Better Sampling of Negatives for Distantly Supervised Named Entity Recognition

Lu Xu[*] [1] [2] **Lidong Bing**[1] **Wei Lu**[2]
[1]DAMO Academy, Alibaba Group
[2]Singapore University of Technology and Design
xu_lu@hotmail.com l.bing@alibaba-inc.com
luwei@sutd.edu.sg

Abstract

Distantly supervised named entity recognition (DS-NER) has been proposed to exploit the automatically labeled training data instead of human annotations. The distantly annotated datasets are often noisy and contain a considerable number of false negatives. The recent approach uses a weighted sampling approach to select a subset of negative samples for training. However, it requires a good classifier to assign weights to the negative samples. In this paper, we propose a simple and straightforward approach for selecting the top negative samples that have high similarities with all the positive samples for training. Our method achieves consistent performance improvements on four distantly supervised NER datasets. Our analysis also shows that it is critical to differentiate the true negatives from the false negatives.[1]

1 Introduction

Named entity recognition (NER) is one of the fundamental tasks in natural language processing, and it aims to extract the mentioned entities in the text. Existing supervised approaches (Lample et al., 2016; Ma and Hovy, 2016; Devlin et al., 2019; Xu et al., 2021b) have achieved great performance on many NER datasets. However, they still heavily rely on the human-annotated training datasets.

Distantly supervised approaches (Ren et al., 2015; Fries et al., 2017; Shang et al., 2018; Yang et al., 2018; Mayhew et al., 2019; Cao et al., 2019; Peng et al., 2019; Liu et al., 2021; Zhang et al., 2021a; Zhou et al., 2022) have been proposed to exploit the automatically labeled training data generated from the knowledge bases (KBs) or dictionaries. For such distantly supervised datasets, the annotated entities mostly have correct labels, but the overall annotations are frequently incomplete due to the limited coverage of entities in KBs. We include comparisons of the distantly-annotated and human-annotated datasets in Appendix A.

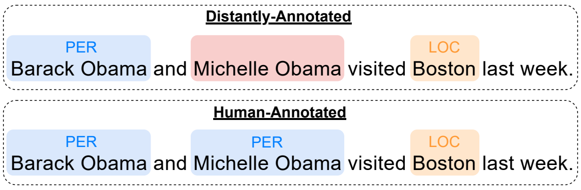
Distantly-Annotated
PER
LOC
Barack Obama and Michelle Obama visited Boston last week.
Human-Annotated
PER
PER
LOC
Barack Obama and Michelle Obama visited Boston last week.

Figure 1: An annotated example with distant supervision. The entity highlighted in red is not recognized.

Self-training has been demonstrated as an effective strategy for addressing the noisy labeled training data (Jie et al., 2019; Liang et al., 2020; Zhang et al., 2021b; Meng et al., 2021; Tan et al., 2022). Specifically, they iteratively refine the entity labels through teacher-student models. In this way, the number of false positive and false negative samples can be reduced. However, such approaches usually require training multiple models with multiple iterations. Another line of work improves the single-stage model by reducing the number of false negative samples during training. In general, the problem of false negatives is more severe than false positives. Li et al. (2021) proposed to sample a portion of negative samples with a uniform sampling distribution for training. By selecting a subset of all the negative samples, fewer false negative samples are involved in the training process. The sampling strategy can be enhanced by using a weighted sampling distribution (Li et al., 2022). Specifically, the negative samples are assigned with different sampling probabilities based on the predicted label distributions. However, this approach also depends on quality of the classifier to derive the distribution.

In this paper, we propose a simple and straightforward approach to sampling the negatives for training. Intuitively, the false negatives are positive samples but are unrecognized based on distant supervision, and they should have high similarities with positive samples that have the same gold entity type. For the example in Fig. 1, "Michelle

[*] This work was done when Lu Xu was under the joint Ph.D. program between Alibaba and SUTD.

[1] Our code is available at https://github.com/xuuuluuu/ds_ner.

Findings of the Association for Computational Linguistics: ACL 2023, pages 4874–4882
July 9-14, 2023 ©2023 Association for Computational Linguistics

Obama" is not identified as *PER* in the distantly labeled dataset. The false negative sample "Michelle Obama" should have high similarity with the positive sample "Barack Obama" that has the *PER* label. Additionally, the false negative sample should not have high similarity with other positive samples of different entity types, such as "Boston". Therefore, when the negative samples have high similarities with all the positive samples, they are more likely to be true negatives. We select these top negative samples for training, and we denote our approach as **Top-Neg**. Unlike the previous approach of relying on a classifier for assigning sampling probability, our approach exploits the encoded representations to derive the similarity scores. Compared to the baseline methods, our approach demonstrates consistent performance improvement on four distantly supervised NER datasets. Our analysis shows that not all the negative samples are required for training, but it is critical to filter the false negatives.

2 Approach

The objective of NER task is to extract all the entities in the text. Given a sentence of length n, $X = \{x_1, x_2, ..., x_n\}$, we denote all possible enumerated spans in X as $S = \{s_{1,1},\ s_{1,2},\ ...,\ s_{i,j},\ ...,\ s_{n,n}\}$, where i and j are the start and end position of span $s_{i,j}$, and the span length is limited as $0 \leq j - i \leq L$. For all the enumerated spans, our approach predicts the corresponding entity types from a predefined label space, including the O label (not an entity).

2.1 Span-based Model

Similar to the previous approaches (Lee et al., 2017; Luan et al., 2019; Zhong and Chen, 2021; Xu et al., 2021a), we adopt a span-based model architecture. First, we encode the input sentence X with a pre-trained language model, such as BERT (Devlin et al., 2019). The encoded contextualized representation for the sentence X is denoted as $\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n]$. Then, the span representation of $s_{i,j} \in S$ can be formed as:

$$\mathbf{s}_{i,j} = [\mathbf{h}_i;\ \mathbf{h}_j;\ f(i,j)] \tag{1}$$

where $f(i,j)$ indicates a trainable embedding to encode the span width feature. The span representation $\mathbf{s}_{i,j}$ is then input to a feed-forward neural network (FFNN) to obtain the distribution of entity type t.

$$P(t|\mathbf{s}_{i,j}) = \text{softmax}(\text{FFNN}(\mathbf{s}_{i,j})) \tag{2}$$

2.2 True Negatives vs. False Negatives

In general, the distantly annotated training datasets contain a significant number of false negative samples and also a small portion of false positives. When the model is trained on such datasets, the performance of precision and recall are affected, and the preliminary experimental results are given in Appendix B. We observe that the recall score is severely affected when compared with the precision score. Such behavior is because the problem of false negatives is more severe than false positives. We also observe a similar phenomenon in the statistics of the datasets in Appendix A.

Regarding the false negative samples, they are actually true positives but cannot be annotated based on only the distantly supervised information. Through our intuition that is described in Section 1, the false negative samples should have high similarities with the positive samples having the same gold entity type and low similarities with other positive samples of different entity types. Note that a vanilla model that is trained on the distantly supervised dataset can still well differentiate the positive labels, as demonstrated by the high precision score in the Appendix B. With the above findings, when a negative sample has a high similarity with all the positive samples, it is likely to be a true negative sample. Therefore, we propose to only utilize the negative samples that have high similarities with all the positive samples for training.

At the training stage, we have the label information so that we can obtain the span set of positive samples $S^{pos} = \{..., s^{pos}, ...\}$, and span set of negative samples $S^{neg} = \{..., s^{neg}, ...\}$. Note that $S = S^{pos} \cup S^{neg}$. Then, we calculate the average similarity score of each negative span $s^{neg} \in S^{neg}$ with respect to all the positive spans in S^{pos}, and the similarity score Φ is defined as:

$$\Phi(\mathbf{s}^{neg}, S^{pos}) = \frac{1}{M} \sum_{\mathbf{s}^{pos} \in S^{pos}} \frac{\mathbf{s}^{neg}}{\|\mathbf{s}^{neg}\|} \cdot \frac{\mathbf{s}^{pos}}{\|\mathbf{s}^{pos}\|} \tag{3}$$

where M denotes the number of positive samples. In practice, we calculate the similarity score at the batch level.

2.3 Training and Inference

We rank the similarity score Φ of all the negative spans in S^{neg}. Note that the number of the negative samples is denoted as N, which has a complexity of $O(n^2)$. To only consider the negative samples that have high similarities with all the positives and also save the computational cost, we select the top Nr negative samples for training and r is the hyper-parameter to control the quantity. We denote the set of the selected negative samples as $\tilde{S}^{neg} = \{..., \tilde{s}^{neg}, ...\}$. Then, we input all the positive samples and the selected negative samples to Eq. 2 to obtain the probability distributions. Our training objective is defined as:

$$\mathcal{L} = -\sum_{\mathbf{s}^{pos} \in S^{pos}} \log P(t^* | \mathbf{s}^{pos}) - \sum_{\tilde{\mathbf{s}}^{neg} \in \tilde{S}^{neg}} \log P(t^* | \tilde{\mathbf{s}}^{neg}) \quad (4)$$

where t^* denotes the corresponding gold entity type of a span. During inference, all the enumerated span representations are passed to Eq. 2 to predict the corresponding entity types.

3 Experiments

Datasets We evaluate our approach on four distantly supervised NER datasets: CoNLL03 (Tjong Kim Sang and De Meulder, 2003), BC5CDR (Wei et al., 2015), WNUT16 (Godin et al., 2015), and WikiGold (Balasuriya et al., 2009). The distantly supervised datasets are obtained from (Liang et al., 2020) and (Shang et al., 2018). We use the distantly supervised data for training and the human-annotated development and test sets for evaluation. The statistics of the datasets are given in Appendix A.

Experimental Setup We use the *bert-base-cased* and *roberta-base* as the base encoders for CoNLL03, WNUT16, and WikiGold datasets. BC5CDR is in the biomedical domain, and we adopt the *biobert-base-cased-v1.1* as the encoder. The maximum span length L is set as 8. The r is set as 0.05. See Appendix C for additional experimental settings. We use the same combination of hyperparameters for all experiments, and the reported results are the average of 5 runs with different random seeds.

Baselines *KB Matching* retrieves the entities based on string matching with knowledge bases. *AutoNER* (Shang et al., 2018) filters the distantly annotated datasets through additional rules and dictionaries, and they also proposed a new tagging scheme for the DS-NER task. *Bond* (Liang et al., 2020) proposed a two-stage approach to adopt self-training to alleviate the noisy and incomplete distantly annotated training datasets. *bnPU* (Peng et al., 2019) formulates the task as a positive unlabelled learning problem with having the mean absolute error as the objective function. *Conf-MPU* (Zhou et al., 2022) is a two-stage approach, with the first stage estimating the confidence score of being an entity and the second stage incorporating the confidence score into the positive unlabelled learning framework. *Span-NS* (Li et al., 2021) and *Span-NS-V* (Li et al., 2022) are the negative sampling approaches, while the latter replaces the previous uniform sampling distribution with a weighted sampling distribution.

As discussed by Zhou et al. (2022), the iterative self-training strategy (Liang et al., 2020; Zhang et al., 2021b; Meng et al., 2021) could be considered as a post-processing technique that is orthogonal to the single-stage approach. We consider the discussion of the self-training (Zoph et al., 2020) approach beyond the scope of this paper.

Experimental Results Table 1 shows the comparisons of our approach with the baseline methods on four datasets. Our model consistently outperforms the previous approaches in terms of the $F1$ score. *AutoNER* achieves good performance on the BC5CDR dataset by mining the phrases with external in-domain knowledge, but it does not show similar performance on the other three datasets. When comparing to the strong baseline *Conf-MPU*, our Top-Neg BERT achieves performance improvement of 0.92 and 3.17 F1 points on CoNLL03 and BC5CDR respectively. Note that the *Conf-MPU* also reported the results with lexicon feature engineering in the original paper, but they are not directly comparable with our approach. Our Top-Neg BERT also outperforms the previous sampling approach *Span-NS-V* by 1.52 $F1$ points on average. As the distantly supervised datasets are noisy in terms of both positive and negative samples, the *Span-NS-V* may not have a good classifier to determine the sampling probabilities. By contrast, our method only relies on the encoded representations of the samples to derive the similarity score for sampling. We also conduct experiments with RoBERTa as the encoder so as to have a fair comparison with the *BOND*. When the stronger pre-trained model is applied to our approach, we observe better performance on all datasets.

4 Analysis

Comparison with human-annotated training data We compare the performance of our *Top-*

Mode	Model	CoNLL03			BC5CDR			WNUT16			WikiGold		
		$P.$	$R.$	$F1$	$P.$	$R.$	$F1$	$P.$	$R.$	$F1$	$P.$	$R.$	$F1$
FS	Existing SOTA	-	-	94.60$^{\clubsuit}$	-	-	90.99$^{\blacklozenge}$	-	-	58.98$^{\blacklozenge}$	62.25	66.12	64.13$^{\spadesuit}$
DS	KB Matching*	63.75	81.13	71.40	51.24	86.39	64.32†	32.22	40.34	35.83	47.63	47.90	47.76
	AutoNER (Shang et al., 2018)*	60.40	75.21	67.00	77.52	82.63	79.99†	18.69	43.26	26.10	52.35	43.54	47.54
	BOND$_{\text{RoBERTa}}$ (Liang et al., 2020)	68.90	83.76	75.71	-	-	-	41.52	53.11	46.61	54.40	49.17	51.55
	bnPU (Peng et al., 2019)†	82.97	74.38	78.44	77.06	48.12	59.24	-	-	-	-	-	-
	Conf-MPU (Zhou et al., 2022)†	79.75	78.58	79.16	86.42	69.79	77.22	-	-	-	-	-	-
	Span-NS (Li et al., 2021)‡	80.41	71.35	75.61	86.90	73.49	79.64	53.51	39.76	45.62	51.05	48.27	49.62
	Span-NS-V (Li et al., 2022)‡	80.19	72.91	76.38	86.67	73.52	79.56	47.78	44.37	46.01	50.91	48.43	49.64
	Top-Neg (BERT)	82.72	77.71	80.08	-	-	-	55.28	40.35	46.55	55.47	48.57	50.65
	Top-Neg (RoBERTa)	81.07	80.23	**80.55**	-	-	-	60.55	45.33	**51.78**	52.30	53.55	**52.86**
	Top-Neg (Bio-BERT)	-	-	-	82.09	78.90	**80.39**	-	-	-	-	-	-

Table 1: Experiment results. ‘FS” and “DS” indicate fully supervised and distantly supervised respectively. The existing SOTA results marked with $^{\clubsuit}$ are retrieved from (Wang et al., 2021a), $^{\blacklozenge}$ are from (Wang et al., 2021b) and $^{\spadesuit}$ are from (Zhang et al., 2021b). The results with * are retrieved from (Liang et al., 2020), and the results with † are retrieved from (Zhou et al., 2022). ‡ indicates the results of our runs with their released code. See Appendix D for the standard deviation of our results based on 5 different runs and also the results on the development sets.

Model	Training	$P.$	$R.$	$F1$
Span	HA	91.14	91.68	91.41
Top-Neg	HA	91.48	91.66	91.57
Span	DS	88.25	63.03	73.54
Top-Neg	DS	82.72	77.71	80.08

Table 2: Comparisons on human-annotated (HA) and distantly supervised (DS) training data of CoNLL03.

Sampling	$P.$	$R.$	$F1$
Top 3%	84.73	78.61	81.55
Top 5%	85.78	79.38	82.42
Top 10%	88.12	75.51	81.33
Bottom 90%	75.33	70.82	73.01
Bottom 95%	80.36	69.00	74.25
Bottom 97%	82.92	71.96	77.05

Table 3: Results on the development set of CoNLL03 with different sampling strategies.

Neg with the standard span-based model[2] on the human-annotated (HA) and distantly supervised (DS) training sets in Table 2. When the HA dataset is used, our *Top-Neg* achieves comparable performance with the standard *Span* approach. This demonstrates that using all the negative samples for training is unnecessary. However, when the noisy DS dataset is used, the performance of the *Span* approach degrades significantly, especially the recall score. Our *Top-Neg* approach achieves better performance with relatively balanced precision and recall scores by sampling the effective negatives. Additionally, the performance gap of our approach on the HA and DS datasets indicates the room to further differentiate the true negatives from false negative samples.

Comparison of sampling strategies As mentioned, we propose to differentiate the false negatives from the true negatives based on the similarity between the negative sample with all positive samples. We conduct additional evaluations to show the effect of different sampling strategies on the performance, and Table 3 shows the comparisons. First, we compare the performance of our approach when only selecting the negative samples with the top similarity score (Eq. 3). We observe that the performance of our *Top-Neg* with 3% of the negative samples is worse than 5%. This indicates that the top 3% of negative samples are not adequate for training. However, when more negatives are selected (10 %), we observe a significant drop in the recall score as the number of false negative samples could become dominant. By contrast, when the negative samples with low similarity scores Φ are selected, the performance shows a significant decrease (lower section of Table 3). Even though the low similarity score indicates a high probability of being a true negative sample, however, these negative samples are less informative.[3]

5 Conclusion

In this work, we propose an improved approach of sampling the negatives to reduce the number of

[2] This span-based model uses all the negative samples.

[3] See Appendix E for the experiment results of using different sampling strategies on the HA dataset.

false negative samples for training. Specifically, we differentiate the true negatives from the false negative samples by measuring the similarity of the negatives with the positive samples. The experiment results have demonstrated the effectiveness of our approach. Future work may focus on clustering the negative samples to further differentiate the true negatives from the false negatives.

Limitations

Our approach is proposed based on the intuition that false negative samples should have high similarities with the positive samples that have the same gold entity type, and they also have low similarities with the positive samples of different entity types. However, our proposed approach does not guarantee the selected negatives are true negatives. Furthermore, when the negative samples are hard false negative samples, they are likely to have high similarities with other positive samples as well. However, such hard false negative samples are not prevalent in the datasets. Another limitation is that there is still a large performance gap between the distantly supervised datasets and the human-annotated datasets, as mentioned in Section 4.

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*.

Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. In *Proceedings of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Jason Alan Fries, Sen Wu, Alexander J. Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *ArXiv*, abs/1704.06360.

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*.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of NAACL*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of EMNLP*.

Yangming Li, Lemao liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *Proceedings of ICLR*.

Yangming Li, Lemao Liu, and Shuming Shi. 2022. Rethinking negative sampling for handling missing entity annotations. In *Proceedings of ACL*.

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 SIGKDD*.

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 NAACL*.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of NAACL*.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL*.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. Named entity recognition with partially annotated training data. In *Proceedings of CoNLL*.

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 EMNLP*.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of ACL*.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of SIGKDD*.

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

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of EMNLP*.

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 NAACL*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021a. Automated concatenation of embeddings for structured prediction. In *Proceedings of ACL*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021b. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. In *Proceedings of ACL*.

C.H. Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, C.J. Mattingly, Jiao Li, T.C. Wiegers, and Zhiyong Lu. 2015. Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021a. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of ACL*.

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021b. Better feature integration for named entity recognition. In *Proceedings of NAACL*.

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 COLING*.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021a. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of ACL*.

Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Xue Mengge, and Hongbo Xu. 2021b. Improving distantly-supervised named entity recognition with self-collaborative denoising learning. In *Proceedings of EMNLP*.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of NAACL*.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of ACL*.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. 2020. Rethinking pre-training and self-training. In *Proceedings of NeurIPS*.

A Dataset Statistics

Table 4 shows the evaluation results of the distantly supervised annotation when compared with human-annotated datasets. We observe that the results often show high precision but low recall scores.

Table 5 presents the statistics of the four distantly annotated datasets. "# Sent." indicates the number of sentences and "# Entity" denotes the number of entities in the datasets. The training set is annotated based on the distant supervision, and the development and test sets are manually annotated.

Datasets	Type	*P.*	*R.*	*F*1
CoNLL03	PER	82.36	82.11	82.23
	LOC	99.98	65.20	78.93
	ORG	90.47	60.59	72.57
	MISC	100.00	20.07	33.43
BC5CDR	Chemical	96.99	63.14	76.49
	Disease	98.34	46.73	63.35

Table 4: Evaluation results of the distantly annotated datasets based on human annotation.

B Preliminary Experiment Result

Figure 2 shows the experiment results of a standard span-based model that is trained on the distantly annotated CoNLL03 dataset. The evaluation is conducted on the human-annotated development and test sets.

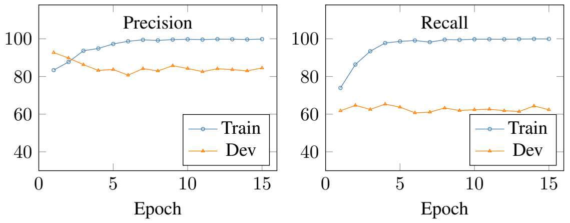

Figure 2: Precision (%) and recall (%) on the training and development sets of CoNLL03. Note that the best performance (F1 score) on the development set is at the 2nd epoch.

C Additional Experimental Setup

We use the *bert-base-cased* and *roberta-base* as the encoders for CoNLL03, WNUT16, and WikiGold datasets. BC5CDR is in the biomedical domain, and we adopt the *biobert-base-cased-v1.1* as the encoder. We use 2 layers of feed-forward neural networks for the classifier and the hidden size is set as 150, and the dropout rate is set as 0.2. The maximum span length L is set as 8. The r is set as 0.05.

Datasets		CoNLL03	BC5CDR	WNUT16	WikiGold
Train	# Sent.	14,041	4,560	2,393	1,142
	# Entity	17,781	6,452	994	2,282
Dev	# Sent.	3,250	4,579	1,000	280
	# Entity	5,942	9,591	661	648
Test	# Sent.	3,453	4,797	3,849	274
	# Entity	5,648	9,809	3,473	607

Table 5: Statistics of datasets.

Datasets	CoNLL03	BC5CDR	WNUT16	WikiGold
Top-Neg $_{\text{BERT}}$	82.42	-	44.34	55.10
Top-Neg $_{\text{RoBERTa}}$	83.67	-	48.78	57.46
Top-Neg $_{\text{BioBERT}}$	-	80.69	-	-

Table 6: Experiment results on the development sets.

We use the same combination of hyperparameters for all the experiments. We select the best model based on the performance on the development sets, and the reported results are the average of 5 runs with different seeds.

The experiments are conducted on Nvidia Tesla A100 GPU with PyTorch 1.10.0. The average running time on the CoNLL03 dataset is 74 seconds/epoch, and the number of model parameters is 108.59M when *bert-base-cased* is adopted.

Datasets	CoNLL03	BC5CDR	WNUT16	WikiGold
Top-Neg $_{\text{BERT}}$	0.94	-	1.09	1.03
Top-Neg $_{\text{RoBERTa}}$	0.63	-	0.42	0.25
Top-Neg $_{\text{BioBERT}}$	-	0.32	-	-

Table 7: Standard deviation of the $F1$ score on the test sets.

D Additional Experiment Results

In this section, we show additional experiment results. Table 6 presents the results of our approach on the development sets of the four datasets. As mentioned that we run our model with different seeds for 5 times, Table 7 shows the standard deviation of the $F1$ scores on the test sets.

E Additional Experiment Results on the HA Dataset

Table 8 shows the experimental results on the development set of CoNLL03 when using different sampling methods on the HA dataset. The experiment with the top 5% of the negative samples achieves comparable performance when using all the negative samples. We observe a large performance gap between the settings of the top 5% and bottom 5%. This indicates that the bottom negative samples are less informative than the top negatives. When the bottom 50% of negative samples are selected, the performance shows improvement, but it still exists a gap when compared with the top 50%.

Sampling	$P.$	$R.$	$F1$
ALL	95.65	95.93	95.79
Top 5%	95.70	95.76	95.73
Bottom 5%	64.85	96.90	77.70
Top 50%	96.10	95.31	95.70
Bottom 50%	89.70	96.13	92.80

Table 8: Comparisons of different sampling strategies on the development set of the HA CoNLL03.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
Section Ethics Statement

☑ 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?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ 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)?
No response.

□ 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?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ 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.
No response.

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3 and Appendix C

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?
Section 3 and Appendix C

☑ 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?
Section 3 and Appendix D

☑ 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.)?
Section 3 and Appendix C

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.