Semi-supervised Relation Extraction via Incremental Meta Self-Training

Xuming Hu[1], Chenwei Zhang[2], Fukun Ma[1], Chenyao Liu[1], Lijie Wen[1] ,Philip S. Yu[1,3]

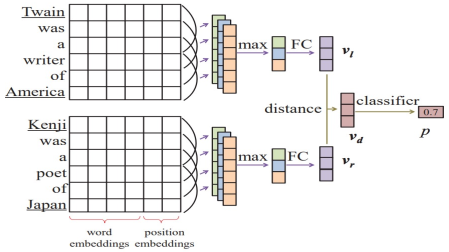
[1] Tsinghua University
[2] Amazon
[3] University of Illinois at Chicago

Relation Extraction

Labor-intensive

Sentence		Relation
Derek Bell was born in *Belfast*. *Donald Trump* was born in *America*. ……	→	Born In
Thomson is based in *Toronto*. *Beijing* is located in *China*. ……	→	Located in

Relation Encoder + Deep Classification Model

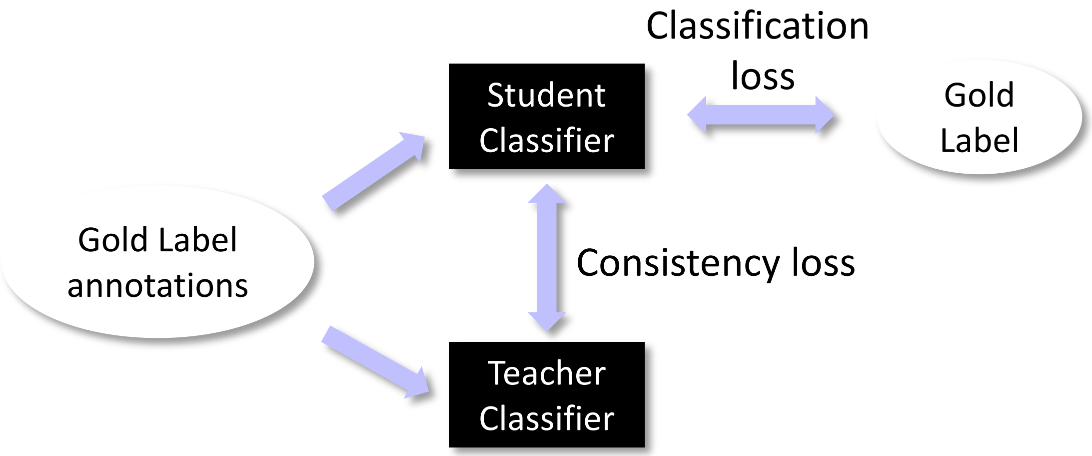


(Stanovsky et al., 2018;
Saha et al., 2018;
Yu et al., 2017)

Semi-supervised Relation Extraction

Leverage unlabeled data

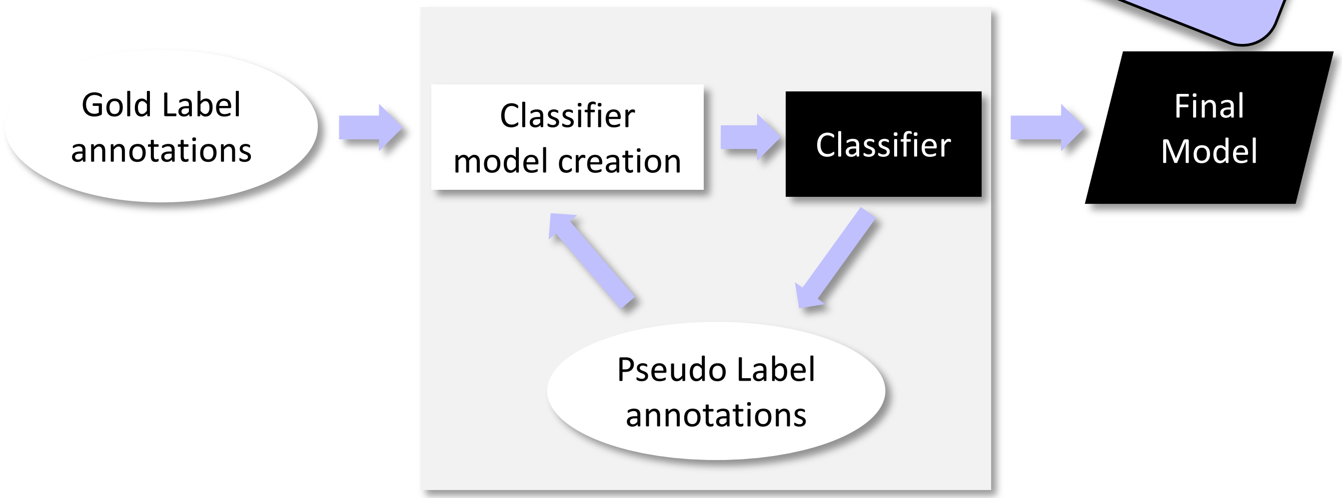
Self-Ensembling



😁 **Robust in model parameters**

😭 **Insufficient supervision**

Self-Training



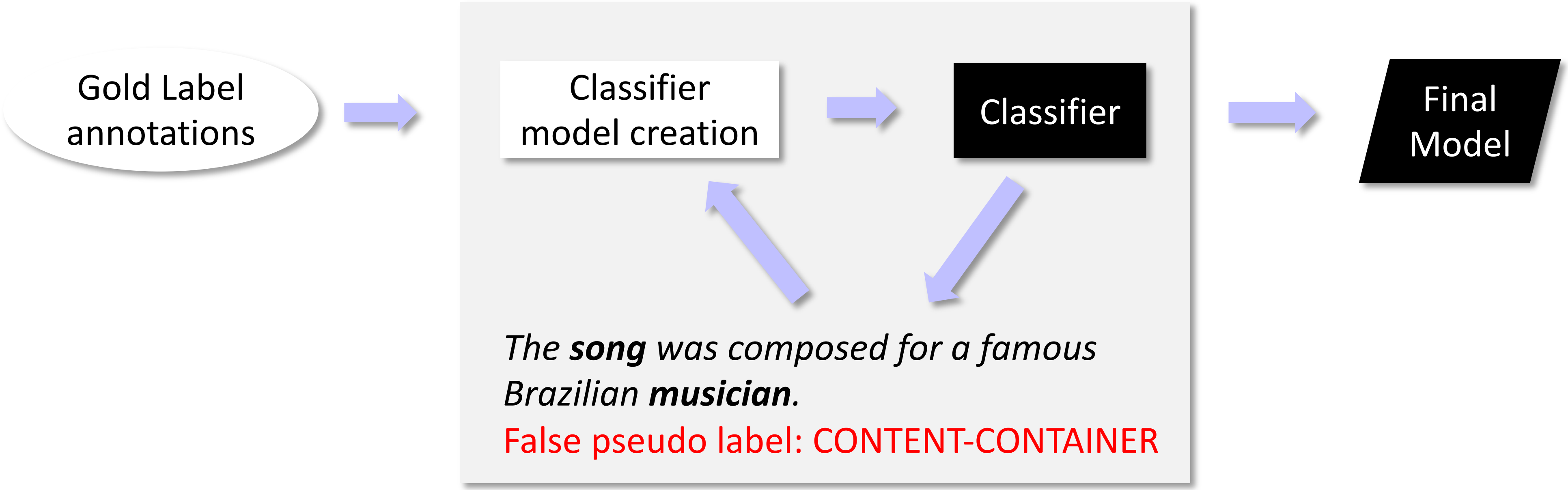
😁 **Supervision from unlabeled data**

😭 **Gradual drift problem**

Gradual Drift problem

Accumulate errors in the subsequent training

Self-Training

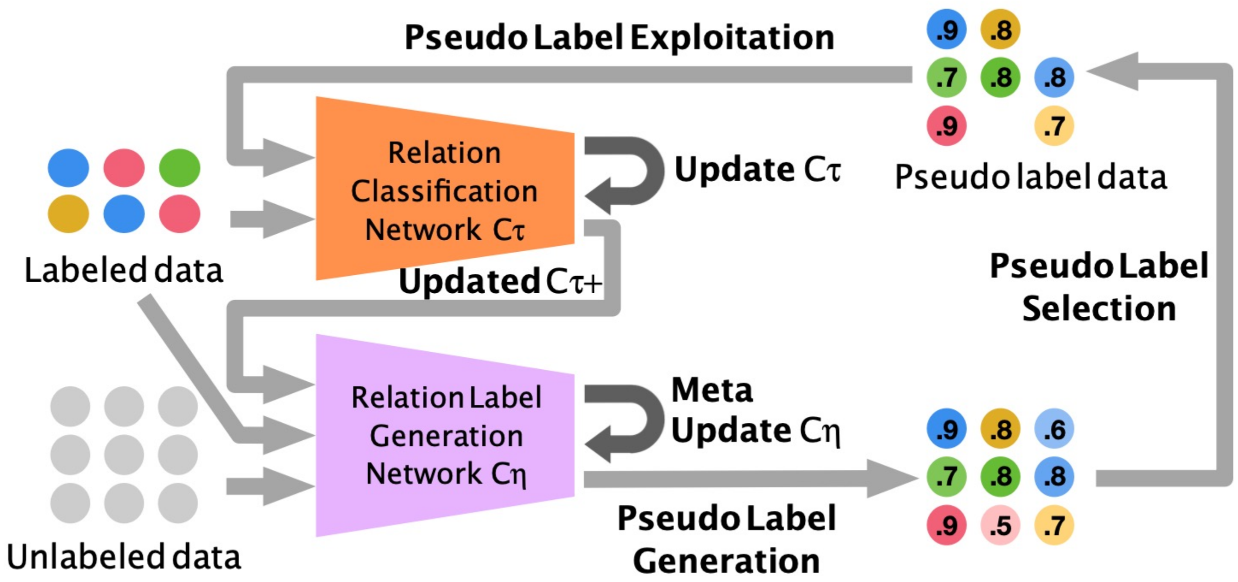


How to alleviate the gradual drift problem?

- Task: Solve the gradual drift problem.
- Goal: Generate high-quality pseudo label from the unlabeled data.
- Methods:

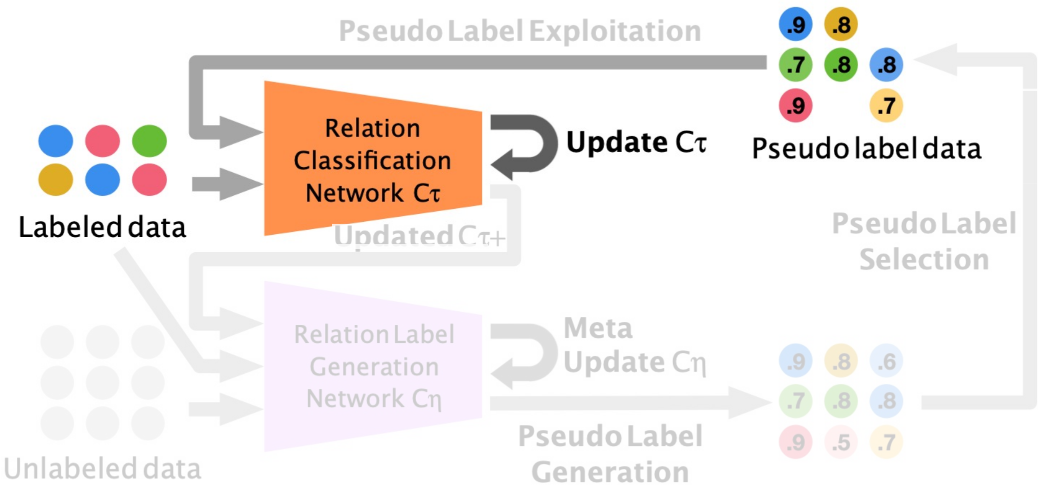
  ① Meta Learning: Learn to assess the quality of pseudo labels by (meta) learning from the **successful and failed attempts**.
  ② Pseudo Label Selection: Select **informative and high-quality** pseudo labels.
  ③ Pseudo Label Exploitation: Exploit pseudo labels with **confidence**.

Framework (MetaSRE)



- Relation Classification Network
  ① Contextualized Relation Encoder
  ② Relation Classification
- Relation Label Generation Network
- Pseudo Label Selection and Exploitation

Relation Classification Network



- Contextualized Relation Encoder

  $[E1_{start}]$ *Derek Bell* $[E1_{end}]$ *was born in* $[E2_{start}]$ *Belfast* $[E2_{end}]$

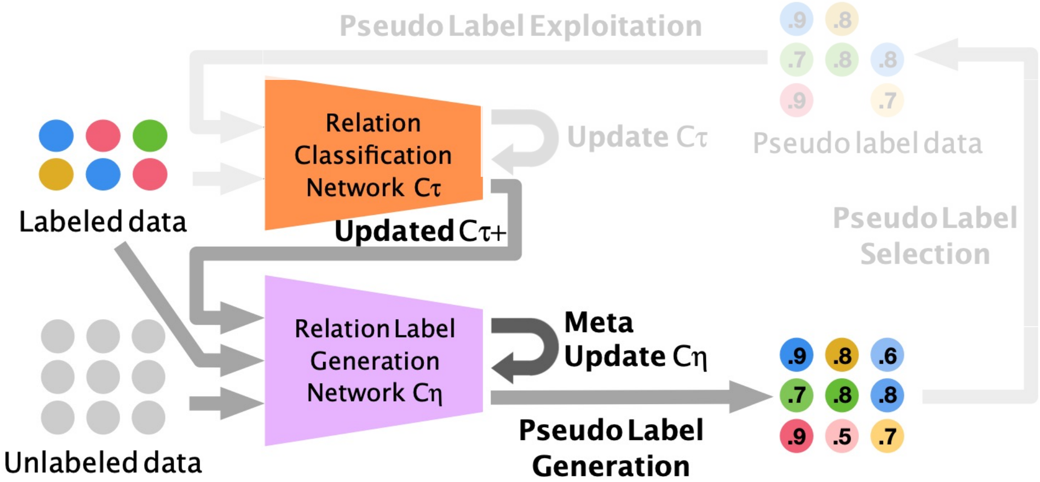
  Get the relation representation of two entities corresponding to $[E1_{start}]$, $[E2_{start}]$ from BERT.

  $$\mathbf{h} = [\mathbf{h}_{[E1_{start}]}, \mathbf{h}_{[E2_{start}]}]$$

- Relation Classification

  Classifiy Labeled data and Pseudo label data representations into specific relations with a fully connected network $C_\tau(X_{n,E1,E2})$.

Relation Label Generation Network



- **Purpose:**

  Prevent the noise contained in the pseudo labels.

- **Updated Relation Classification Network:**

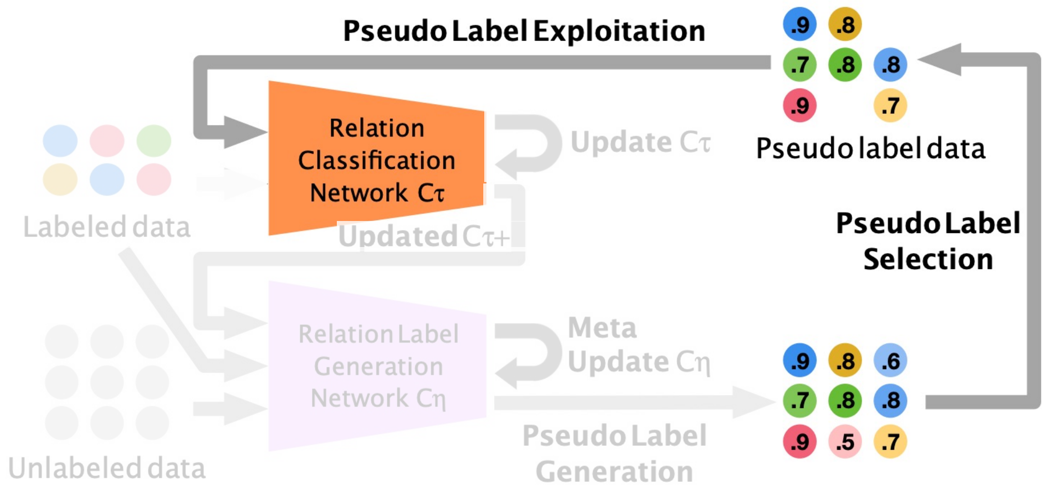
$$\tau^{+} \leftarrow \tau - \alpha \nabla_{\tau} \mathcal{L}_{C_\tau}$$

- **Meta Objective:**

  Perform a derivative over the parameters on the updated Relation Classification Network using labeled data g_n.

$$\underset{\eta}{\operatorname{argmin}}\, loss\left(\boxed{C_{\tau^{+}}(X_{n,E1,E2})}, \mathrm{one\_hot}(g_n)\right)$$

Pseudo Label Selection and Exploitation



- ## Pseudo Label Selection

We treat maximum probability after softmax as the confidence score. Sort them in a descending order and select top Z%.

$$\text{argmax}\left(C_\eta\left(X_{m',E1,E2}\right)\right)$$

- ## Pseudo Label Exploitation

We use the maximum probability value as the weight of the pseudo label data to optimize Relation Classification Network.

$$w_m = \max_m\left(C_\eta\left(X_{m,E1,E2}\right)\right)$$

Experiments

Datasets

Datasets	SemEval	TACRED
Relation mentions	7199/800/1864	75049/25763/18659
Relation	19	42
No_relation rate	17.4%	78.7%

Implementations

Datasets	SemEval	TACRED
Labeled set Unlabeled set	5%/10%/30% 50%	3%/10%/15% 50%

Baselines

- Relation Encoders
  LSTM (Hochreiter and Schmidhuber, 1997)
  PCNN (Zeng et al., 2015)
  PRNN (Zhang et al., 2017)
  BERT (Devlin et al., 2019)
- Self-Training (Rosenberg et al., 2005)
- Mean-Teacher (Tarvainen and Valpola, 2017)
- DualRE (Lin et al., 2019)
- MRefG (Li and Qian, 2020)
- BERT w. gold labels

Does meta learning and pseudo label selection give better quality pseudo labels?

Yes!

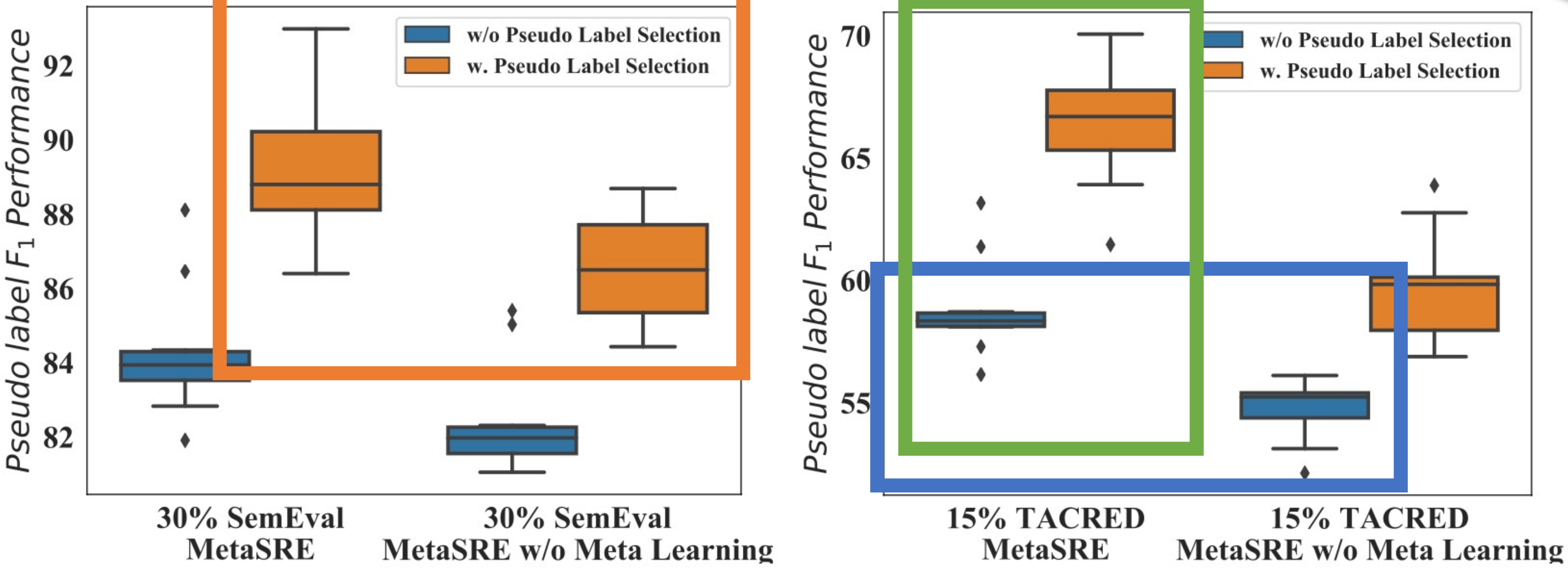


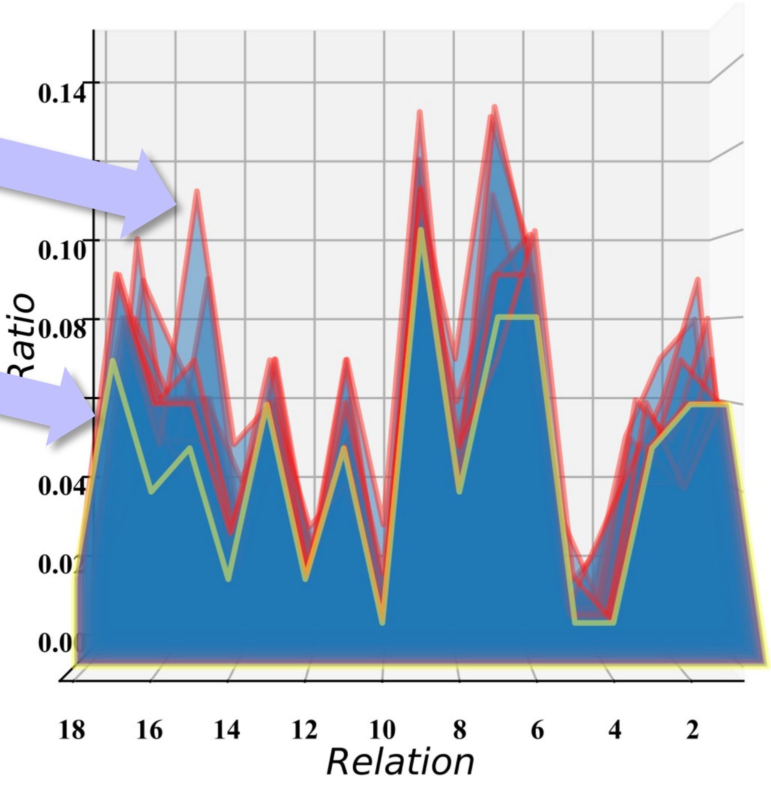
Figure 2: Pseudo label F_1 Performance with different modules based on SemEval (left) and TACRED (right).

Does meta learning prevent the gradua lem?

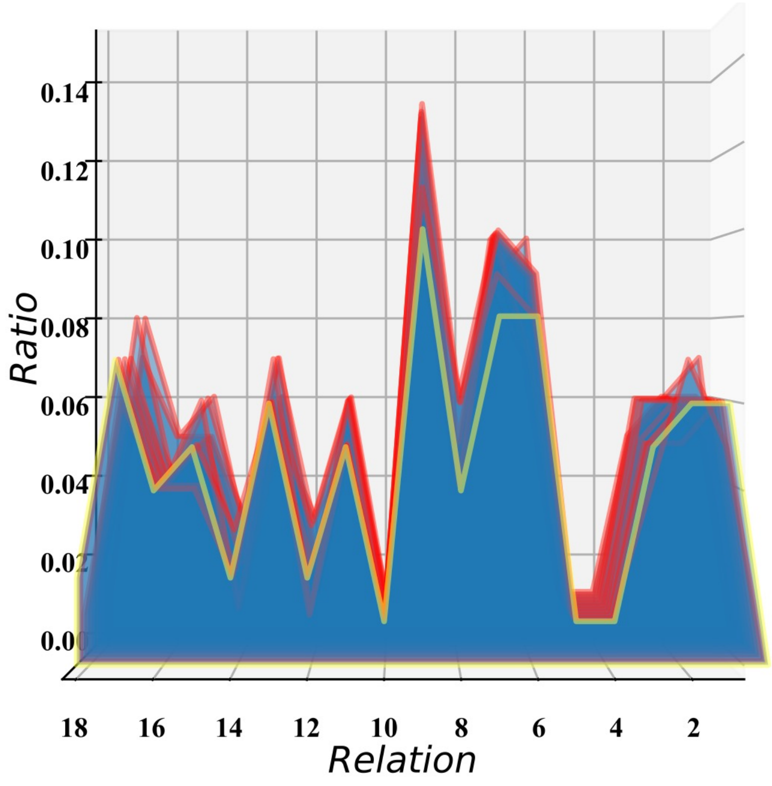
Yes!

Red line is the pseudo label distribution.

Yellow line is the gold label distribution.



MetaSRE w/o Meta Learning



MetaSRE

How much unlabeled data is needed?

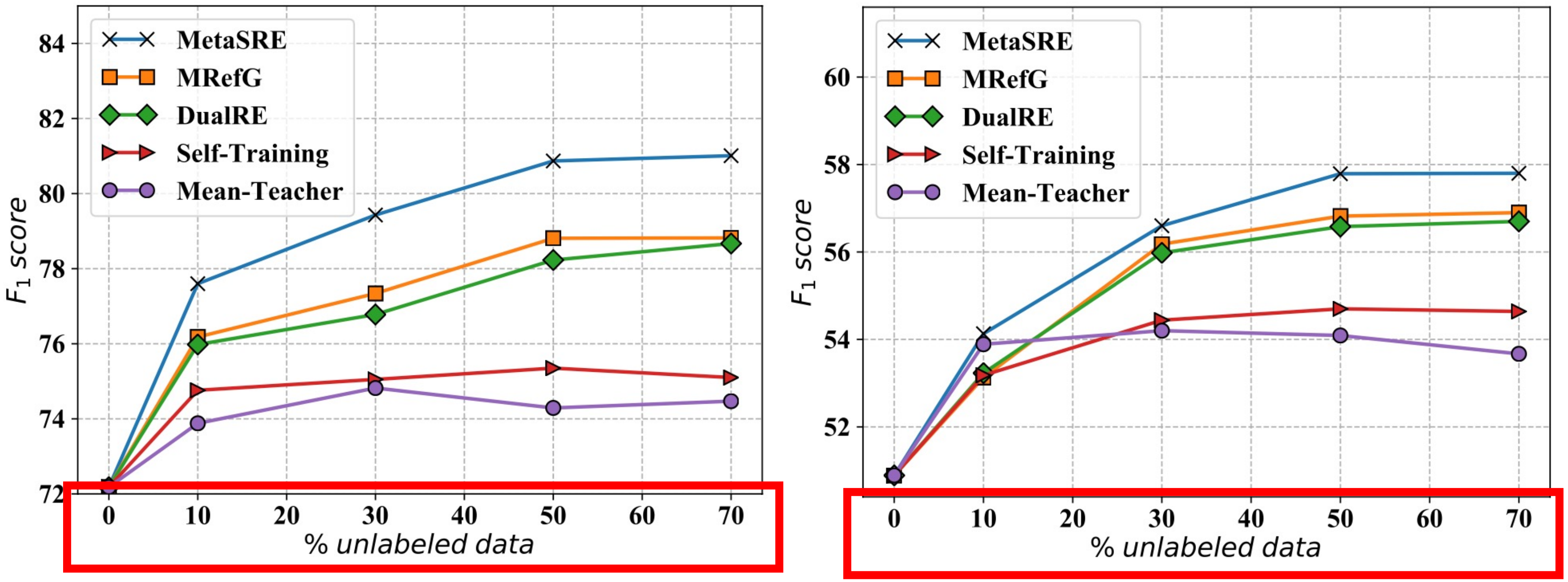


Figure 4: F_1 Performance with various unlabeled data and 10% labeled data on SemEval (left) and TACRED (right).

THANK YOU!

Code + Data are Available at:
http://github.com/THU-BPM/MetaSRE
https://arxiv.org/abs/2010.16410