Noise-Robust Training with Dynamic Loss and Contrastive Learning for Distantly-Supervised Named Entity Recognition

Zhiyuan Ma and **Jintao Du** and **Shuheng Zhou**
Tiansuan Lab, Ant Group Co., Ltd.
{mazhiyuan.mzy,lingke.djt,shuheng.zsh}@antgroup.com

Abstract

Distantly-supervised named entity recognition (NER) aims at training networks with distantly-labeled data, which is automatically obtained by matching entity mentions in the raw text with entity types in a knowledge base. Distant supervision may induce incomplete and noisy labels, so recent state-of-the-art methods employ sample selection mechanism to separate clean data from noisy data based on the model's prediction scores. However, they ignore the noise distribution change caused by data selection, and they simply excludes noisy data during training, resulting in information loss. We propose to (1) use a dynamic loss function to better adapt to the changing noise during the training process, and (2) incorporate token level contrastive learning to fully utilize the noisy data as well as facilitate feature learning without relying on labels. We conduct extensive experiments on multiple datasets and our method obtains 4.3%, 1.5%, 0.9% F1 score improvements over the current state-of-the-art on Wikigold, CoNLL03 and OntoNotes5.0.

1 Introduction

Named entity recognition (NER) is a fundamental task in natural language processing, which aims at locating entity mentions in a given sentence and assign them to certain types, and it has a wide range of applications (Khalid et al., 2008; Etzioni et al., 2005; Aramaki et al., 2009; Bowden et al., 2018). In recent years, deep neural networks have achieved great performance on NER due to their strong ability to learn from large amount of labeled data. However, the acquisition of abundant high-quality human annotated data is costly and difficult. A major solution to this problem is distant supervision, a method for automatic generation of entity labels, the common practice of which is to match entity mentions in an unlabeled dataset with typed entities in external gazetteers or knowledge bases. Unfortunately, this method inevitably introduces noise while generating labeled dataset, leading to a deterioration of the NER models' performance.

There are many works that try to improve the performance of NER networks on distantly supervised dataset with the existence of such noise. Some studies use training tricks like applying early stopping (Liang et al., 2020) and labeling entities with multiple types (Shang et al., 2018) to handle the noise, and some studies build an additional classification model to distinguish noisy labels from the ground truth labels (Onoe and Durrett, 2019) relying on additional labeled data. Sample separation (Li et al., 2020; Yu et al., 2019) is a dominant method in noise-robust learning, trying to filter out the noisy samples from the clean ones based on the small-loss criterion (Li et al., 2020). RoSTER (Meng et al., 2021) applies sample separation to distantly-supervised NER, and they also uses generalized cross entropy (GCE) (Ghosh et al., 2017) loss for noise-robust training.

In this paper, we also use the framework of sample separation and propose a distantly-supervised NER training scheme that uses dynamic GCE loss to optimize the model training, and we incorporate contrastive learning to lower the risk of noisy label overfitting and fully utilize the data with untrusted labels. The contributions of this paper are as follows:

1. We propose to use dynamic GCE loss during training with sample separation steps, and we adjust the loss function automatically based on the prediction entropy, which benefits the training process with changing noise distribution.

2. We propose to apply contrastive learning to facilitate feature learning without relying on labels, so that the risk of noisy label overfitting can be mitigated and the noisy data can be fully utilized.

3. We conduct experiments on three benchmark datasets and our method outperforms existing distantly-supervised NER approaches by significant margins.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10119–10128

2 Methodology

In this section, we (1) briefly describe how to obtain distantly-labeled data, (2) introduce our noise-robust learning scheme with dynamic GCE loss and (3) present how to incorporate contrastive learning into our framework. We use the pre-trained RoBERTa (Liu et al., 2019) as our backbone model, but our proposed methods can be integrated with other architectures as well. The overall framework of our noise-robust training is showed in Figure 1.

2.1 Distant Label Generation

Distant labels can be obtained by matching entities in an unlabeled corpus with those in external knowledge bases or gazetteers with typing information. In this work, instead of introducing new distant label generation methods, we follow the previous work (Liang et al., 2020; Meng et al., 2021) for these steps: (1) potential entities are determined via POS tagging and hand-crafted rules, (2) their types are acquired by querying Wikidata using SPARQL (Vrandecic and Krötzsch, 2014), and (3) additional gazetteers from multiple online resources are used for matching more entities in the corpus.

2.2 Noise-Robust Learning with Dynamic GCE loss

We apply sample separation steps to remove wrong labels and handle noise. Specifically, we decide which labels are eliminated based on the model prediction: At first, all tokens participate in the training process; later, those tokens whose distant label does not agree with the model prediction (i.e., $f_{i,y_i}(x;\theta) <= \tau$ where f_{i,y_i} is the model's predicted probability of token x_i belonging to the label class, θ is the model parameter and τ is a threshold value) will be excluded from the training set.

GCE loss. The purpose of NER is to classify each token in a sentence to a tag, and cross entropy (CE) loss is most commonly used for such a purpose:

$$L_{CE} = -\sum_{i=1}^{n} log f_{i,y_i}(x,\theta) \quad (1)$$

The logarithmic function in CE loss makes the tokens on which the model's prediction is less congruent with the provided labels be weighted more during the gradient update. This mechanism grants better model convergence, but also brings more attention to noisy labels when the dataset is not clean. The mean absolute error (MAE) loss has been shown inherently noise-tolerant when used for classification (Ghosh et al., 2017) and is defined as follows

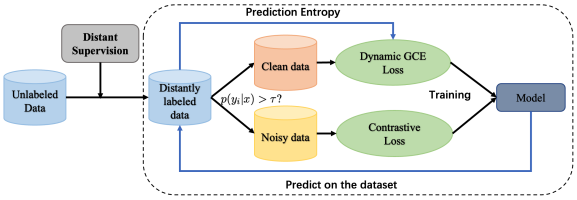

Figure 1: The overall framework of our method.

$$L_{MAE} = \sum_{i=1}^{n}(1 - f_{i,y_i}(x,\theta)) \quad (2)$$

However, MAE loss treats every token equally for gradient update, and this mechanism is not suitable for deep learning, causing lower convergence efficiency and suboptimal model performance. (Zhang and Sabuncu, 2018) proposes generalized cross entropy (GCE) loss to balance between model convergence and noise-robustness, which is defined as follows

$$L_{GCE} = \sum_{i=1}^{n} \frac{1 - f_{i,y_i}(x,\theta)^q}{q} \quad (3)$$

where

$$0 < q < 1$$

is a hyperparameter: When $q \to 1$, L_{GCE} approximates L_{MAE}; when $q \to 0$, L_{GCE} approximates L_{CE} (using L'Hôpital's rule). RoSTER (Meng et al., 2021) is the first to use GCE loss in distantly-supervised NER and it fixes q during training.

Dynamic GCE loss. However, we argue that the static GCE loss is not suitable for a training process with change of noise distribution caused by sample selection, so we propose to use dynamic GCE loss in our training scheme. We perform noisy sample removal certain times during training. At first, all tokens along with their distant labels are used in the model training, bringing a lot of noise. Later, the training set is dynamically adjusted according to the consistency of tokens' distant labels and their prediction scores, and noise labels are gradually discarded from the training set as training progresses. Therefore, the balance between noise-robustness and model convergence is constantly changing, and the GCE loss should also be adjusted accordingly. Specifically, the GCE loss should approximates MAE loss first as the

noise is greatest at the beginning of the training process; then it should gradually change towards the CE loss as our noisy sample removal strategy decreases noisy labels. In practice, we can make the hyperparameter q drop stepwise from 1 to 0, but we propose a more flexible method that dynamically adjusts q based on the prediction entropy.

The model's prediction of a token's label is a probability distribution, and the entropy of this distribution reflects the uncertainty of the model prediction. We average this entropy over all data to get the prediction entropy of the model on this dataset. The prediction entropy gradually decreases during the training process as the model tends to be more and more confident of its predictions. We let q vary according to the following formula

$$q = q_0 + \frac{lnE - lnE_0}{N} \tag{4}$$

where q_0 is the initial value of q which we set to 1 in our experiments, E is the prediction entropy, E_0 is the initial value of E, and N is the total number of entity types. Practically, q will decrease along with E, and the loss function will trend from higher noise robustness to better model convergence.

2.3 Noise-Robust Contrastive Learning

Generally, a noise-robust learning scheme using sample separation treats samples with untrusted labels as unlabeled data. The unlabeled data may be excluded during training, or pseudo-labels can be generated for the data. But pseudo-labels introduce new noise into the training set, and simply ignoring the unlabeled data leads to information loss, especially in distant supervision where noisy labels are generated because some entities are ambiguous or not covered by the knowledge base. These entities are harder to learn and contain more useful information. Besides, the clean data selected based on the model predictions may still contain noisy labels, so the risk of noisy label overfitting still exists.

To handle these problems, we propose to apply contrastive learning to the unlabeled tokens, which can boost the model performance for two reasons. First, contrastive learning facilitates feature learning without relying on labels, which further mitigates the risk of noisy label memorization since it does not rely on imperfect separation of clean and noisy samples as well as incorrect pseudo-labels generated during training (Karim et al., 2022). Second, contrastive learning is a perfect way to fully utilize unlabeled data as it does not require any pseudo-labels, avoiding introducing new noise. Practically, for each token with untrusted label, we perturb its word embedding in two random directions to obtain two new vectors. The two vectors are augmented copies of the original word embedding, and although they cannot be mapped back to any real words, we argue that they have similar semantics because of the continuity of semantic spaces. It means that the two augmented copies should share the same label although we do not know what the label is since the original label is not trusted. Therefore, we employ the projection head $g(.;\phi)$ to obtain feature projections $z_i = g(f(x_{i,1}, \theta), \phi)$, and $z_j = g(f(x_{i,2}, \theta), \phi)$ of the differently augmented copies $(x_{i,1}, x_{i,1})$ of one input x_i, and they are positive samples for contrastive learning. The contrastive loss function can be expressed as

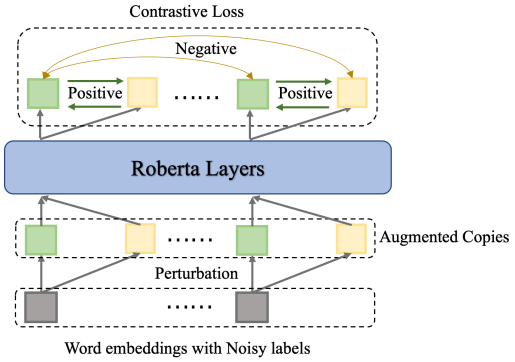

Figure 2: The illustration of our contrastive learning.

$$l_{i,j} = -log\frac{exp(sim(z_i, z_j)/\kappa)}{\sum_{b \neq i} exp(sim(z_i, z_b)/\kappa)} \tag{5}$$

$$L_C = \frac{1}{2B}\sum_{b=1}^{2B}[l_{2b-1,2b}, l_{2b,2b-1}] \tag{6}$$

where κ is a temperature constant, B is the number of samples in mini-batch, and $sim(z_i, z_j)$ can be expressed as the cosine similarity between z_i and z_j. The illustration of our token level contrastive learning method is presented in Figure 2.

The total loss function we minimize is

$$L = L_{GCE} + \lambda_C L_C \tag{7}$$

where λ_C is contrastive loss coefficient.

2.4 Co-Teaching

Apart from noise-robust learning, we also use model ensemble and self-training like RoSTER

	Methods	CoNLL03			OntoNotes5.0			Wikigold		
		Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Distant-Sup.	**Distant Match**	0.811	0.638	0.714	0.745	0.693	0.718	0.479	0.476	0.478
	Distant RoBERTa	0.837	0.633	0.721	0.769	0.715	0.737	0.603	0.532	0.565
	AutoNER	0.752	0.604	0.670	0.731	0.712	0.721	0.435	0.524	0.475
	BOND	0.821	0.809	0.815	**0.774**	0.701	0.736	0.534	0.686	0.600
	RoSTER	0.859	0.849	0.854	0.753	0.789	0.771	0.649	0.710	0.678
	DGCE&CL (Ours)	**0.875**	**0.862**	**0.869**	0.770	**0.790**	**0.780**	**0.676**	**0.773**	**0.721**

Table 1: Performance of all methods on three datasets measured by precision (Pre.), recall (Rec.) and F1 scores. Baseline results are reported by (Meng et al., 2021).

(Meng et al., 2021) to further improve the model performance. However, instead of training N models $\{\theta_k\}_{k=1}^{N}$ independently as in RoSTER, we propose a co-teaching (Han et al., 2018) framework in which the N models select training data not for themselves but for one another. Specifically, all models are trained simultaneously with different random seeds. At the sample separation steps mentioned above, each model selects data based on the label consistency with the model prediction from the original dataset as clean data, which will be used by another model for training in the next training stage.

If a model has already overfit some noisy data, it will treat them as clean data in the next sample separation step. Therefore, using its own selected data in a model's next training stage would cause noise enhancement. Co-teaching can alleviate the problem of noise enhancement, so we propose to combine our learning scheme with the co-teaching framework. Other implementation details are presented in Section 3.3.

3 Experiments

3.1 Datasets

We conduct experiments on three NER datasets: CoNLL03 (Sang and Meulder, 2003), OntoNotes5.0 (Weischedel et al., 2013) which we follow the pre-processing of (Chiu and Nichols, 2016), and Wikigold (Balasuriya et al., 2009). Details are presented in Appendix A.

3.2 Baselines

We compare our method with a wide range of distantly supervised methods using the distantly-labeled training set obtained as in (Meng et al., 2021): **Distant Match** uses the distant supervision to generate predictions. **Distant RoBERTa** (Liu et al., 2019) fine-tunes a pre-trained RoBERTa model on distantly-labeled data as if they are ground truth. **AutoNER** (Shang et al., 2018) uses a new tagging scheme that assigns ambiguous tokens with all possible labels. **BOND** (Liang et al., 2020) first trains a RoBERTa model on distantly-labeled data with early stopping, and then uses a teacher-student framework to iteratively self-train the model. **RoSTER** (Meng et al., 2021) proposes a noise-robust learning framework with GCE loss and label removal steps, followed by a self-training method that uses contextualized augmentations created by pre-trained language models.

3.3 Implementation Details

We use the pre-trained RoBERTa-base model as our backbone model. For the three datasets CoNLL03, OntoNotes5.0, and Wikigold, the maximum sequence lengths are set to be 150, 180, and 120 tokens; E_0 in Eq. 4 are 2, 1.5 and 0.24; the noise-robust training epochs are set to be 3, 5, 5; the temperature in contrastive learning is set to be 1, 0.5, 0.5. For all three datasets: The training batch size is 32 and the number of models for ensemble $K = 5$. The threshold value τ for sample separation is 0.7. We use Adam (Kingma and Ba, 2015) as the optimizer, and the peak learning rate is $3e{-}5$, $1e-5$, $5e-7$ for noise-robust training, ensemble model training and self-training respectively with linear decay. The warmup proportion is 0.1. We train the model on 1 NVIDIA A100 Tensor Core GPU.

3.4 Main Results

Table 1 presents the performance of all methods measured by precision, recall and F1 scores. On all datasets, our method with Dynamic GCE loss and Contrastive Learning (DGCE&CL) achieves the best performance among distantly-supervised methods. RoSTER applies the original GCE loss and noisy label removal, and it achieves better re-

Methods	CoNLL03	OntoNotes5.0	Wikigold
DGCE&CL	**0.869**	**0.780**	**0.721**
w/o CL	0.863 (0.6%↓)	0.774 (0.6%↓)	0.710 (1.1%↓)
w/o DL	0.857 (1.2%↓)	0.772 (0.8%↓)	0.693 (2.8%↓)
w/o CT	0.860 (0.9%↓)	0.778 (0.2%↓)	0.704 (1.7%↓)

Table 2: The F1 Scores of different variants in the ablation study.

Methods	CoNLL03	OntoNotes5.0	Wikigold
DGCE&CL	20.466	23.500	29.640
w/o CL	11.163	13.435	27.235
w/o DL	5.581	7.101	13.167
w/o CT	6.928	18.448	25.342

Table 3: The t-values of our method and its variants in the significance test.

sults than other baselines, implying that the sample separation strategy and a noise-robust loss are useful in distantly-supervised NER. Our method consistently outperforms RoSTER, showing the superiority of our proposed dynamic GCE loss and token level contrastive learning when trained on distantly-labeled data. Specifically, our method achieves 4.3%, 1.5%, 0.9% absolute F1 scores gain over RoSTER on Wikigold, CoNLL03 and OntoNotes5.0. For OntoNotes5.0, the precision of BOND is a little higher than our method, but we have much better recall rate, mainly because our noise-robust training can learn entities that are not covered by the knowledge base.

3.5 Ablation Study

To verify the validity of different modules in our proposed method, we introduce the following variants of our method to further carry out an ablation study: 1) w/o CL (Contrastive Learning), which removes the contrastive learning and simply ignores the noisy data during training; 2) w/o DL (Dynamic Loss), which removes the dynamic loss and only uses a static GCE loss with q fixed as 0.7 like RoSTER does; 3) w/o CT (Co-Teaching), which removes the co-teaching strategy and let each model select its own training data.

The performance of each variant is shown in Table 2. From the listed results we can see that all modules in our proposed method can boost the model performance and removing any of them leads to a performance drop. Specifically, removing the dynamic loss always brings the biggest decrease of F1 scores, indicating that dynamically adjusting the hyperparameter q of GCE loss is very important in distantly-supervised NER, and our tuning strategy based on the prediction entropy is effective. Removing the contrastive learning also deteriorates the model performance, showing its effectiveness in feature learning. Training independent models without co-teaching also leads to suboptimal results on all three datasets, showing that co-teaching can further mitigate the noise label memorization.

3.6 Significance Test

To verify the significance of the advantages of our proposed method over the baseline, we conduct a t-test to check the statistical significance and report the results on Table 3. Specifically, we unfreeze the random seed and repeat each experiment five times, and the t-test shows the results of our method are higher than the baseline with a 99% confidence interval ($\alpha = 0.01$) on all datasets. The significance test experiments are conducted on both our DGCE&CL and the model variants from ablation study, and the t-values of t-test on all datasets are shown in Table 3.

4 Conclusion

In this paper, we study the distantly-supervised NER problem. We propose a noise-robust training scheme using dynamic loss to adapt to changing noise distributions caused by sample selection mechanism. To further improve the model's generalization and robustness, we incorporate contrastive learning to facilitate feature learning without relying on labels and fully utilize the data with noisy labels. We conduct experiments on three datasets and our method outperforms all previous distantly-supervised NER methods.

Limitations

Sample separation based on model predictions can only eliminate part of the noise, and it costs extra time in training. Moreover, although our dynamic GCE loss based on prediction entropy works well in distantly-supervised NER, Eq. 4 is determined mainly because it has superior experiment results and it lacks theoretical proof.

Acknowledgements

This research work has been sponsored by Ant Group Security and Risk Management Fund.

References

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, and Kazuhiko Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Named entity recognition in wikipedia. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (People's Web)*, pages 10–18.

Kevin Bowden, JiaQi Wu, Shereen Oraby, Amita Misra, and Marilyn A. Walker. 2018. Slugnerds: A named entity recognition tool for open domain dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Trans. Assoc. Comput. Linguistics*, 4:357–370.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. 2022. UNICON: combating label noise through uniform selection and contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9666–9676. IEEE.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: bert-assisted open-domain named entity recognition with distant supervision. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1054–1064. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10367–10378. Association for Computational Linguistics.

Yasumasa Onoe and Greg Durrett. 2019. Learning to denoise distantly-labeled data for entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2407–2417. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*. ACL.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 2054–2064. Association for Computational Linguistics.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7164–7173. PMLR.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8792–8802.

A Datasets and Metrics

We conduct experiments on three datasets, and the dataset statistics are shown in Table 4. All datasets are in English language, and the entity type numbers of CoNLL, OntoNotes5.0 and Wikigold are 4, 18 and 4. Following previous work (Sang, 2002), we calculate the entity level F1-score and use the Micro F1 over all entity types as the metric of evaluation. We train 5 models with different random seeds and report the F1-score of their ensemble.

B Strategies for Tuning q

As mentioned above, we propose to dynamically change the hyperparameter q in Eq. 3, and in this section we compare different strategies for tuning q. q is changed at regular intervals of training process, and (1) a straightforward way is decreasing q stepwise from 1 to 0. (2) The second strategy also reduces q by the same value each time, but we regard the value of each reduction as a hyperparameter and search for its optimal value. This strategy has better performance but it requires lots of extra experiments to find an optimal parameter. (3) The third way is to use Eq. 4 to automatically adjust q. 4) We also use a fixed q as a baseline.

Dataset	Type	Train	Test
CoNLL03	4	14,041	3,453
OntoNotes5.0	18	59,924	8,262
Wikigold	4	1,142	274

Table 4: Dataset statistics with the number of entity types and the number of training/test sequences.

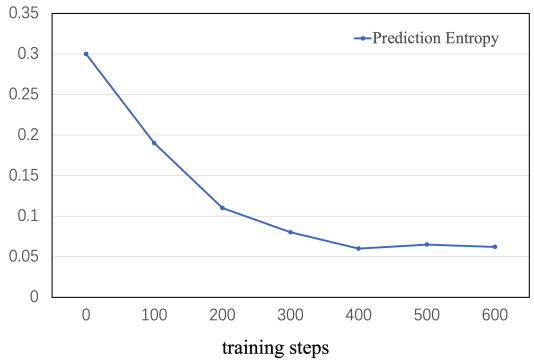

Figure 3: The model's prediction entropy on Wikigold during training.

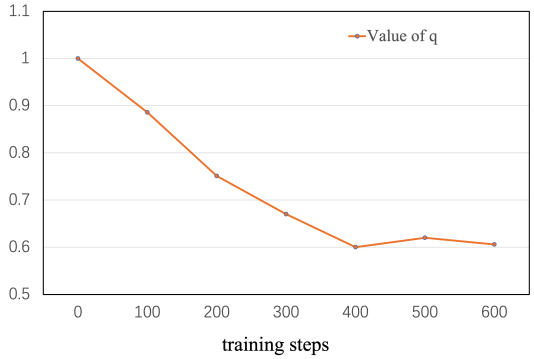

Figure 4: The changing line of q during training on Wikigold.

Methods	CoNLL03	OntoNotes5.0	Wikigold
Fixed q	0.854	0.771	0.678
Decreasing Stepwise	0.863	0.773	0.691
Optimal Reduction	0.870	0.778	0.723
Entropy Based	0.869	0.780	0.721

Table 5: The F1 scores of different strategies for tuning q on three datasets.

Table 5 presents the F1 scores of different strategies on three datasets, and we can draw the following conclusions. No matter what strategy is used to adjust q, dynamic GCE loss always performs better than static GCE loss, which shows the necessity of tuning q to adapt to the noise change caused by sample separation during training. Automatically tuning q based on the prediction entropy achieves competitive performance to decreasing q with the optimal reduction value each time, and no extra experiments are required.

We record the prediction entropy during model training on Wikigold and present the changing line in Figure 3. It can be seen that the prediction entropy drops sharply in the early period of training, but then it stabilizes and only fluctuates in a small range. Figure 4 presents the according changing line of q. It is close to a linear decline in the early stage, and also stabilizes afterwards. This is reasonable as we assume that the noise in training set gradually decreases in the early stage of training, and the loss function should be more stable and more conductive to model convergence later in the training process.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☒ A2. Did you discuss any potential risks of your work?
Our work is a foundational study on distantly-supervised NER, which includes no potential malicious or unintended harmful effects and uses.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2.4

☑ B1. Did you cite the creators of artifacts you used?
Section 2.4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 2.4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2.4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets we use are widely used in named entity recognition, and they contain no information that names or uniquely identifies individual people nor offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.1

C ☑ Did you run computational experiments?

Section 2.4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix A

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.