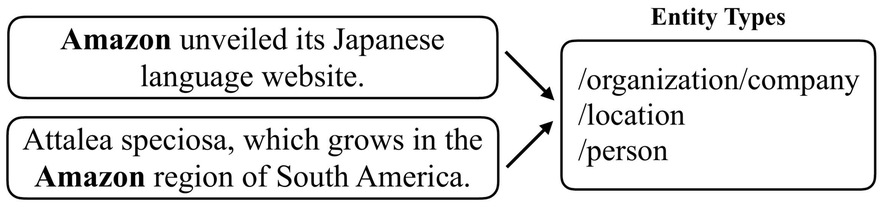
Learning with Noise: Improving Distantly-Supervised Fine-grained Entity Typing via Automatic Relabeling

Haoyu Zhang[1], Dingkun Long[2], Guangwei Xu[2], Muhua Zhu[2], Pengjun Xie[2], Fei Huang[2], Ji Wang[1]

[1]HPCL, College of Computer, National University of Defense Technology, China; [2]Alibaba Group, China;
Correspondence to: zhanghaoyu10@nudt.edu.cn

Introduction

Fine-grained entity typing (FET) is a fundamental task for various entity-leveraging applications. Although great success has been made, existing systems still have challenges in handling noisy samples in training data introduced by distant supervision method. In this paper, we propose a probabilistic automatic relabeling method which treats all training samples uniformly. Our method aims to estimate the pseudo-truth label distribution of each sample, and the pseudo-truth distribution will be treated as part of trainable parameters which are jointly updated during the training process. The proposed approach does not rely on any prerequisite or extra supervision, making it effective on real applications. Experiments on several benchmarks show that our method outperforms previous competitive approaches and indeed alleviates the noisy labeling problem.



Noisy samples introduced by distant supervision. The entity mention "Amazon" in two different sentences will be labeled with same entity type set, in which some types are inappropriate given the context.

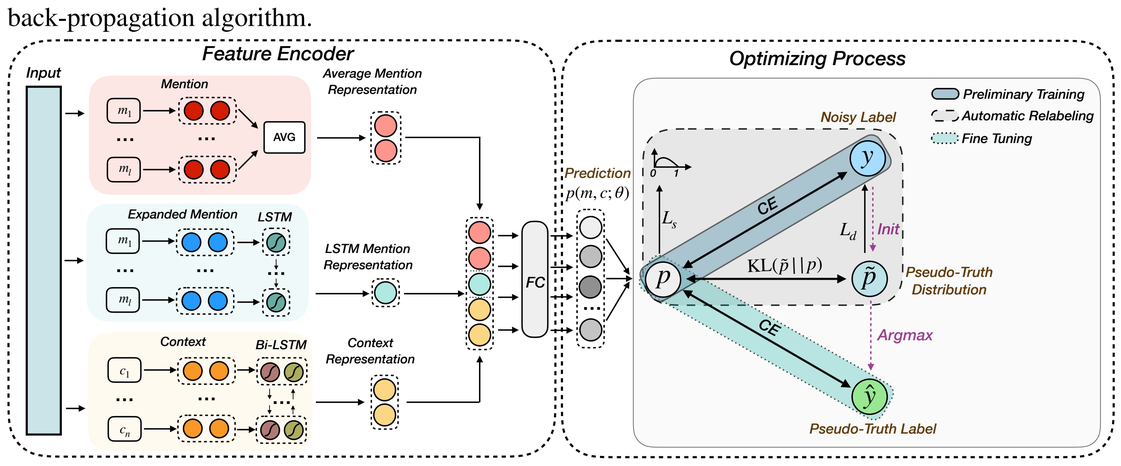
Motivation

To address the issue of noisy labeling, most of previous studies try to model the samples with *only one label* and samples *with multiple labels* separately, or to detect and weight out noise based on the assumption that the distantly-supervised label set *must contain the correct type*. The weaknesses of these studies are:

- Previous work rely on overly strong assumptions.
- Existing methods cannot deal with false positive one type "clean" samples.

Method

A probabilistic automatic relabeling (AR) method which handles the above two limitations simultaneously. As the ground-truth label distribution is not available, our method aims at estimating the pseudo-truth label distribution during the training process. In detail, each sample is assigned a continuous label distribution $\tilde{p}$ over all candidate labels, and $\tilde{p}$ will be jointly updated as trainable parameters through the back-propagation algorithm.



The architecture of our proposed **NFETC-A**utomatic-**R**elabeling (**NFETC-AR**) model. The left part is the encoder of the backbone model NFETC; the right part illustrates the automatic relabeling process along with the three-phrase training strategy.

The learning objective of the AR process includes:

- Self-learning objective: KL divergence between pseudo distribution and predictive distribution.
- Information in noisy labels: initialization, cross entropy and deviation constraints.
- A distribution sharpen constraint to control the shape of the pseudo distribution.

Finally, we utilize the AR process with a three-phrase training strategy: 1) pre-training, 2) relabeling and 3) fine-tuning using the estimated pseudo labels.

Overall Results

- Three datasets Wiki, OntoNotes, BBN. (two shown in poster)
- Several competitive baselines, including noisy learning based FET methods (NFETC-CLSC, NDP).
- methods with $_{hier}$ denotes for variants with hierarchical loss proposed by NFETC.

Model	Wiki			OntoNotes		
	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
AAA	65.8	81.2	77.4	52.2	68.5	63.3
Attentive	59.7	80.0	75.4	51.7	71.0	64.91
NDP	67.7	81.8	78.0	58.0	71.2	64.8
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
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
NFETC-CLSC	-	-	-	59.6 ± 0.3	75.5 ± 0.4	69.3 ± 0.4
NFETC-CLSC$_{hier}$	-	-	-	**62.8 ± 0.3**	**77.8 ± 0.3**	**72.0 ± 0.4**
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
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**

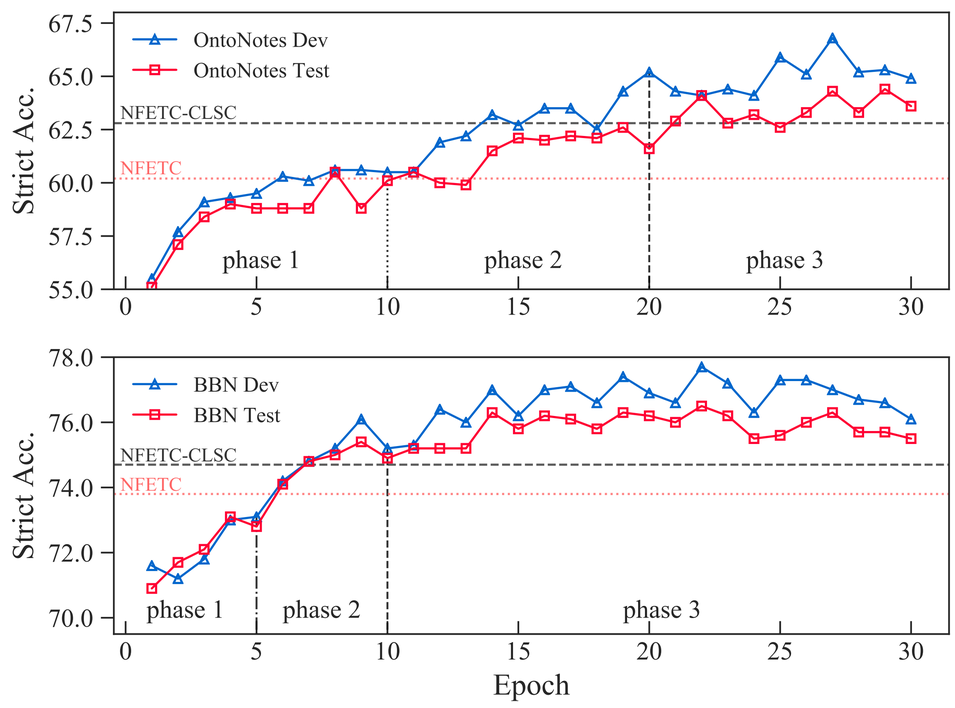
Ablation study of AR

- Self-learning (KL divergence) is the key component.
- Noise information (Noisy label initialization) is also important.

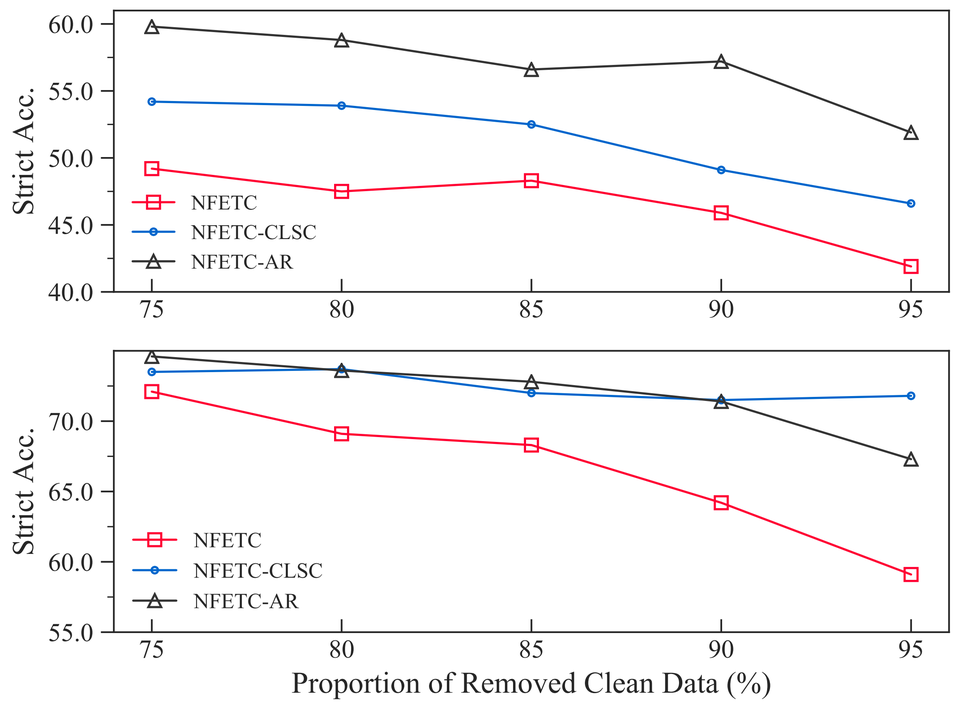
	Acc	Macro F1	Micro F1
NFETC-AR$_{hier}$	**64.0 ± 0.3**	**78.8 ± 0.3**	**73.0 ± 0.3**
w/o $\mathcal{L}_{kl}$	61.2 ± 0.5	76.6 ± 0.4	70.4 ± 0.5
w/o $\mathcal{L}_{d}$	63.8 ± 0.3	78.4 ± 0.2	72.6 ± 0.3
w/o $\mathcal{L}_{ce}$	61.1 ± 0.3	76.1 ± 0.3	69.9 ± 0.5
w/o noisy label init	55.0 ± 0.3	67.1 ± 0.4	60.3 ±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

Auxiliary Experiments

How does the three phrase training impacts the performance?



How robust is our method training on more noisy data?



Treat samples with one type label as "clean" data as most of them are correct.

Conclusion

- A probabilistic automatic relabeling method, verified on fine-grained entity typing.
- Do not rely on extra prerequisite or supervision.
- Backbone and task agnostic, making it general and flexible.