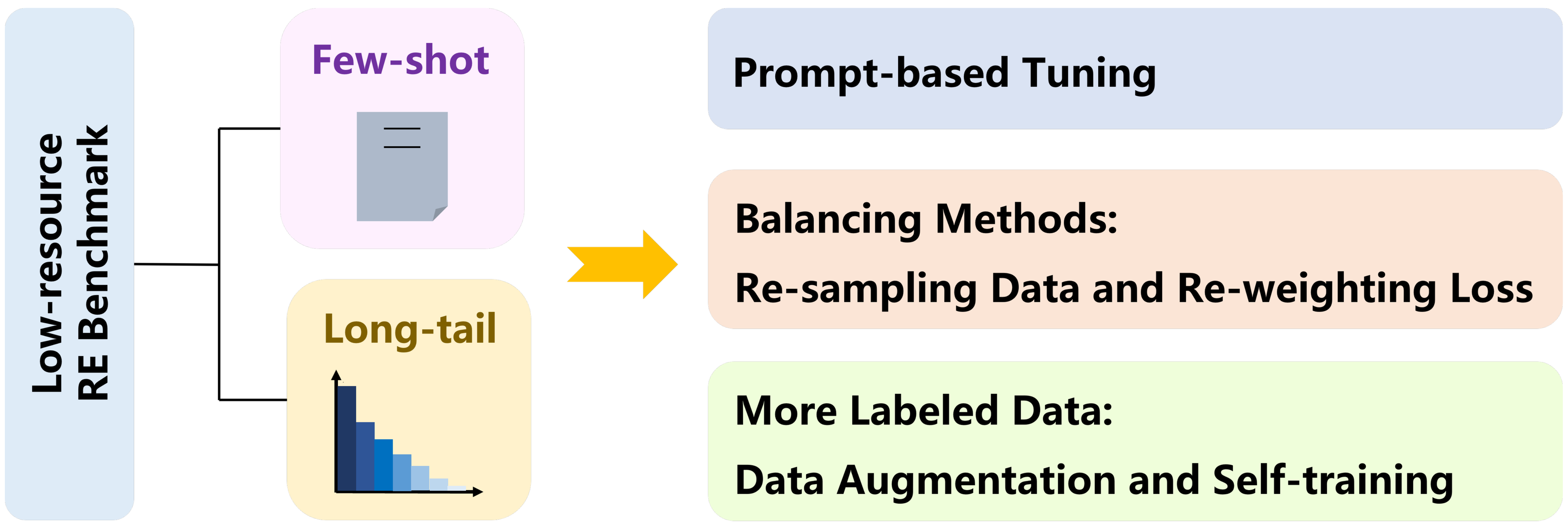
Towards Realistic Low-resource Relation Extraction: A Benchmark with Empirical Baseline Study

Xin Xu[1,2*], Xiang Chen[1,2*], Ningyu Zhang[1,2], Xin Xie[1,2], Xi Chen[3], Huajun Chen[1,2]
[1]Zhejiang University & AZFT Joint Lab for Knowledge Engine
[2]Hangzhou Innovation Center, Zhejiang University, [3]Tencent



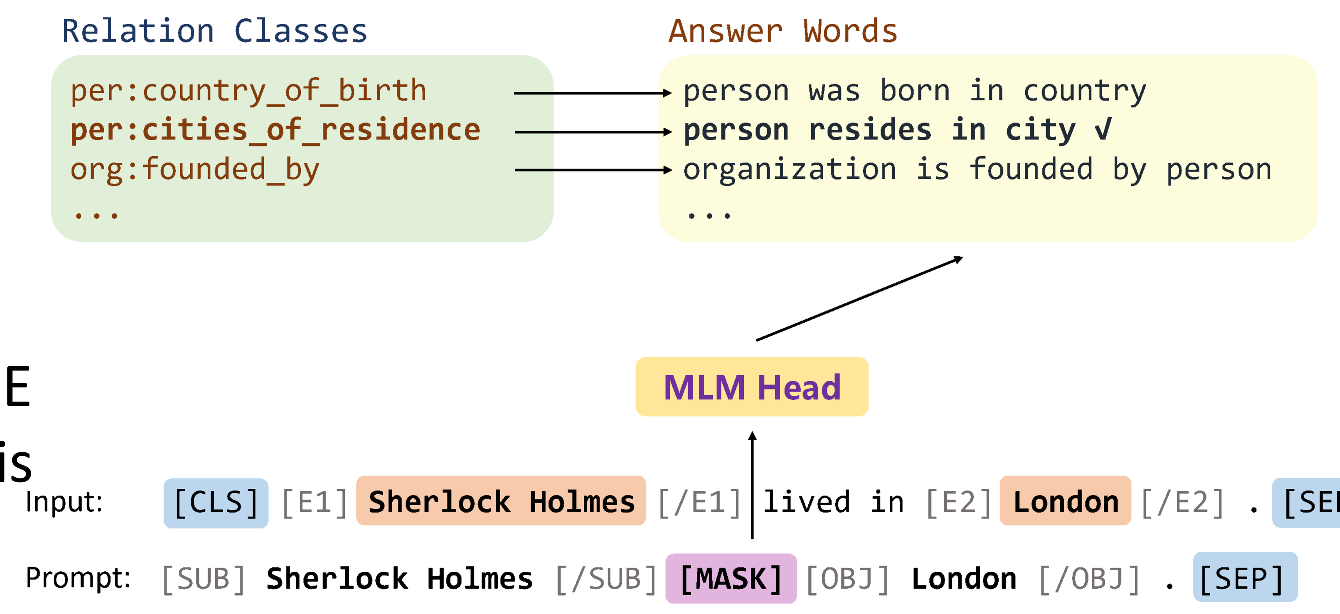
LREBench (https://github.com/zjunlp/LREBench)

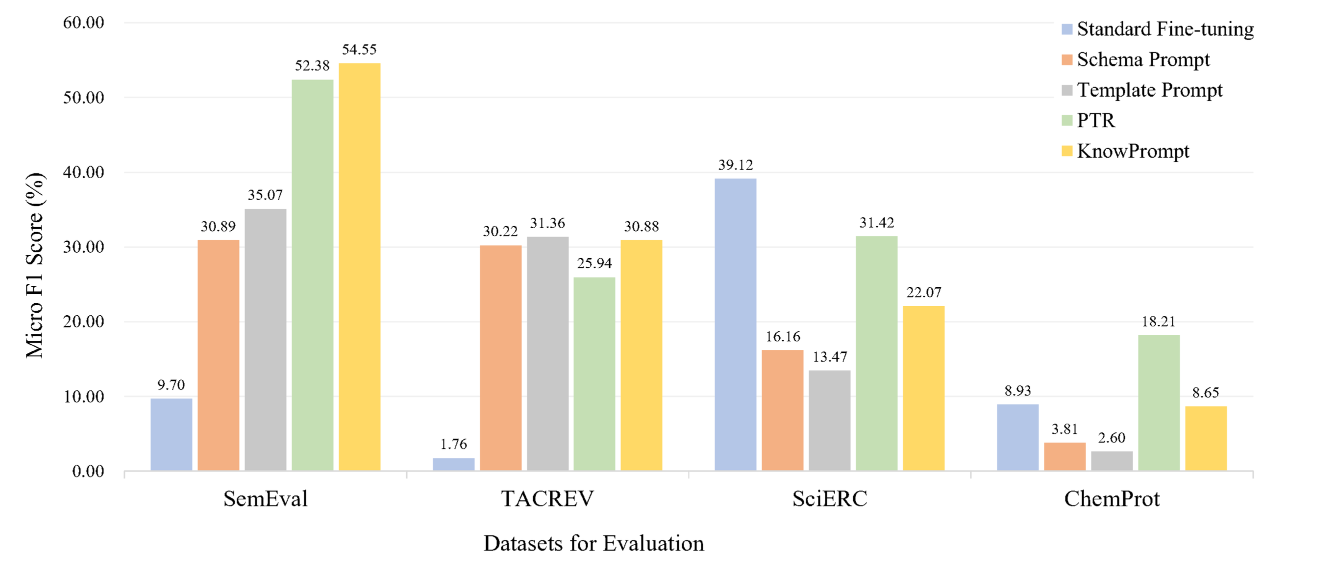
This paper presents an empirical study with 3 schemes to build relation extraction systems in low-resource settings:

- **prompt**-based methods with **few-shot** labeled data
- **balancing** methods for the **long-tailed** distribution issue
- **data augmentation** and **self-training** to generate more **labeled** in-domain data from easy-collected unlabeled data

Prompting for Few-shot Instances

- Prompt-based tuning is more beneficial in general domains than specific domains for low-resource RE
- Entity type information in prompts is helpful for low-resource RE





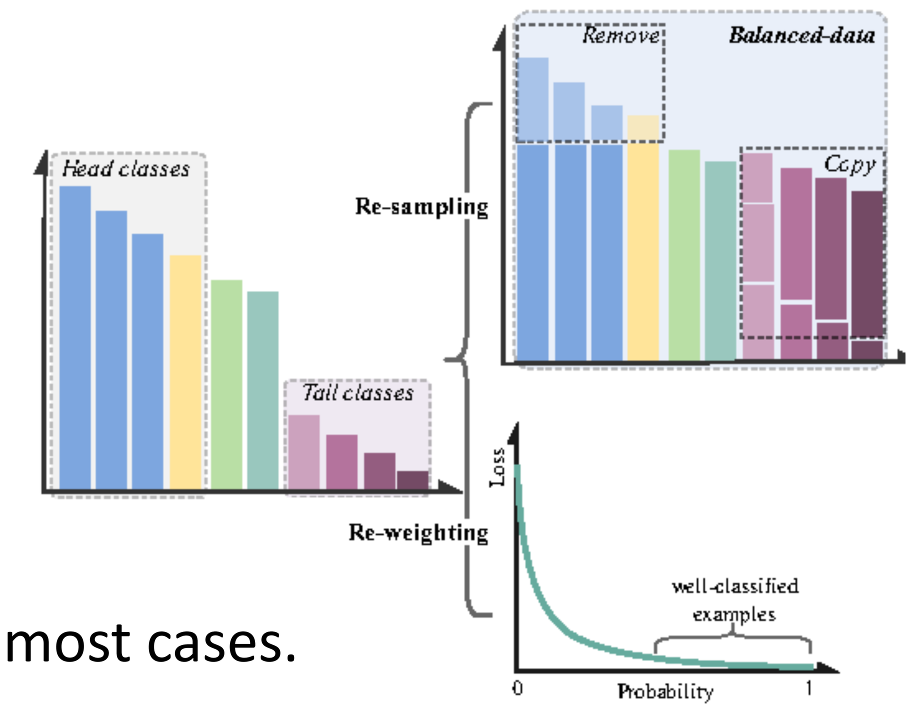
Balancing for Long-tailed Distribution

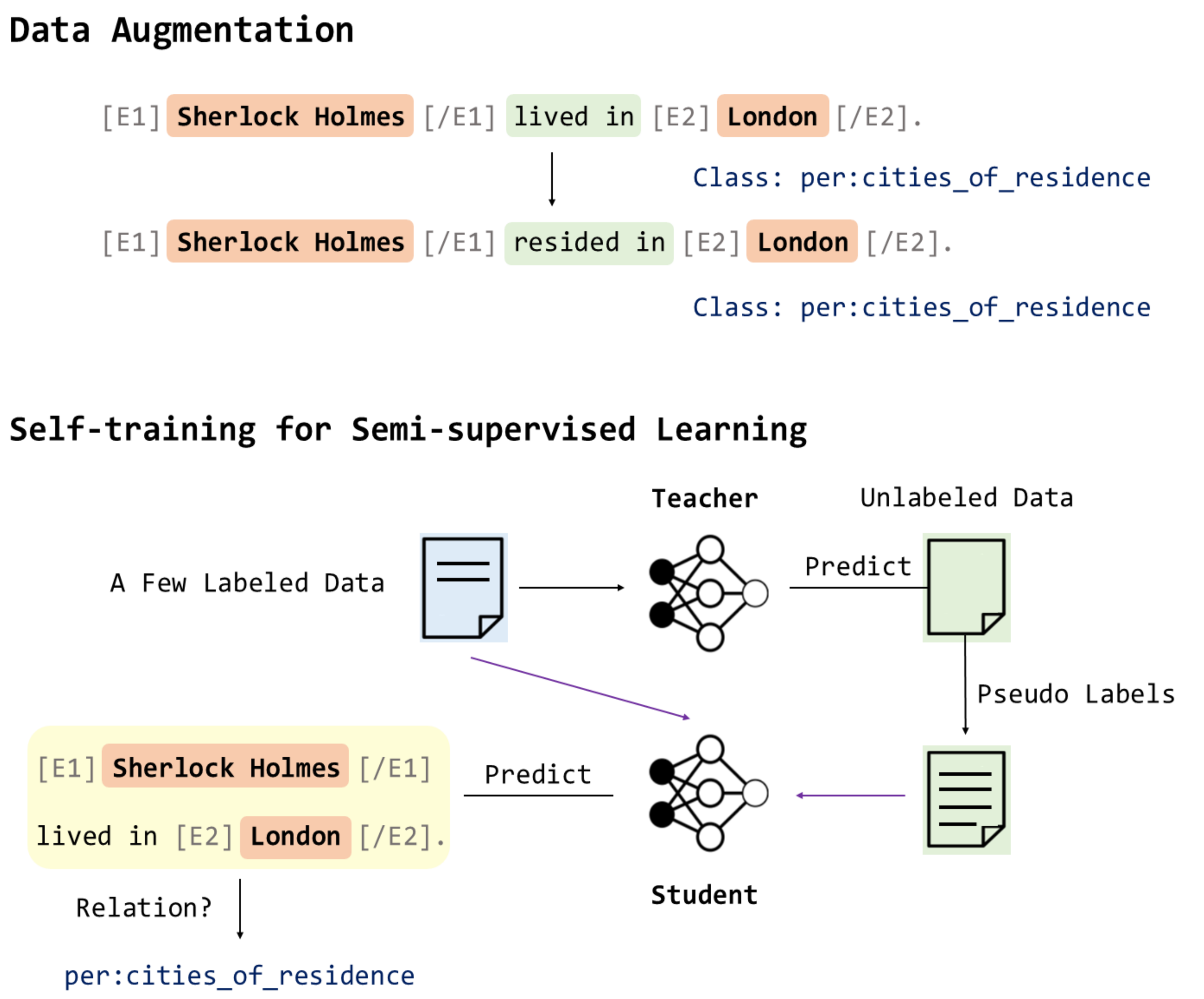
- Re-sampling Data
- Re-weighting Loss
- The tail relations can yield better performance on both general and domain-specific datasets with re-balancing methods (e.g., Focal Loss and LDAM Loss

Method	SemEval								SciERC							
	Few		Medium		Many		Overall		Few		Medium		Many		Overall	
	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1
Normal	50.42	74.58	**89.53**	89.02	90.17	90.59	83.40	90.01	69.98	67.78	88.05	87.52	92.98	91.93	83.27	89.01
Re-sample	38.17	56.18	70.13	70.56	71.22	71.54	65.37	71.31	71.79	69.64	88.49	87.83	92.96	92.25	82.61	87.58
DSC	49.80	73.87	87.84	88.00	88.97	89.52	82.19	89.00	71.57	69.90	89.94	89.51	93.51	92.88	83.09	88.91
Focal	53.31	77.69	89.50	**89.57**	**90.71**	**91.06**	**84.21**	**90.55**	**73.47**	**72.38**	**91.88**	**91.54**	**94.83**	**94.08**	**84.83**	**90.04**
GHM-C	00.00	00.00	3.39	6.27	70.42	75.81	43.79	70.99	71.34	69.28	89.42	88.82	93.90	93.33	82.95	88.81
LDAM	**53.53**	**79.66**	88.71	88.98	90.32	90.60	83.83	90.15	72.32	70.55	88.48	87.73	94.61	93.98	83.31	89.22

Leveraging More Instances via Data Augmentation and Self-training

- Data Augmentation
- Self-training
- Data augmentation is beneficial to few-shot RE and performs better in general domains than vertical domains.
- Self-training performs poorly on few-shot RE in most cases.





Method	SemEval						TACRED-Revisit					
	30%			100%			30%			100%		
	Context	Entity	All	Context	Entity	All	Context	Entity	All	Context	Entity	All
WordNet's Synonym	**75.49**	75.50	**76.47**	**83.54**	**83.50**	**82.56**	76.54	**76.87**	76.63	**76.12**	**76.59**	**76.37**
TF-IDF Similarity	73.93	**76.23**	74.30	82.92	82.61	82.33	**76.63**	76.05	**76.90**	75.44	75.80	75.15
Contextual Word Embedding (RoBERTa)	73.84	-	74.41	81.63	-	81.31	75.86	76.76	76.35	75.98	76.12	75.92
KnowPrompt (RoBERTa)	69.90						77.00					
	SciERC						**ChemProt**					
WordNet's Synonym	77.70	76.98	**77.54**	**79.36**	**79.40**	**79.92**	**57.37**	57.56	**57.03**	**53.36**	**57.11**	**54.27**
TF-IDF Similarity	**78.50**	**77.33**	73.92	78.30	79.38	79.38	41.22	**58.26**	47.95	43.06	54.60	43.63
Contextual Word Embedding (BioBERT)	76.24	73.55	74.62	75.50	77.35	76.59	56.01	53.48	56.28	45.95	53.26	46.68
KnowPrompt (BioBERT)	74.00						56.96					

Datasets and Settings

- 8 benchmark RE datasets in diverse domains and with different languages
- Training without validation and hyperparameter tuning

Dataset	Metric	Fine-Tune							Prompt						
		Normal			Balance		DA	ST	Normal			Balance		DA	ST
		8-shot	10%	100%	10%	100%	10%	10%	8-shot	10%	100%	10%	100%	10%	10%
SemEval	MaF1	2.69	34.63	81.88	41.84	82.44	69.84	60.10	48.54	44.71	83.40	54.54	83.20	71.73	63.55
	MiF1	9.70	54.61	89.10	58.26	89.44	78.98	74.12	54.55	69.90	90.01	76.53	92.31	83.54	76.81
TACREV	MaF1	1.02	47.32	63.41	48.64	63.38	50.68	48.84	29.46	61.40	67.08	63.09	69.63	62.20	7.32
	MiF1	1.76	65.43	71.68	67.19	73.86	65.99	66.89	30.88	77.00	78.30	76.25	81.41	76.90	32.93
Wiki80	MaF1	37.89	37.82	71.31	44.37	73.36	49.40	37.47	75.11	60.67	82.79	63.99	83.72	63.40	60.86
	MiF1	44.85	46.50	72.82	49.74	74.20	55.00	45.91	76.34	64.86	82.96	67.86	83.86	66.96	65.04
SciERC	MaF1	10.41	10.31	83.41	10.11	81.17	30.09	31.48	23.26	51.71	83.27	60.55	84.83	65.98	56.94
	MiF1	39.12	54.66	89.12	54.72	87.78	61.79	64.07	22.07	74.00	89.01	76.90	90.04	79.92	76.32
ChemProt	MaF1	2.18	27.96	47.35	33.38	47.35	36.31	30.67	6.17	36.43	47.16	38.99	47.07	37.44	33.62
	MiF1	8.93	49.20	68.81	54.98	68.77	56.58	54.17	8.65	56.96	69.14	57.28	69.12	58.26	53.55
DialogRE	MaF1	1.13	2.17	25.31	5.84	27.28	9.74	0.00	44.96	45.51	64.49	46.22	71.73	49.47	34.70
	MiF1	3.92	23.37	41.52	24.53	41.24	27.40	0.00	45.70	54.16	73.66	55.65	73.52	57.53	46.54
DuIE2.0	MaF1	36.62	90.46	95.01	92.91	96.00	91.47	89.27	80.31	93.48	95.73	93.70	96.01	93.66	90.49
	MiF1	39.00	94.42	96.22	94.46	96.13	94.46	93.81	82.14	95.09	96.43	95.23	96.44	95.11	93.35
CMeIE	MaF1	13.68	62.30	84.37	67.22	86.31	63.82	58.46	36.54	67.59	86.42	67.84	86.68	69.95	65.79
	MiF1	17.05	79.82	90.48	80.43	90.56	80.14	78.92	38.02	83.38	92.08	83.40	92.14	83.71	81.26

Findings

- **Prompt-based tuning** largely **outperform**s standard fine-tuning for RE, especially more effective in the low-resource scenario
- Though **balancing methods** obtain advancement with long-tailed distribution, they may still **fail on challenging RE datasets**
- **Data augmentation** achieves much gain on RE and even better performance than prompt-based tuning
- RE systems **struggle** against difficulty in obtaining correct relations from **cross-sentence contexts and among multiple triples**
- **Self-training** with unlabeled in-domain data may **not always show an advantage** for low-resource RE

Xiang Chen, et al. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. Proceedings of the ACM Web Conference 2022.
Xu Han, et al. OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. Proceedings of the 2019 EMNLP-IJCNLP: System Demonstrations.
Xu Han, et al. PTR: Prompt Tuning with Rules for Text Classification. arXiv: 2105.11259.