Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training

EMNLP 2021
발표: Baek, Hyeongryeol

Abstract

1. Task: Distantly-supervised Named Entity Recognition
2. Contribution
    a. Noise robust learning
    b. Self-training methods

NER

1. Named Entity: a named entity is a real-world object, such as **a person, location, organization, product, etc.,** that can be denoted with a proper name.
2. Distantly Labeled: matching **entity mentions** in the target corpus with **typed entities** in **external gazetteers** or **knowledge bases**.
   a. gazetteers: **a collection of common entity names**, e.g., Random Name, US First Names Database , Word Lists, etc. for the type PER*
   b. Knowledge base (KB): A knowledge base is any system in which **knowledge is stored**, maintained, and accessed.

*Source: Chan Hee Song, Dawn Lawrie, Tim Finin, & James Mayfield. (2020). Improving Neural Named Entity Recognition with Gazetteers.

KB example

- Predicate: 술어

e.g. In the sentence "We went to the airport", "**went to the airport**" is the predicate.

*Source: https://slideplayer.com/slide/17212176/

Distantly-labeled data

- Problem: incomplete and noisy entity labels -> yielding deteriorated performance

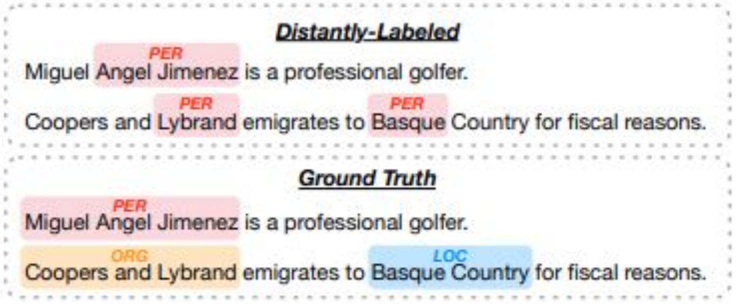


Figure 1: Distant labels obtained with knowledge bases may be incomplete and noisy, resulting in wrongly-labeled tokens.

Contribution

1. Noise-robust learning scheme
   a. Introducing noise-robust loss function
   b. Removing a noisy label
2. Proposing unsupervised contextualized augmentation approach
   a. Novel self-training methods

*self-training: “Self-training is a simple semi-supervised learning approach: Unlabelled examples that attract high-confidence predictions are labeled with their predictions and added to the training set”
(Sahito, A., Frank, E., & Pfahringer, B. (2021). Better Self-training for Image Classification through Self-supervision.)

1. Noise-robust learning scheme

   a. Introducing noise-robust loss function: giving **less weights** to tokens on which the model prediction is **less consistent with the given labels**

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{n} \log f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta}),$$

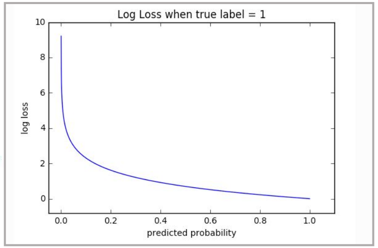
$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{n} \frac{\nabla_{\boldsymbol{\theta}} f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})}{\boxed{f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})}}.$$

$$\frac{\partial L}{\partial O_n} = -\sum_k y_k \frac{\partial \log(P(O_k))}{\partial O_n}$$

$$= -\sum_k y_k \frac{\partial \log(P(O_k))}{\partial P(O_k)} \times \frac{\partial P(O_k)}{\partial O_n}$$

$$= -\sum_k y_k \frac{1}{P(O_k)} \times \boxed{\frac{\partial P(O_k)}{\partial O_n}}$$

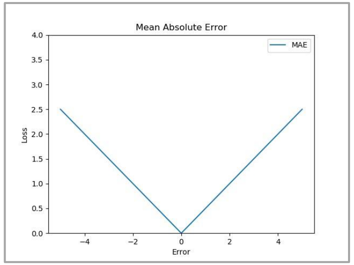
$$\begin{cases} p_n(1-p_n) & (n=k) \\ -p_n p_k & (n\neq k) \end{cases}$$



CE
: Less consistent with given y_i is weighed more -> sensitive to noisy labels

$$\mathcal{L}_{\mathrm{MAE}} = \sum_{i=1}^{n} \left(1 - f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})\right),$$

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{MAE}} = -\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta}).$$



MAE
: noise-tolerant but worsening convergence

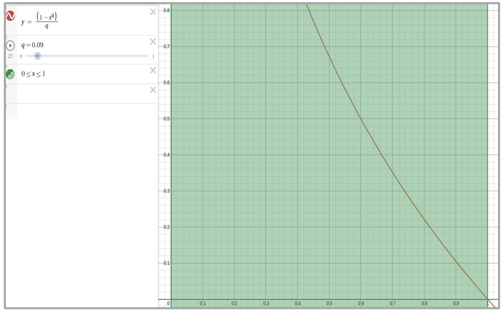
<Cross entropy loss vs Mean absolute error>

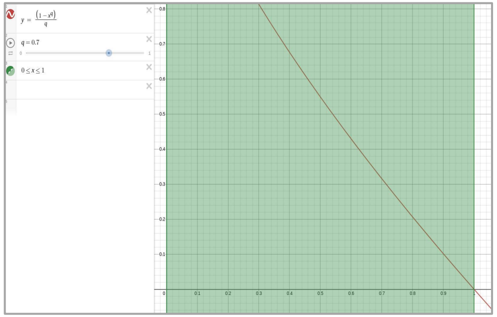
1. Noise-robust learning scheme
    a. Introducing noise-robust loss function: giving **less weights** to tokens on which the model prediction is **less consistent with the given labels**
        - GCE vs CE: more noise robust
        - GCE vs MAE: giving more attention to difficult tokens

$$\mathcal{L}_{\mathrm{GCE}} = \sum_{i=1}^{n} \frac{1 - f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})^q}{q}, \qquad (3)$$

where $0 < q < 1$ is a hyperparameter: When $q \to 1$, $\mathcal{L}_{\mathrm{GCE}}$ approximates $\mathcal{L}_{\mathrm{MAE}}$; when $q \to 0$, $\mathcal{L}_{\mathrm{GCE}}$ approximates $\mathcal{L}_{\mathrm{CE}}$ (using L'Hôpital's rule; see Appendix A for the derivation). The gradient is computed as:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{GCE}} = -\sum_{i=1}^{n} \frac{\nabla_{\boldsymbol{\theta}} f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})}{f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})^{1-q}}. \qquad (4)$$





<Generalized Cross entropy>

1. Noise-robust learning scheme
   b. Removing a noisy label: excluding tokens with lower prediction probability than predefined threshold ($f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta}) \leq \tau$ where τ is a threshold value)

$$\mathcal{L}_{\mathrm{GCE}} = \sum_{i=1}^{n} w_i \frac{1 - f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})^q}{q}, \qquad (5)$$

where $w_i = 1$ at the start of training and is periodically updated once every several batches as $w_i = \mathbb{1}\left(f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta}) > \tau\right)$, where $\mathbb{1}(\cdot)$ is the indicator function.

<Final GCE>

1. Noise-robust learning scheme
   + Model ensemble: multiple models' prediction are likely to be consistent on clean data. Training Model_ENS using K (K=5) models' prediction.

$$\mathcal{L}_{\text{ENS}} = \sum_{i=1}^{n} \text{KL}\left(\bar{f}_i\left(\boldsymbol{x}; \{\boldsymbol{\theta}_k\}_{k=1}^{K}\right) \| f_i(\boldsymbol{x}; \boldsymbol{\theta}_{\text{ENS}})\right), \tag{6}$$

where $\bar{f}_i\left(\boldsymbol{x}; \{\boldsymbol{\theta}_k\}_{k=1}^{K}\right) = \frac{1}{K}\sum_{k=1}^{K} f_i(\boldsymbol{x}; \boldsymbol{\theta}_k)$ is the K models' averaged prediction, and we find that $K = 5$ is sufficient to provide stable ensembled model performance.

2. Proposing unsupervised contextualized augmentation approach

   a. Novel self-training methods

   - Augmented seq.: token-level perturbation; but likely to be preserving entity type

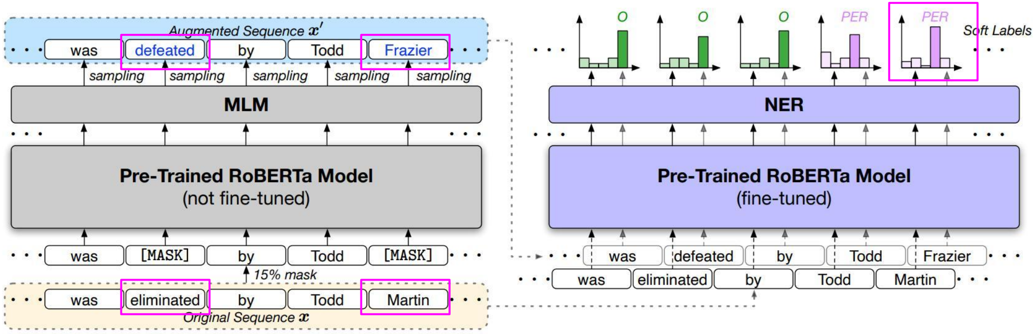


Figure 2: Overview of language model augmented self-training. Only a part of the sequence is shown; the original sequence is "Renzo Furlan was eliminated by Todd Martin in the tournament." We feed the partially masked original sequence into a pre-trained RoBERTa model and sample from its MLM output probability to obtain an augmented sequence (replaced tokens are marked in blue). Then the NER model is trained with both original and augmented sequences as inputs to approximate the soft labels.

*self-supervised learning: Self-supervised Learning is an unsupervised learning method where the supervised learning task is created out of the unlabelled input data. (https://towardsdatascience.com/self-supervised-learning-methods-for-computer-vision-c25ec10a91bd#:~:text=Self%2Dsupervised%20learning%20is%20a,are%20based%20on%20contrastive%20learning.)

2. Proposing unsupervised contextualized augmentation approach

   a. Novel self-training methods

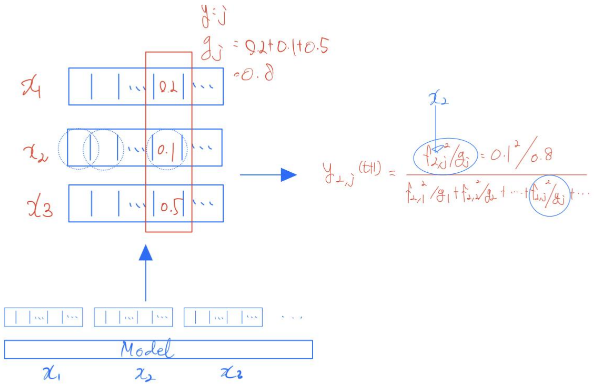
   - Soft Labels: enhancing high-confidence predictions while demote low-confidence ones

$$y_{i,j}^{(t+1)} = \frac{f_{i,j}\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)^2 / g_j}{\sum_{j'}\left(f_{i,j}\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)^2 / g_{j'}\right)}, \qquad (8)$$

$$g_j = \sum_i f_{i,j}\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right).$$

Then the model $\boldsymbol{\theta}^{(t+1)}$ at the next iteration is updated by approximating the soft labels with both the original sequence and the augmented sequence as inputs, via the KL divergence loss:

$$\mathcal{L}_{\mathrm{ST}} = \sum_{i=1}^{n} \mathrm{KL}\left(y_i^{(t+1)} \middle\| f_i\left(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}\right)\right) + \sum_{i=1}^{n} \mathrm{KL}\left(y_i^{(t+1)} \middle\| f_i\left(\boldsymbol{x}';\boldsymbol{\theta}^{(t+1)}\right)\right). \qquad (9)$$



*self-supervised learning: Self-supervised Learning is an unsupervised learning method where the supervised learning task is created out of the unlabelled input data. (https://towardsdatascience.com/self-supervised-learning-methods-for-computer-vision-c25ec10a91bd#:~:text=Self%2Dsupervised%20learning%20is%20a,are%20based%20on%20contrastive%20learning.)

2. Proposing unsupervised contextualized augmentation approach

   a. Novel self-training methods

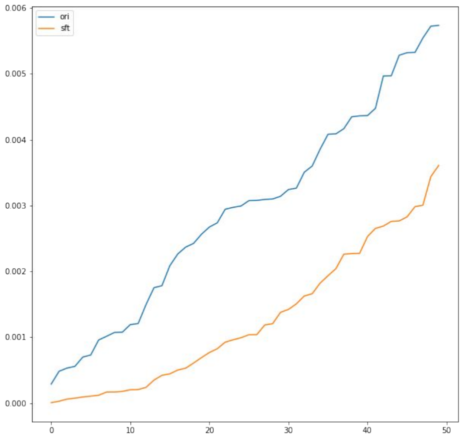
   - Soft Labels: enhancing high-confidence predictions while demote low-confidence ones

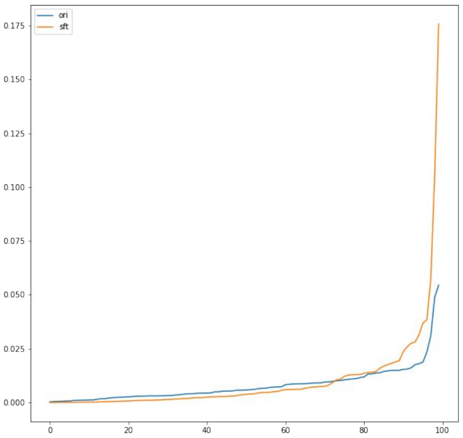
$$y_{i,j}^{(t+1)} = \frac{f_{i,j}\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)^2 / g_j}{\sum_{j'}\left(f_{i,j}\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)^2 / g_{j'}\right)}, \quad (8)$$

$$g_j = \sum_i f_{i,j}\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right).$$

Then the model $\boldsymbol{\theta}^{(t+1)}$ at the next iteration is updated by approximating the soft labels with both the original sequence and the augmented sequence as inputs, via the KL divergence loss:

$$\mathcal{L}_{\mathrm{ST}} = \sum_{i=1}^{n} \mathrm{KL}\left(y_i^{(t+1)} \middle\| f_i\left(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}\right)\right) + \sum_{i=1}^{n} \mathrm{KL}\left(y_i^{(t+1)} \middle\| f_i\left(\boldsymbol{x}';\boldsymbol{\theta}^{(t+1)}\right)\right). \quad (9)$$





*self-supervised learning: Self-supervised Learning is an unsupervised learning method where the supervised learning task is created out of the unlabelled input data. (https://towardsdatascience.com/self-supervised-learning-methods-for-computer-vision-c25ec10a91bd#:~:text=Self%2Dsupervised%20learning%20is%20a,are%20based%20on%20contrastive%20learning.)

Experiments

Ablations	Pre.	Rec.	F1
RoSTER	0.859	0.849	0.854
w/o GCE	0.817	0.843	0.830
w/o NR	0.830	0.836	0.833
w/o ST	0.844	0.812	0.828

Table 3: Ablation study on CoNLL03 dataset. We compare our full method with ablations (see texts for the abbreviation meanings).

Ablations	Mean (Std.) F1
w. ensemble	0.828 (0.009)
w/o ensemble	0.817 (0.025)

Table 4: Mean and standard deviation (std.) F1 scores of 5 runs (before self-training) with and without model ensemble on CoNLL03 dataset.

Distant Match: Shanghai-Ek [Chor]$_{\text{PER}}$ is jointly owned by the Shanghai Automobile Corporation and [Ek Chor]$_{\text{PER}}$ China Motorcycle.
Ground Truth: [Shanghai-Ek Chor]$_{\text{ORG}}$ is jointly owned by the [Shanghai Automobile Corporation]$_{\text{ORG}}$ and [Ek Chor China Motorcycle]$_{\text{ORG}}$.

AutoNER: Shanghai-Ek [Chor]$_{\text{PER}}$ is jointly owned by the Shanghai Automobile Corporation and [Ek Chor]$_{\text{PER}}$ [China]$_{\text{LOC}}$ Motorcycle.
BOND: [Shanghai-Ek Chor]$_{\text{PER}}$ is jointly owned by the [Shanghai]$_{\text{LOC}}$ [Automobile Corporation]$_{\text{ORG}}$ and [Ek Chor]$_{\text{PER}}$ [China Motorcycle]$_{\text{ORG}}$.
RoSTER: [Shanghai-Ek Chor]$_{\text{ORG}}$ is jointly owned by the [Shanghai Automobile Corporation]$_{\text{ORG}}$ and [Ek Chor China Motorcycle]$_{\text{ORG}}$.

Table 6: Case study with **RoSTER** and baselines. The sentence is from CoNLL03.

Model Components
: Generalized cross entropy, noisy label removal, self-training, ensemble

*self-training: “Self-training is a simple semi-supervised learning approach: Unlabelled examples that attract high-confidence predictions are labeled with their predictions and added to the training set”
(Sahito, A., Frank, E., & Pfahringer, B. (2021). Better Self-training for Image Classification through Self-supervision.)

Conclusion

- studying the distantly-supervised NER problem without using any human annotations but only distantly-labeled data
- proposing a **noise-robust learning scheme**, consisting of a new loss function and a noisy label removal step.
- proposing a **self-training method** that guides model refinement with its own **high-confidence predictions** and enforces the model to make consistent predictions on original and augmented sequences generated by PLMs