Distantly Supervised Named Entity Recognition with Category-Oriented Confidence Calibration

Anonymous ACL submission

Abstract

In this work, we study the noisy-labeled named entity recognition under distant supervision setting. Considering that most NER systems based on confidence estimation deal with noisy labels ignoring the fact that model has different levels of confidence towards different categories, we propose a **C**ategory-**o**riented **c**onfidence c**a**libration (**Coca**) strategy with an automatically confidence threshold calculation module. We integrate our method into a teacher-student framework to improve the model performance. Our proposed approach achieves promising performance among advanced baseline models, setting new state-of-the-art performance on three existing distantly supervised NER benchmarks[1].

1 Introduction

Named entity recognition(NER) task is one of the core tasks in natural language processing(NLP), which lays foundations to many NLP applications including summarization(Erera et al., 2019; Hassel, 2003), relation extraction(Hou et al., 2019; Wang et al., 2020), question answering(Kandasamy and Cherukuri, 2020) and knowledge base population(Shen et al., 2014).

In many cases, annotated datasets for training supervised NER models are not available and it's labor-expensive and time-consuming to manually label the datasets. Distant supervision uses knowledge bases, domain ontologies and gazetteers to automatically generate annotated datasets, alleviating the need for hand-crafted datasets.

Although distant supervision(Mintz et al., 2009) provides an efficient way to annotate training data, entity type labels induced by distant supervision ignore entities' local context and may have limited usage in context-sensitive applications. In addition, since knowledge bases are inherently incomplete(Min et al., 2013), existing KBs only include limited entity mentions. Thus models trained on distantly supervised datasets fail to generalize to unseen entities.

To address these challenges, it's essential to denoise the distant labels for training the robust NER system. Many existing works propose various methods such as partial annotation learning(Mayhew et al., 2019), self-training(Jie et al., 2019), reinforcement learning(Yang et al., 2018), positive-unlabeled learning(Peng et al., 2019), causal intervention(Zhang et al., 2021a) to tackle this problem.

In this work, we introduce a category-oriented confidence calibration approach accompanying with an automatically confidence threshold calculation module and apply our method to a self-training teacher-student framework to process the unreliable labels in distantly supervised NER. The inspiration of our method is that traditional NER systems usually deal with noise by setting a confidence threshold to filter out uncertain labels. While we assume that it's inconsistent of the model's certainty for distinct label types, particularly under the label imbalance scenario. It could be detrimental to the model performance if there is only one confidence threshold for all the label types, especially for highly-skewed sequence labeling tasks.

To summarize, our major contribution includes:

1. We integrate category-oriented confidence calibration with teacher-student framework to tackle the problem of noisy labels in distantly supervised NER, which allows the model to adjust labels during self-training according to the confidence threshold of each label type.

2. Our proposed approach can automatically calculate the threshold of confidence estimation score for each label type instead of treating it as a hyper-parameter, which can be easily adapted to different NER tasks.

[1] Our code is publicly available at: `https://github.com/possible1402/BOND_Coca`

2 Methodology

In this section, we introduce our teacher-student framework with category-oriented confidence calibration. Figure 1 provides a graphical overview of our model.

2.1 Task Formulation

We formally formulate the named entity recognition task as a sequence labeling task. Given a sequence of tokens $\boldsymbol{X} = [x_1, ...x_i, ..., x_n]$, the aim is to predict a sequence of labels $\boldsymbol{Y} = [y_1, ...y_i, ..., y_n]$ that encodes the named entities, where n is the length of sequence and y_i is the corresponding label for token x_i. Let $L_E = [1, 2, ..., C]$ represent the label set and C is the number of classes.

2.2 Model Architecture

Our model consists of four modules, termed confidence threshold calculation, category-oriented confidence calibration, teacher network, and student network, where the student network and teacher network are structurally identical.

We use pretrained language model RoBERTa to implement the teacher-student framework, which acts as the encoder and a linear classification layer is atop it to compute the probability distribution of labels among L_E for each corresponding token x_i. Specifically, the RoBERTa layer f_θ maps the input sequence $\boldsymbol{x}$ into a sequence of hidden vectors $\boldsymbol{h} = \{h_1, ...h_i, ..., h_n\}$. After that, the classifier takes in token-wise hidden vector from the RoBERTa layer and gives the probabilities on all label types for each token x_i through $SoftMax$ function.

$$h_i = f_\theta(x_i) \tag{1}$$

$$p(x_i, \Theta) = SoftMax(Wh_i + b) \tag{2}$$

where $f_\theta(\cdot)$ produces context-sensitive representations for the input token sequence, h_i is the hidden vector of the final hidden layer corresponding to the i-th token x_i, $p(x_i, \Theta) \in \mathbb{R}^{|C|}$, and $\Theta = \{\theta, W, b\}$ denotes all the model parameters to be learned.

2.3 Self-Training with Category-Oriented Confidence Calibration

Under the teacher-student framework, our model modulates the parameter update to the student network according to the posterior confidence in its label-quality estimated by a teacher network based on category-oriented confidence calibration. The whole self-training process can be mainly split into three phases.

Step 1: Confidence Threshold Calculation.

Given the distant labeled NER training set $D = \{(\boldsymbol{X^{(m)}}, \tilde{\boldsymbol{Y}}^{\boldsymbol{(m)}})\}_{m=1}^{M}$, where $\tilde{\boldsymbol{Y}}^{\boldsymbol{(m)}} = [\tilde{y}_1^{(m)}, ..., \tilde{y}_i^{(m)}, ..., \tilde{y}_n^{(m)}]$ represents a sequence of distant labels for sample m and $\tilde{y}_i^{(m)} \in L_E$. We remark that self-training behaves poorly when encountering unreliable predicted labels namely pseudo labels, which will cause the student network to be updated towards the wrong direction. To alleviate this issue, we firstly initialize the student network and teacher network using the parameters of RoBERTa and then adapt them to acquire task dependent representation on the minibatch T_K from D whose size is K:

$$\Theta^{init} = \arg\min_{\Theta} \frac{1}{K} \sum_{m=1}^{K} l(\tilde{\boldsymbol{Y}}^{\boldsymbol{(m)}}, f(\boldsymbol{X}^{(m)}; \Theta)) \tag{3}$$

$$\Theta_{tea}^{0} = \Theta_{stu}^{0} = \Theta^{init} \tag{4}$$

During the adaptation, RoBERTa model assigns the probabilities to all label types for the i-th token in the m-th sample, and the output probability simplex over C classes is denotes as: $\boldsymbol{P_{x_i}^{(m)}} = [p(x_i^{(m)}, \Theta)_1, ..., p(x_i^{(m)}, \Theta)_C]$. Here $p(x_i^{(m)}, \Theta)_c$ denotes the predicted probability corresponding to the label type c. The formulas to get the pseudo label and the predicted confidence score for token x_i in m are as follows:

$$\hat{y}_i^{(m)} = \arg\max_{c} \boldsymbol{P_{x_i}^{(m)}} \tag{5}$$

$$\hat{s}_i^{(m)} = \max \boldsymbol{P_{x_i}^{(m)}} \tag{6}$$

To calculate confidence thresholds, we average over all confidence scores where the pseudo labels are same as the distant labels for the corresponding label type. Specifically, we firstly create an empty number sequence $\{\boldsymbol{S_j^{c,m}}\}_{\boldsymbol{j=1}}^{\boldsymbol{n_{c,m}}}$ to gather all the predicted confidence scores corresponding to label type c for the m-th sample, in which $n_{c,m}$ represents the length of number sequence. And then for sample m in minibatch T_K, if the distant label is same as the pseudo label in the corresponding position i and the entity type is c, we add the predicted confidence score in position i to the number

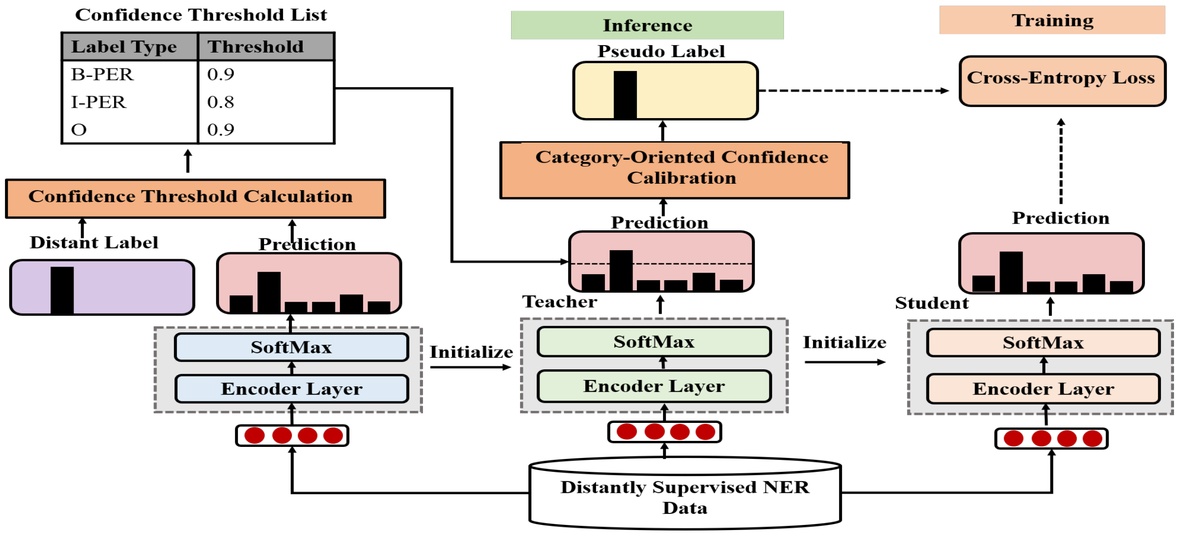

Figure 1: The Framework of Our Model

sequence. Finally, the confidence threshold for label type c, denoted as $\boldsymbol{V_c}$, is calculated as follows:

$$\boldsymbol{V_c} = \frac{\sum_{m=1}^{K} \sum_{j=1}^{n_{c,m}} \boldsymbol{S_j^{c,m}}}{\sum_{m=1}^{K} n_{c,m}} \tag{7}$$

Step 2: Category-Oriented Confidence Calibration.

Instead of discarding all distant labels(Liang et al., 2020), we only replace the distant labels over the confidence threshold with pseudo labels, while other distant labels remain constant. Intuitively, we consider that most distant labels are correct, and it's reasonable to do the replacement only if the model has strong confidence to believe the predicted pseudo labels.

To be precise, $p(x_i^{(m)}, \Theta)_{\tilde{y}_i^{(m)}}$ denotes the predicted probability corresponding to the distant label $\tilde{y}_i^{(m)}$ for the i-th token in the m-th sample. If the label type of the pseudo label $\hat{y}_i^{(m)}$ is c and the confidence score $\hat{s}_i^{(m)}$ is greater than or equal to $\boldsymbol{V_c}$, we assign $p(x_i^{(m)}, \Theta)_{\tilde{y}_i^{(m)}}$ to 1, and other elements in $\boldsymbol{P_{x_i}^{(m)}}$ are set to 0. After that, we update $\hat{y}_i^{(m)}$ and $\hat{s}_i^{(m)}$ according to the calibrated confidence simplex by Eq.(5) and Eq.(6). The final pseudo label for token $x_i^{(m)}$ after calibration is calculated as follows:

$$\hat{y}_i^{(m)} = \begin{cases} c & if\ \hat{s}_i^{(m)} \geq V_c \\ \tilde{y}_i^{(m)} & if\ \hat{s}_i^{(m)} < V_c \end{cases} \tag{8}$$

Step 3: Self-Training

We integrate our approach into a teacher-student self-training framework BOND (Liang et al., 2020). The teacher network is used to generate pseudo labels in the named entity recognition system. The student network is solving a surrogate task of approximating the output probability distribution of the entity types by the teacher network, transferring the knowledge from teacher network to student network.

The teacher network generates pseudo labels by Eq.(8) based on category-oriented confidence calibration. And then the student network is trained to fit these pseudo labels. Specifically, at the t-th iteration round, we learn the student network $\hat{\Theta}_{stu}^{t}$ by minimizing Eq.(3) with $\tilde{\boldsymbol{Y}}^{(\boldsymbol{m})}$ replaced by the pseudo labels counterparts $\hat{\boldsymbol{Y}}^{(\boldsymbol{m})}$, in which $\hat{\boldsymbol{Y}}^{(\boldsymbol{m})} = [\hat{y}_1^{(m)}, ..., \hat{y}_i^{(m)}, ..., \hat{y}_n^{(m)}]$. The final student model after self-training iteration is treated as the final model.

The student network is updated using stochastic gradient descent(SGD). Note that back-propagation is only through student network and the parameters of the teacher network are kept frozen during each self-training iteration. At the end of t-th iteration, we update the teacher model and student model as follows:

$$\Theta_{tea}^{t+1} = \Theta_{stu}^{t+1} = \hat{\Theta}_{stu}^{t} \tag{9}$$

3 Experiments

3.1 Experimental Setup

We consider five benchmark datasets including CoNLL03(Sang and De Meulder, 2003), Tweet(Godin et al., 2015), Wikigold(Balasuriya et al., 2009), Webpage(Ratinov and Roth, 2009)

Table 1: Experiment Results(F1%) on Fully Supervised Datasets and Distantly Supervised Datasets

Method	**Datasets**				
	conll2003	twitter	webpage	wikigold	BC5CDR
Roberta-base(Fully Supervised)	90.1	52.2	72.4	86.4	-
AutoNER(Shang et al., 2018)	67.0	26.1	51.4	47.5	**84.8**
LRNT (Cao et al., 2019a)	69.7	23.8	47.4	46.2	-
Noisy NER(Liu et al., 2021)	78.9	47.3	61.9	57.7	-
BOND(Liang et al., 2020)	81.5	48.0	65.7	60.1	78.7
BOND+BA+CIR(Zhang et al., 2021b)	81.5	**49.0**	64.7	61.5	-
BOND+Coca(Ours)	**82.7**	48.7	**68.2**	**63.0**	79.6

and BC5CDR(Li et al., 2016). The first four datasets are processed in KB-Matching annotation setting, where the distantly supervised labels are generated following BOND(Liang et al., 2020). And the BC5CDR dataset is processed following AutoNER(Shang et al., 2018). We compare with a wide range of state-of-the-art distantly-labeled NER models including AutoNER(Shang et al., 2018), LRNT(Cao et al., 2019b), BOND(Liang et al., 2020), Noisy NER(Liu et al., 2021), BOND+BA+CIR(Zhang et al., 2021b).

3.2 Experimental Results

Table 1 shows our primary results on fully supervised datasets and distantly supervised datasets. Experimental results demonstrate that our model is effective under distant supervision setting; it achieves state-of-the-art performance on three existing named entity recognition benchmarks.

Our proposed method is mostly closely aligned with the BOND framework. We observe that integrating our category-oriented confidence calibration strategy into BOND exceeds the performance without calibration on all of the five datasets by $\{1.3, 0.7, 2.5, 3.0, 0.9\}$ in terms of F1-score.

In addition, We demonstrate that our proposed category-oriented confidence estimation method is beneficial not only to open-domain NER tasks, but also to specific domain such as medical domain, which can be seen from the result on BC5CDR dataset. We note that one of the reasons why our method is worse than AutoNER on BC5CDR dataset is that they use additional lexicons to boost the model performance.

4 Related Work

Distant supervision is a particular instance of weak supervision, which relies on external resources such as knowledge bases to automatically label documents with entities that are known to belong to a particular category(Shang et al., 2018; Ritter et al., 2013). NER systems achieve high performance on clean text, while their performance dramatically degrades when moved to noisy scenarios such as distant supervision(Aguilar et al., 2018).

Denoising is an essential step in many distantly supervised NER systems. Peng et al. (2019) formulated the task as a positive-unlabeled (PU) learning problem to obtain unbiased estimation of the loss value. Shang et al. (2018) employed fuzzy CRF layer which assigned ambiguous tokens with all possible labels and maximized the overall likelihood. Yang et al. (2018) designed an instance selector based on reinforcement learning to distinguish positive sentences.

Our proposed method is most closely aligned with the BOND framework(Liang et al., 2020), in which a self-training algorithm is applied to guide the training of teacher-student network. They cast confidence threshold as a hyper-parameter to tune and prevented the low-confidence labels to be involved into loss calculation. Zhang et al. (2021a) integrated backdoor adjustment and causal invariance regularizer into BOND to conduct debiased method via causal interventions(Wu et al., 2020).

5 Conclusions

In this work, we propose a category-oriented confidence calibration approach accompanying with an automatically confidence threshold calculation module to tackle the problem of noisy supervision in distantly supervised NER. We integrate our method into Roberta under a teacher-student self-training framework. Extensive experiments demonstrate the effectiveness of our method. We evaluate our model on five datasets where our method outperforms state-of-the-art alternative distantly labeled learning methods in three datasets and shows promising results on another two datasets.

References

Gustavo Aguilar, Adrian Pastor López-Monroy, Fabio A González, and Thamar Solorio. 2018. Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1401–1412.

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.

Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, and Heng Ji. 2019a. Low-resource name tagging learned with weakly labeled data. *arXiv preprint arXiv:1908.09659*.

Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019b. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 261–270, Hong Kong, China. Association for Computational Linguistics.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 211–216, Hong Kong, China. Association for Computational Linguistics.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab@ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the workshop on noisy user-generated text*, pages 146–153.

Martin Hassel. 2003. Exploitation of named entities in automatic text summarization for swedish. In *NODALIDA'03–14th Nordic Conferenceon Computational Linguistics, Reykjavik, Iceland, May 30–31 2003*, page 9.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 729–734, Minneapolis, Minnesota. Association for Computational Linguistics.

Saravanakumar Kandasamy and Aswani Kumar Cherukuri. 2020. Query expansion using named entity disambiguation for a question-answering system. *Concurrency and Computation: Practice and Experience*, 32(4):e5119.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021. Noisy-labeled NER with confidence estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3437–3445, Online. Association for Computational Linguistics.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. Named entity recognition with partially annotated training data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 645–655, Hong Kong, China. Association for Computational Linguistics.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on*

Natural Language Processing of the AFNLP, pages 1003–1011.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Alan Ritter, Luke Zettlemoyer, Mausam Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

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*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021a. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4803–4813, Online. Association for Computational Linguistics.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021b. De-biasing distantly supervised named entity recognition via causal intervention. *arXiv preprint arXiv:2106.09233*.

A Example Appendix

This is an appendix.