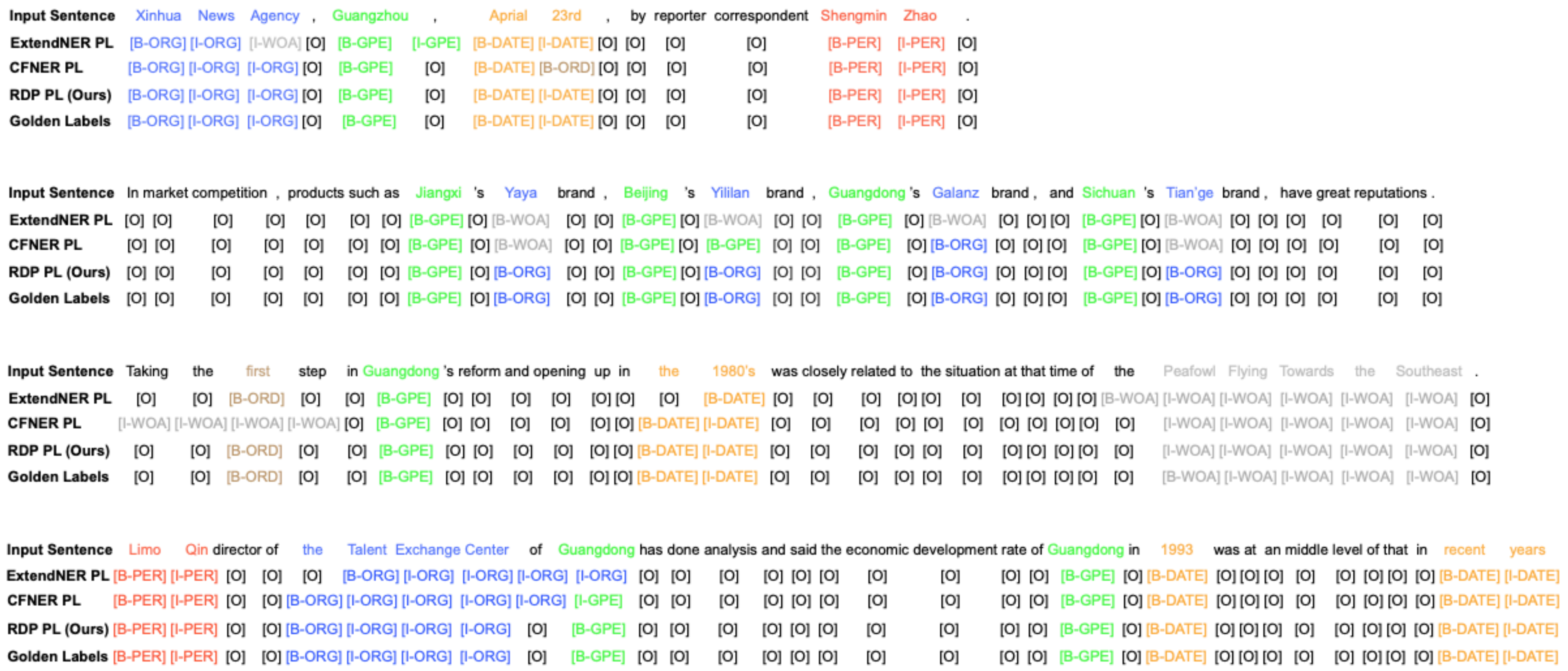

Figure 4: Four real NER cases sampled from the OntoNotes5 [15] test set. PL denotes the predicted labels. B- and I- distinguish begin/inside of named entities. [O], [DATE], [GPE], [ORD], [ORG], [PER], and [WOA] denote non-entity type, `Date`, `Countries, Cities, or States`, `Ordinals`, `Organization`, `Person`, and `Work of art`, respectively. All the prediction results are from the last task of the FG-8-PG-2 setting. These visualization NER cases qualitatively demonstrate the superiority and effectiveness of our proposed RDP method.

Conclusion

- In this paper, we present the RDP method as a solution to address the challenges of catastrophic forgetting and background shift in INER.
- We begin by introducing a task relation distillation scheme to explore the semantic relations between old and new tasks, leading to a suitable trade-off between stability and plasticity for INER and, ultimately, mitigating catastrophic forgetting.
- Additionally, we propose a prototypical pseudo label strategy to label old entity types contained in the non-entity type, effectively tackling the background shift problem by correcting the prediction error from the old model.
- We conduct extensive experiments on ten INER settings of three datasets: CoNLL2003, I2B2, and OntoNotes5. The results clearly show the superiority of our RDP method, outperforming previous SOTA methods by a significant margin. Our method offers a promising direction for advancing INER techniques and overcoming the challenges posed by incremental learning scenarios.

Thanks !