Learning with Noise: Improving Distantly-Supervised Fine-grained Entity Typing via Automatic Relabeling

Haoyu Zhang*, Dingkun Long*, Guangwei Xu
Muhua Zhu, Pengjun Xie, Fei Huang, Ji Wang

Alibaba Group
阿里巴巴集团

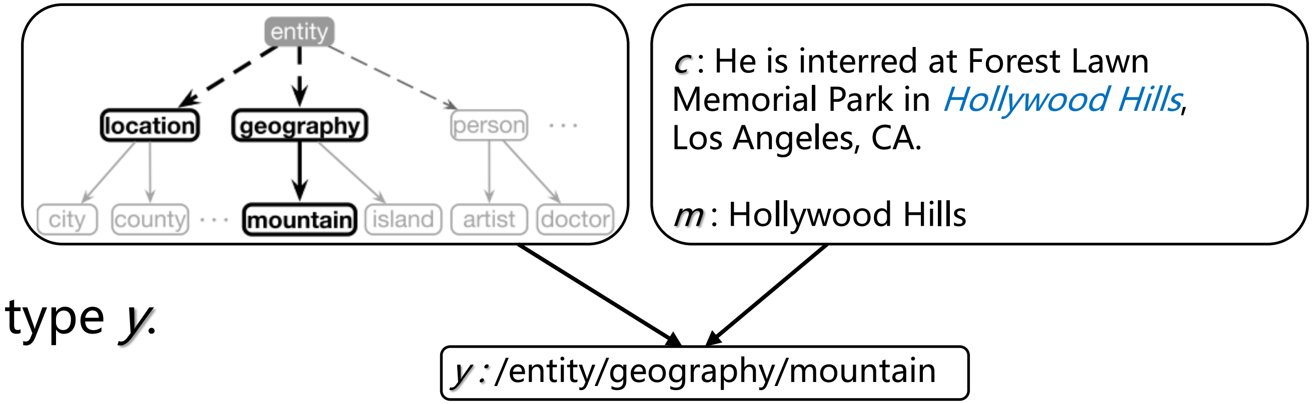
国防科技大学

*: equal contribution

Background

Fine-Grained Entity Typing (FET):

- A pre-defined type hierarchy.
- Given sentence *c* with entity mention *m* .
- Answer one most appropriate fine-grained entity type *y*.



distant supervision: Construct supervised corpus for FET

- entity mention link> knowledge base nodes
- node types ---> the distant labels
- provide supervision & NOISE

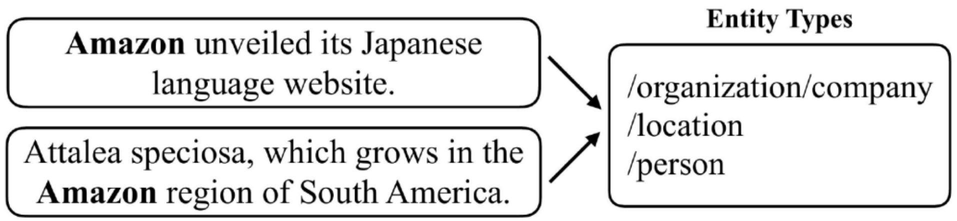


Figure from (Chen et al. Hierarchical Entity Typing via Multi-level Learning to Rank. 2020.)

Recent Approaches

NDP (Wu et al. 2019):

- Weight out noisy samples.
- Assumption: distant labels always contain the correct type <- Overly Strong

NFETC-CLSC (Chen et al. 2019):

- Classification objective on "clean" samples
- Unsupervised cluster objective on "clean" + "noisy" samples <- Not always valid
- Prerequisites: judgeable "clean" training set
- Assumption: samples with one type label contains no noise <- Overly Strong

Wu et al. Modeling Noisy Hierarchical Types in Fine-Grained Entity Typing: A Content-Based Weighting Approach. 2019.
Chen et al. Improving Distantly-supervised Entity Typing with Compact Latent Space Clustering. 2019.

Recent Approaches

NDP (Wu et al. 2019):

- Weight out noisy samples.
- Assumption: distant labels always contain the correct type

NFETC-CLSC (Chen et al. 2019):

- Classification objective on "clean" samples
- Unsupervised cluster objective on "clean" + "noisy" samples
- Prerequisites: judgeable "clean" training set
- Assumption: samples with one type label contains no noise

```
In a word:
-> Rely on Overly Strong assumptions/prerequisites
-> × × false positive one type "clean" samples × ×
```

Wu et al. Modeling Noisy Hierarchical Types in Fine-Grained Entity Typing: A Content-Based Weighting Approach. 2019.
Chen et al. Improving Distantly-supervised Entity Typing with Compact Latent Space Clustering. 2019.

Our Method

Use *NFETC* with parameters θ as the backbone network.

Pseudo Label Distribution Estimation:

- $\tilde{p}$: Continuous pseudo distribution $\xrightarrow{\text{assign}}$ each training sample
- Trainable, update through BP.
- $\theta^* = \mathrm{argmin}_\theta \mathcal{L}(m, c, y; \theta)$ ---> $\theta^*, \tilde{p}^* = \mathrm{argmin}_{\theta,\tilde{p}} \mathcal{L}(m, c, y; \theta, \tilde{p})$

$$\tilde{p} \leftarrow \mathcal{KL} \rightarrow p$$

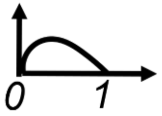
Automatic Relabeling (AR) Process:

- Assumption: each sample has exact one true type label
- Objective: $\mathcal{KL}(\tilde{p}||p)$

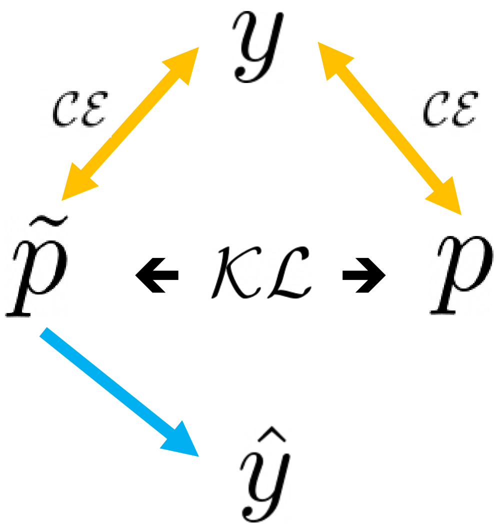
Confirmation Bias!
Ignore the noisy labels

Our Method

Automatic Relabeling (AR) Process:

- Assumption: each sample has exact one true type label
- Main Objective: $\mathcal{KL}(\tilde{p}||p)$
- Information in noisy labels:
  - ××Totally different from noisy labels ××: $\mathrm{L}_d = \mathcal{CE}(\tilde{p}, y)$
  - Noisy labels contain valuable information, using it for pseudo distribution initialization
  - keep predictive distribution reasonable: $\mathrm{L}_{ce} = \mathcal{CE}(p, y)$
- Distribution Sharpen Constraint: 

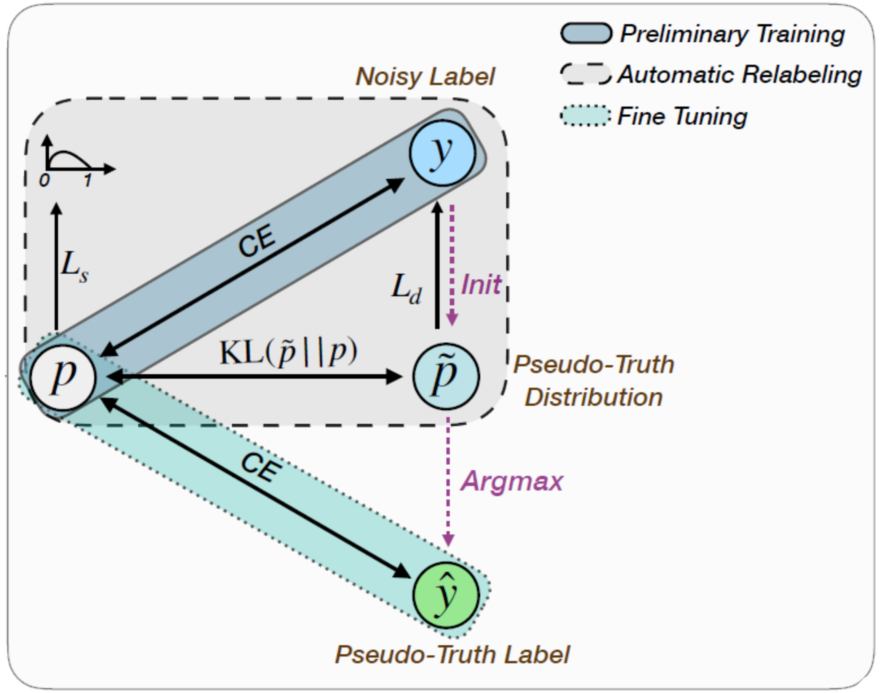
- Estimated distribution $\tilde{p}$ -> one-hot pseudo labels $\hat{y}$



Our Method

Three phrase Training:

- Preliminary Training, use y, update θ.
- Automatic Relabeling.
- Fine-Tuning, use $\hat{y}$, update θ.



NFETC-AR Framework

Exp: Baselines & Evaluation

Baselines:

- NFETC - (Xu et al., 2018, our backbone). [variants w/wo hier loss]
- NDP – (Wu et al., 2019).
- NFETC-CLSC - (Chen et al., 2019).
- NFETC-AR w/o KL/Noisy-Info/Sharpen - (ablation study).

Datasets:

- BBN, Wiki, OntoNotes

Evaluation:

- Classification Accuracy
- Macro F1
- Micro F1

Exp: Results

	Acc	Macro F1	Micro F1
NFETC-AR$_{hier}$	**64.0 ± 0.3**	**78.8 ± 0.3**	**73.0 ± 0.3**
w/o $\mathcal{L}_{kl}$	61.2 ± 0.5	76.6 ± 0.4	70.4 ± 0.5
w/o $\mathcal{L}_{d}$	63.8 ± 0.3	78.4 ± 0.2	72.6 ± 0.3
w/o $\mathcal{L}_{ce}$	61.1 ± 0.3	76.1 ± 0.3	69.9 ± 0.5
w/o noisy label init	55.0 ± 0.3	67.1 ± 0.4	60.3 ± 0.4
w/o $\mathcal{L}_{s}$	63.7 ± 0.6	78.3 ± 0.4	72.3 ± 0.6
w/o AR	60.2 ± 0.2	76.4 ± 0.1	70.2 ± 0.2

ablation on OntoNotes

Model	Wiki			OntoNotes			BBN		
	Strict Acc	Macro F1	Micro F1	Strict Acc	Macro F1	Micro F1	Strict Acc	Macro F1	Micro F1
AFET	53.3	69.3	66.4	55.3	71.2	64.6	68.3	74.4	74.7
AAA	65.8	81.2	77.4	52.2	68.5	63.3	65.5	73.6	75.2
Attentive	59.7	80.0	75.4	51.7	71.0	64.91	48.4	73.2	72.4
NDP	67.7	81.8	78.0	58.0	71.2	64.8	72.7	76.4	77.7
NFETC	56.2 ± 1.0	77.2 ± 0.9	74.3 ± 1.1	54.8 ± 0.4	71.8 ± 0.4	65.0 ± 0.4	73.8 ± 0.6	78.4 ± 0.6	78.9 ± 0.6
NFETC$_{hier}$	**68.9 ± 0.6**	**81.9 ± 0.7**	**79.0 ± 0.7**	60.2 ± 0.2	76.4 ± 0.1	70.2 ± 0.2	**73.9 ± 1.2**	78.8 ± 1.2	**79.4 ± 1.1**
NFETC-CLSC	-	-	-	59.6 ± 0.3	75.5 ± 0.4	69.3 ± 0.4	**74.7 ± 0.3**	**80.7 ± 0.2**	**80.5 ± 0.2**
NFETC-CLSC$_{hier}$	-	-	-	**62.8 ± 0.3**	**77.8 ± 0.3**	**72.0 ± 0.4**	73.0 ± 0.3	79.8 ± 0.4	79.5 ± 0.3
NFETC-AR	58.1 ± 1.1	79.0 ± 0.4	76.1 ± 0.4	62.8 ± 0.4	77.8 ± 0.4	71.8 ± 0.5	**76.7 ± 0.2**	**81.4 ± 0.3**	**81.5 ± 0.3**
NFETC-AR$_{hier}$	**70.1 ± 0.9**	**83.2 ± 0.7**	**80.1 ± 0.6**	**64.0 ± 0.3**	**78.8 ± 0.3**	**73.0 ± 0.3**	74.9 ± 0.6	80.4 ± 0.6	80.3 ± 0.6

Classification performances on three benchmark datasets

Exp: Analysis

Source	Type	Context & Mention	Original Label	After Relabeling
OntoNotes	Multi-to-one-in	Jennifer Laden , NPR News , Jerusalem	{/org/company/news, /person}	/org/company/news
Wiki	Multi-to-one-in	the Plains of Abraham in Quebec City , Quebec , Canada .	{/location/country, /person/director}	/location/country
BBN	Multi-to-one-in	Federal researchers said lung-cancer mortality rates for people under 45	{/person, /org/government}	/person
OntoNotes	Multi-to-one-out	Because along with Haditha comes Jesse Macbeth , allegedly ...	{/other/art/writing, /other/art/stage}	/person
OntoNotes	One-to-one-out	... Alcee Hastings of Florida of eight impeachment articles ,...	/other/health/malady	/other/art/writing

Three types of relabeling

- multi-to-one-in: 97.49%
- multi-to-one-out: 0.12%
- one-to-one-out: 2.39%
- Relabeling/All-Samples: 27.52%

Conclusion

Automatic Relabeling Process:

- Helps infer pseudo labels and use the corrected labels for fine-tuning.
- Verified on noisy label FET task.
- NOT rely any prerequisite or extra supervision. (robust)
- Backbone agnostic, noise agnostic. (general & flexible)

Thanks for listening!

For more details, please refer to our paper/slides/poster:

Paper

Slides

Poster