H-FND: Hierarchical False-Negative Denoising for Distantly Supervision Relation Extraction

Jhin-Wei Chen, Tsu-Jui Fu, Chen-Kang Lee and Wei-Yun Ma

Findings ACL-IJCNLP 2021
Challenges in Information Extraction

- The lack of labeled data
- Labeling data is very expensive and time consuming

Distant supervision was proposed to generate training data by aligning triples in knowledge bases to unannotated sentences.
# Noise from Distantly Supervised Relation Extraction Datasets

<table>
<thead>
<tr>
<th>Knowledge base</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Jobs, San Francisco</td>
<td>PoB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Relation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs was born in San Francisco</td>
<td>PoB (✓)</td>
<td>TP</td>
</tr>
<tr>
<td>Jobs moved back to San Francisco</td>
<td>PoB (✗)</td>
<td>FP</td>
</tr>
<tr>
<td>Manuela was born in New York</td>
<td>NA (✗)</td>
<td>FN</td>
</tr>
</tbody>
</table>
The Ternary Policy of the Agent

To retrieve positive instance from negative samples

For each negative sample $s$ in the dataset:

- **Keep**: maintain $s$ as a negative instance for training/validation
- **Discard**: remove $s$ to prevent it from misleading the model
- **Revise**: predict a new relation type for $s$ and treat it as a positive for the following training/validation.
The H-FND Framework
Experiments on Synthetic Noise

![Graphs showing F1 score vs. FN ratio for CNN and PCNN on SemEval and TACRED datasets.](image)
Experiment on Distantly Supervised Dataset

**CNN On NYT**

**PCNN On NYT**