Semi-supervised Relation Extraction via Incremental Meta Self-Training

Xuming Hu¹, Chenwei Zhang², Fukun Ma¹, Chenyao Liu¹, Lijie Wen¹, Philip S. Yu¹,³

¹ Tsinghua University
² Amazon
³ University of Illinois at Chicago
Relation Extraction

Sentence

Derek Bell was born in Belfast.
Donald Trump was born in America.

Thomson is based in Toronto.
Beijing is located in China.

Relation

Born In
Located in

Relation Encoder + Deep Classification Model

(Stanovsky et al., 2018; Saha et al., 2018; Yu et al., 2017)
Semi-supervised Relation Extraction

**Self-Ensembling**

- Student Classifier
- Teacher Classifier
- Gold Label annotations
- Consistency loss
- Classification loss

- Robust in model parameters
- Insufficient supervision

**Self-Training**

- Gold Label annotations
- Classifier model creation
- Pseudo Label annotations

- Supervision from unlabeled data
- Gradual drift problem

Leverage unlabeled data
Gradual Drift problem

Self-Training

Gold Label annotations

Classifier model creation

Classifier

Final Model

Accumulate errors in the subsequent training

The song was composed for a famous Brazilian musician.
False pseudo label: CONTENT-CONTAINER
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**
• **Contextualized Relation Encoder**

\[
[ E_{1\text{start}} ] \text{ Derek Bell } [ E_{1\text{end}} ] \text{ was born in } [ E_{2\text{start}} ] \text{ Belfast } [ E_{2\text{end}} ]
\]

Get the relation representation of two entities corresponding to \([ E_{1\text{start}} ], [ E_{2\text{start}} ]\) from BERT.

\[
h = [ h_{[ E_{1\text{start}} } ], h_{[ E_{2\text{start}} ]} ]
\]

• **Relation Classification**

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

\[
\arg\min_{\eta} \text{loss} \left( C_{\tau^+}(X_n, E_1, E_2), \text{one_hot}(g_n) \right)
\]
• Pseudo Label Selection

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

\[
\text{argmax} \left( \eta(X_{m}, E_{1}, E_{2}) \right)
\]

• Pseudo Label Exploitation

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

\[
w_{m} = \max_{m} \eta(X_{m}, E_{1}, E_{2})
\]
# Experiments

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SemEval</th>
<th>TACRED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation mentions</td>
<td>7199/800/1864</td>
<td>75049/25763/18659</td>
</tr>
<tr>
<td>Relation</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>No_relation rate</td>
<td>17.4%</td>
<td>78.7%</td>
</tr>
</tbody>
</table>

## 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**

## Implementations

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SemEval</th>
<th>TACRED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled set</td>
<td>5%/10%/30%</td>
<td>3%/10%/15%</td>
</tr>
<tr>
<td>Unlabeled set</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

11/7/21
Does meta learning and pseudo label selection give better quality pseudo labels?

Yes!

Figure 2: Pseudo label F1 Performance with different modules based on SemEval (left) and TACRED (right).
Does meta learning prevent the gradual drift problem?

Yes!

Red line is the pseudo label distribution.

Yellow line is the gold label distribution.
How much unlabeled data is needed?

Figure 4: F₁ 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