# Enhancing Distantly Supervised Named Entity Recognition with Strong Label Guided Lottery Training

**Zhiyuan Ma, Jintao Du, Changhua Meng, Weiqiang Wang**
Tiansuan Lab, Ant Group Co., Ltd.
{mazhiyuan.mzy, lingke.djt, changhua.mch, weiqiang.wwq}@antgroup.com

**Abstract**

In low-resource Named Entity Recognition (NER) scenarios, only a limited quantity of strongly labeled data is available, while a vast amount of weakly labeled data can be easily acquired through distant supervision. However, weakly labeled data may fail to improve the model performance or even harm it due to the inevitable noise. While training on noisy data, only certain parameters are essential for model learning, termed safe parameters, whereas the other parameters tend to fit noise. In this paper, we propose a noise-robust learning framework where safe parameters can be identified with guidance from the small set of strongly labeled data, and non-safe parameters are suppressed during training on weakly labeled data for better generalization. Our method can effectively mitigate the impact of noise in weakly labeled data, and it can be easily integrated with data level noise-robust learning methods for NER. We conduct extensive experiments on multiple datasets and the results show that our approach outperforms the state-of-the-art methods.

**Keywords:** Named Entity Recognition, Information Extraction, Weakly-supervised learning

## 1. Introduction

Named entity recognition (NER) is a fundamental task in natural language processing aiming 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). Nevertheless, the acquisition of abundant high-quality human annotated data is costly, and in many cases there is only a small amount of strongly labeled data. Fortunately, the entity labels can be automatically generated by distant supervision, the common practice of which is to match entity mentions in an unlabeled dataset with typed entities in external gazetteers or knowledge bases. However, this approach inevitably introduces label noise into the training set, which may lead to deterioration of the model performance without proper treatment. There are many methods proposed to improve the performance of NER networks on datasets with the existence of noise, such as sample separation (Li et al., 2020; Yu et al., 2019; Li et al., 2020; Meng et al., 2021) and training tricks like early stopping (Liang et al., 2020). Some works leverages the small set of human annotated data to handle the distantly supervised data more effectively, but their studies remain in the data level, such as building an additional classification model to distinguish noisy labels from the ground truth labels (Onoe and Durrett, 2019) or training a model on the strongly labeled data to revise the weak label (Jiang et al., 2021).

In this paper, we shift our focus to the model parameter level. Researches show that only certain parameters are essential for model learning during training on noisy data, termed safe parameters, whereas the other parameters tend to fit noise (Xia et al., 2021). Inspired by recent works (He et al., 2022; Chen et al., 2021a), we present a novel insight: rather than using the strongly labeled data to train a teacher model and generate pseudo labels, it is better to discover safe parameters relying on the limited trusted data. Driven by this insight, we propose a noise-robust learning framework consisting of three stages, where safe parameters can be identified with guidance from the small set of strongly labeled data. The contributions of our work can be summarized as follows:

1. We propose a novel framework for noise robust learning in low-resource NER scenarios. To our knowledge, it is the first time to solve this problem from the perspective of parameter level.

2. We propose a novel strategy to identify the safe parameters and introduce an effective optimization strategy to suppress the other parameters during distantly supervised training.

3. The results of extensive experiments show that our approach outperforms the state-of-the-art methods. Moreover, it can be easily integrated with data level noise-robust NER methods and further enhance their performance.

## 2. Methodology

### 2.1. Safe Parameters Learning

The concept of safe parameters comes from the lottery ticket hypothesis (LTH), which was originally proposed to advocate the existence of an independently trainable sparse sub-network from a dense network (Frankle and Carbin, 2019). LTH has been



*LREC-COLING 2024*, pages 5922–5929
20-25 May, 2024. 

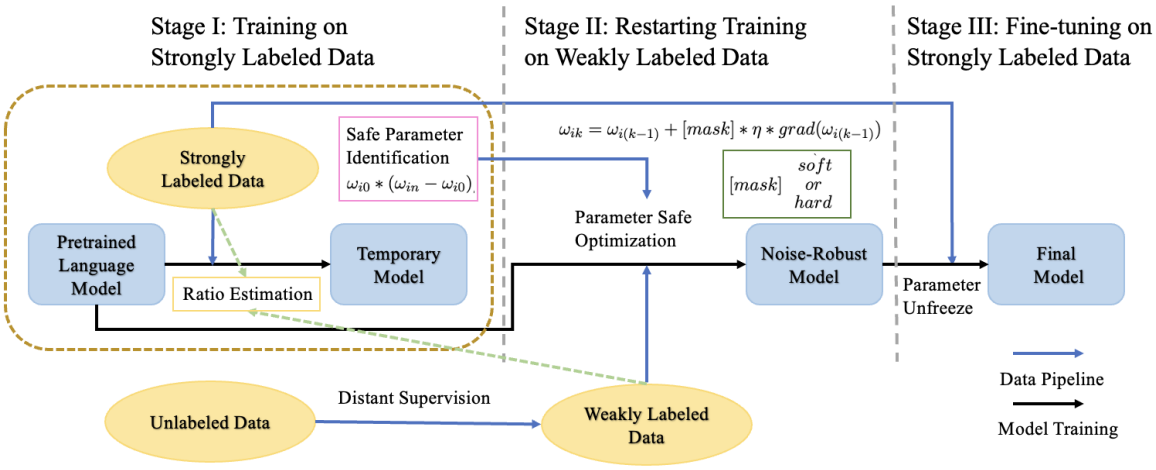


Figure 1: The overall framework of our method.

explored widely in numerous contexts, such as image classification (Chen et al., 2021a,b) and natural language processing (Chen et al., 2020). Xia et al. (2021) extend LTH to the field of noise-robust learning, suggesting that only partial parameters are crucial for model learning and generalization, while the other parameters tend to fit noisy labels.

For NER, large weakly labeled data can be easily acquired through distant supervision while human annotated data is usually limited. Inspired by the works of LTH, we propose a novel method to identify safe parameters with guidance of the limited strongly labeled data, and enhance distantly supervised NER with lottery training, where the non-safe parameters are suppressed.

## 2.2. Noise-Robust Learning Framework

We divide our learning approach into three stages as illustrated in Figure 1. In the first stage, we train the model merely on the small set of strongly labeled data, and we utilize the training information to identify the safe parameters. In the second stage, we reset the model parameters to their initial values, and train the model on the large distantly supervised data from scratch. We do not train from the temporary model checkpoint at the end of the first training stage because training on a very small clean dataset may result in the model getting trapped in a local optimum that is hard to escape from. In this stage, we apply a different optimization method and suppress the updating of non-safe parameters during training, which effectively mitigates the impact of noise and ultimately leads to better generalization. In the third stage, we unfreeze the non-safe parameters and fine-tune the model on the small set of strongly labeled data.

The importance of parameters is determined by two factors: the magnitude of the parameters and their gradients during the first-stage training. The significance of parameters has an active correlation with the magnitude of parameters in the pre-trained model (Han et al., 2015), and if the value is zero or close to zero, the parameter is inactivated and non-critical for optimization (Lee et al., 2019). Meanwhile, the safe parameters should play an important role in the first-stage training, where the model is trained merely on clean data and the safe parameters should be actively updated to fit the objective function. Thus, the magnitude of gradients during the first-stage training is also crucial.

A simple and straightforward way to combine the two factors is applying the product of the magnitude of parameters and their gradients. We denote the $i$-th parameter as $\omega_i$, and the initial value of $\omega_i$ is $\omega_{i0}$. In our first-training stage, we record the gradients of all parameters at each step. The training loss of the first stage is denoted as $L$, and the gradient of $\omega_i$ at the training step $j$ is $\frac{\partial L}{\partial \omega_{ij}}$. The importance score of $\omega_i$ is defined as

$$f(\omega_i) = \sum_{j=0}^{N} |\omega_{i0}| \cdot |\frac{\partial L}{\partial \omega_{ij}}| \quad (1)$$

where $N$ is the number of training steps in the first stage. We use $f(\omega_i)$ as a criteria to rank the parameters and those with the top ranking are considered safe parameters. Based on the ranking of parameters, we can generate a parameter mask $W$, where the elements are assigned a value of $1$ if the corresponding parameter is determined as a safe parameter, and $0$ otherwise.

$$\frac{\|W\|_0}{k} = p \quad (2)$$

where $\|\cdot\|_0$ means the standard $l_0$-norm, $k$ is the number of parameters and $p$ is the ratio of safe parameters.

| Methods | CoNLL03 | | | OntoNotes5.0 | | | Wikigold | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Pre.** | **Rec.** | **F1** | **Pre.** | **Rec.** | **F1** | **Pre.** | **Rec.** | **F1** |
| **Ours** | 0.848 | 0.874 | 0.861 | 0.806 | 0.830 | 0.818 | 0.700 | 0.769 | 0.733 |
| **NEEDLE** | 0.843 | 0.857 | 0.850 | 0.795 | 0.819 | 0.807 | 0.673 | 0.758 | 0.713 |
| Supervised Baselines | | | | | | | | | |
| **Supervised RoBERTa** | 0.824 | 0.773 | 0.798 | 0.765 | 0.807 | 0.785 | 0.640 | 0.725 | 0.680 |
| **Smooth Bound** | 0.845 | 0.789 | 0.816 | 0.771 | 0.812 | 0.791 | 0.655 | 0.731 | 0.691 |
| Noise-Robust Distantly-Supervised Methods | | | | | | | | | |
| **BOND** | 0.821 | 0.809 | 0.823 | 0.785 | 0.808 | 0.797 | 0.674 | 0.739 | 0.705 |
| **RoSTER** | **0.859** | 0.849 | 0.855 | 0.802 | 0.824 | 0.813 | 0.697 | 0.758 | 0.726 |
| Method Integration | | | | | | | | | |
| **NEEDLE + Ours** | 0.851 | 0.873 | 0.862 | **0.812** | 0.828 | 0.820 | **0.704** | 0.765 | 0.733 |
| **RoSTER + Ours** | 0.854 | **0.876** | **0.865** | 0.808 | **0.834** | **0.822** | 0.701 | **0.773** | **0.735** |

Table 1: Performance of all methods on three datasets measured by precision (Pre.), recall (Rec.) and F1 scores.

### 2.3. Estimating the Ratio of Safe Parameters

We have presented how to judge the safety of parameters and then divide them into safe and harmful ones. However, how to obtain the ratio of safe parameters $p$ remains an issue. Fortunately, previous works (Xia et al., 2021) show that the model performance is not sensitive to the variation of $p$, so the accuracy of safe prameters' ratio is not crucial and a rough estimate of $p$ is enough. We exploit the entity distribution difference between strongly labeled data and distantly supervised data to help estimate the ratio of safe parameters. Intuitively, if the difference is large, the label distribution of distantly supervised data is deviated a lot from the ground truth. Therefore, the number of safe parameters has a negative correlation with the entity distribution difference. We use the Kullback-Leibler (KL) divergence to calculate the distance between two distributions. However, most words in an NER dataset are not entities, whose labels are set to $O$. The proportion of $O$ is very high in both strongly labeled dataset and weakly labeled dataset, making the KL divergence of their label distribution quite small. So we ignore the label $O$ and only take those words labeled as entities into account. We estimate the probability of entity labels by their frequency, and then calculate the KL divergence. Noting that the value of KL divergence can be anywhere from 0 to infinity theoretically, we set a very small threshold $\tau > 0$, representing the minimum proportion of safe parameters. But practically, the values of KL divergence in most NER datasets are smaller than $1$. Therefore, the ratio $p$ of safe parameter is determined by

$$p = \begin{cases} 1 - KL(p||q), & 1 - KL(p||q) > \tau \\ \tau, & 1 - KL(p||q) \leq \tau \end{cases} \quad (3)$$

There might be a more accurate way to estimate $p$, but it is not the focus of our work and we leave it to the future work.

### 2.4. Optimization Method

By calculating the importance scores of model parameters and estimating the ratio of safe parameters, we can finally determine which parameters are safe. In order to combat label noise and prevent the model from fitting noisy data, we suppress the non-safe parameters during training on weakly labeled data and block the updates of non-safe parameters. Specifically, we apply to mask the gradient of parameter $\omega$ at each step by $W$

$$grad(\omega) \Leftarrow W \odot grad(\omega) \quad (4)$$

where $\odot$ denotes the element-wise multiplication. We propose two masking strategy termed hard mask and soft mask. For hard mask strategy, the element of $W$ is either '0' or '1'. '1' means that the parameter is safe and '0' means that the parameter is not essential for model generalization. For soft mask strategy, we set the mask value of non-safe parameters as a small number greater than 0, and set the mask value of safe parameters as a number slightly less than one. We cannot guarantee that the identification of safe parameters is absolutely accurate, so we provide the non-safe parameters a chance to slowly update and also reduce the confidence of updates of safe-parameters.

## 3. Experiments

### 3.1. Datasets

We conduct experiments on CoNLL03 (Sang and Meulder, 2003), OntoNotes5.0 (Weischedel et al., 2013) , and Wikigold (Balasuriya et al., 2009). For each dataset, we select a specific portion of data as strongly labeled data, while discarding labels of the remaining data and generate distant labels for it. In this work, instead of introducing new distant

| | CoNLL03 | OntoNotes5 | Wikigold |
|---|---|---|---|
| **S-Mask** | **0.861** | **0.818** | **0.733** |
| **H-Mask** | 0.859 | 0.817 | 0.729 |
| **R-Mask** | 0.820 | 0.801 | 0.698 |
| **w/o Mask** | 0.818 | 0.794 | 0.691 |

Table 2: The F1 scores of different model variations.

label generation methods, we follow the previous work (Liang et al., 2020; Meng et al., 2021) and use the distant labels provided by (Meng et al., 2021). We randomly selected 2 percent of the CoNLL03 and OntoNotes5.0 datasets as clean samples. Given the considerably smaller size of the Wikigold dataset, we increased this proportion to 10 percent.

### 3.2. Baselines

We compare our method with the following baselines. **NEEDLE** (Jiang et al., 2021) also studies the low-resource scenario where limited strongly labeled data and a large amount of weakly labeled data are available. We directly fine-tune the pre-trained **RoBERTa** (Liu et al., 2019) on the strongly labeled data as a supervised baseline. **Smooth Bound** (Zhu and Li, 2022) is also a state-of-the-art supervised method for NER. **Roster** (Meng et al., 2021) and **Bond** (Liang et al., 2020) are noise-robust learning methods for distantly supervised NER, and to be fairly compared with our method, we first apply them to the weakly labeled data and then fine-tune the model on the clean data.

### 3.3. Experimental 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; the number of the first-stage training epochs and the third-stage training epochs are the same, which are set to be 30, 20, 30; the number of the second-stage training epochs are set to be 3, 2, 5. For all three datasets: The training batch size is 32 and the threshold $\tau$ in Eq.3 is 0.2. We use Adam (Kingma and Ba, 2015) as the optimizer, and the peak learning rate is $3e-5$, $1e-5$, $5e-7$ for the first, sencond, and third training stage respectively with linear decay. The warmup proportion is $0.1$. We conducted each experiment five times and reported the mean precision and recall scores. We train the model on 1 NVIDIA A100 Tensor Core GPU.

### 3.4. Main Results

The main results of baselines and our method are shown in Table 1. Our method outperforms all baselines, which proves the effectiveness of our framework. The results of supervised methods are significantly lower, indicating that incorporating distantly labeled data with appropriate noise handling techniques can enhance the model performance. In addition, we also integrate our method with two data-level noise-robust training approaches by applying them to our second-stage training, and the results show that our method can further improve their performances, suggesting that the parameter-level approach and data-level approach can complement each other to achieve superior results.

### 3.5. Ablation Studies and Analysis

In this paper, we propose two gradient masking strategies, namely hard mask (**H-Mask**) and soft mask (**S-Mask**). Zhu et al. (2023) claim that simply fine-tuning the data with noisy labels followed by fine-tuning on clean samples can achieve strong performance, so we conduct experiments without parameter masking (**w/o Mask**) to testify the effectiveness of parameter suppression during distantly supervised training. We also experiment with a random parameter masking (**R-Mask**) strategy, where safe parameters are randomly selected. The results are shown in Table 2. It can be seen that the evaluation results of soft mask and hard mask are very close, and the F1 score of soft mask is slightly higher than hard mask, indicating that identification of safe parameters of our method is not absolute accurate, and a more relaxed standard for suppressing non-safe parameters leads to better performance. Random parameter suppression strategy performs better than simply fine-tuning the data with distant labels without any parameter suppression followed by fine-tuning on clean samples, but their F1 scores are both significantly lower than hard mask or soft mask. This indicates that parameter freezing can combat noise memorization as a regularization method, but our lottery training with safe parameter selection is more noise robust and can achieve better results.

In order to examine the influence of noise magnitude on our approach, we randomly change a certain proportion of labels in the distantly supervised data, and the proportion is called corruption rate. In Figure 2, we can see that the performance gain from incorporating distantly supervised training diminishes with increasing corruption rate and distantly supervised training even reduces the model performance since the noise is too strong. However, our lottery training is less affected by label noise, allowing the utilization of distantly labeled data even under higher levels of noise.

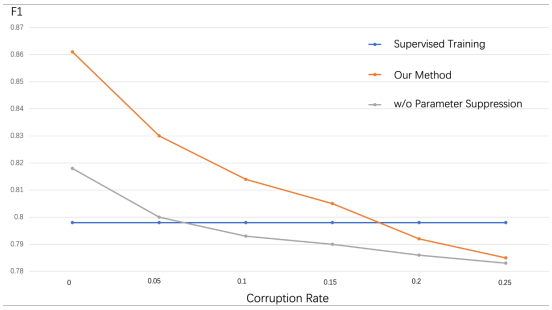


Figure 2: The model performance on CoNLL03 under different levels of noise.

## 4. Related Works

The idea of separating parameters into different levels and treating them differently during the training process came from the outgrowth of Lottery Ticket Hypothesis (LTH) (Frankle and Carbin, 2019), and subsequent research, including studies by He et al. (2022) and our own, applied this theory to various contexts and designed distinct approaches to address specific challenges. We apply LTH to a common low-resource scenario where strongly labeled data is limited while a large amount of weakly labeled data is available, which is especially common in NER as distant supervision is a widely used way to acquire data labels. Our work distinguishes itself from previous works in several key aspects:

1. The core idea of our work is leveraging the small amount of strongly labeled data to facilitate noise-robust training with weakly labeled data based on LTH, and we devised a novel three-stage training framework that fully capitalizes on weakly labeled data without degradation of model performance caused by noise.

2. Given that the application of LTH to different contexts all require the estimation of the proportion of different parameter types, we introduced an innovative technique tailored specifically for NER.

3. We avoid complex optimization strategies by carefully designing our three-stage training framework. This design enables us to more accurately identify safe parameters with the assistance of the limited clean data, while He et al. (2022) used a bi-level optimization strategy to ensure the reliability of safe parameter.

Our study primarily focuses on the noise introduced by distant supervision, but even human-annotated data may not be completely noise-free, which is a limitation in our approach. Some researchers propose to identify overly ambiguous or mislabeled samples and mitigate their impact when training neural networks by exploiting differences in the training dynamics of clean and mislabeled samples (Pleiss et al., 2020).

## 5. Conclusion

In this paper, we propose a three-stage noise-robust learning framework for low-resource NER from the perspective of parameter level, which identify the safe parameters leveraging insight provided by the training process on a small set of strongly labeled data, and suppress the non-safe parameters to impact the noise during training on distantly labeled data. Experiments on three representative datasets show that our method outperforms the state-of-the-art methods and it can be integrated with other methods for superior results.

## 6. Acknowledgements

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

## 7. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of $L_1$-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

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).

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. 2021a. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16306–16316. Computer Vision Foundation / IEEE.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tianlong Chen, Yongduo Sui, Xuxi Chen, Aston Zhang, and Zhangyang Wang. 2021b. A unified lottery ticket hypothesis for graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1695–1706. PMLR.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

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.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1135–1143.

Rundong He, Zhongyi Han, Yang Yang, and Yilong Yin. 2022. Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 6874–6883. AAAI Press.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Haoming Jiang, Danqing Zhang, Tianyu Cao, Bing Yin, and Tuo Zhao. 2021. Named entity recognition with small strongly labeled and large weakly labeled data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1775–1789. Association for Computational Linguistics.

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*.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2019. Snip: single-shot network pruning based on connection sensitivity. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

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.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. 2021. Robust early-learning: Hindering the memorization of noisy labels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14229–14253. Association for Computational Linguistics.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings*

*of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7096–7108. Association for Computational Linguistics.

## 8. Language Resource References

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.

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.

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.