Learning from Noisy Labels for Entity-Centric Information Extraction

Wenxuan Zhou, Muhao Chen
University of Southern California

Problem

Background:
Labeling on large corpora inevitably introduces noisy labels, which leads to degradation of model performance.

Goal:
develop a model that is robust to noisy training labels.

Motivation

Properties of noisy labels:
1. Take longer time to be learned.
2. Frequently forgotten by model.

Model prediction is often inconsistent or oscillates on noisy labels in later epochs.

Given $M(\geq 2)$ models, (with high probability) at least one prediction is inconsistent to the noisy label.

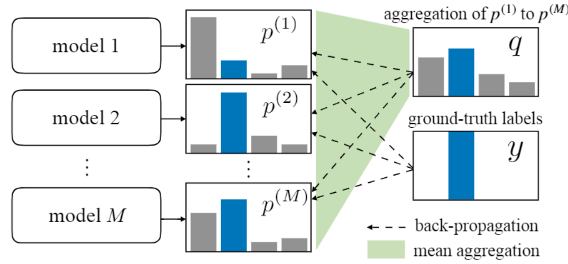
Co-regularization Framework

Step 1: Create $M(\geq 2$; 2 is enough) identical neural models with **different initialization**.

Step 2: Train the models with **the task loss** for certain steps (warm-up phase).

Step 3: Train the models with **the task loss** and an additional **agreement loss.**

Step 4: Return a random model.



Agreement loss:

$$\boldsymbol{q}_i = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{p}_i^{(k)},$$

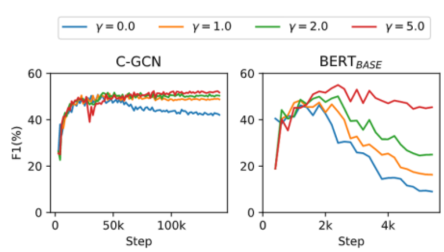
$$d(\boldsymbol{q}_i || \boldsymbol{p}_i^{(k)}) = \sum_{j=1}^{C} \boldsymbol{q}_{ij} \log\left(\frac{\boldsymbol{q}_{ij}+\epsilon}{\boldsymbol{p}_{ij}+\epsilon}\right),$$

$$\mathcal{L}_{\text{agg}} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{k=1}^{M} d(\boldsymbol{q}_i || \boldsymbol{p}_i^{(k)}),$$

KL div between avg model prediction and all model predictions

Experiments

Main Task (Bert-large)	TACREV	CoNLL03
Baseline	77.9	92.2
CrossWeigh	79.8	92.5
Co-regularization	**82.0**	**92.8**



Test F1 on noisy labels

Flipped labels (%)	10	30	50	70	90
BERT$_{\text{BASE}}$	74.2	70.8	62.9	48.6	0
BERT$_{\text{BASE}}$-CrossWeigh	77.3	75.6	71.6	61.3	25.1
BERT$_{\text{BASE}}$-CR	79.3	78.3	73.2	63.5	34.1
BERT$_{\text{BASE}}$ w/o flipped labels	76.5	74.9	72.9	70.8	57.4

TACREV test F1 w.r.t. different noise rates