SCL-RAI: Span-based Contrastive Learning with Retrieval Augmented Inference for Unlabeled Entity Problem in NER

Shuzheng Si[1,2]*, **Shuang Zeng**[1,2]*, **Jiaxing Lin**[1,2], **Baobao Chang**[1]†

[1]Key Laboratory of Computational Linguistics, Peking University, MOE, China
[2]School of Software and Microelectronics, Peking University, China
{sishuzheng,jxlin}@stu.pku.edu.cn {zengs,chbb}@pku.edu.cn

Abstract

Named Entity Recognition is the task to locate and classify the entities in the text. However, *Unlabeled Entity Problem* in NER datasets seriously hinders the improvement of NER performance. This paper proposes **SCL-RAI** to cope with this problem. Firstly, we decrease the distance of span representations with the same label while increasing it for different ones via span-based contrastive learning, which relieves the ambiguity among entities and improves the robustness of the model over unlabeled entities. Then we propose retrieval augmented inference to mitigate the decision boundary shifting problem. Our method significantly outperforms the previous SOTA method by 4.21% and 8.64% F1-score on two real-world datasets.

1 Introduction

As a fundamental task in NLP, Named Entity Recognition aims to locate and classify named entities in the text. Due to the large-scale well-annotated datasets, deep-learning based methods (Li et al., 2022b; Devlin et al., 2019) have achieved great success. However, in real-world datasets, such as Ling and Weld (2012) with 112 fine-grained named entity tags, a large set of entity classes may cause inevitable missing annotations. Moreover, to obtain large NER datasets in practical scenarios, the distant supervision approach (Ren et al., 2015; Fries et al., 2017) may make this problem even worse, since the entity dictionary cannot cover all entities. Previous work (Li et al., 2021; Shang et al., 2018) find that this problem seriously hinders the performance of the NER model and name this problem as *Unlabeled Entity Problem*. As shown in Figure 1, the unlabeled second "NBA" may confuse model and introduce unnecessary noise.

To cope with this problem, several attempts from different perspectives have been proposed. Inspired

Figure 1: A toy case to show *Unlabeled Entity Problem*. The labeled entities are underlined with red lines while the unlabeled entities are underlined with dashed lines.

by positive-unlabeled (PU) learning (Li and Liu, 2005), Peng et al. (2019) use a weighted loss to assign low weights to false negative words and build distinct binary classifiers for different entity types. However, they require prior information or heuristics (Li et al., 2022c) and the unlabeled entities still misguide the classifiers, bringing ambiguity among neighboring entities (Li et al., 2021). Yang et al. (2018); Jie et al. (2019) introduce the Partial CRF (Lafferty et al., 2001) to marginalize the instances that are consistent with the incomplete annotation. However, they require additional well-annotated corpus to get ground truth negative instances, which are usually unavailable in practice. Recently, Li et al. (2021) perform down-sampling among non-entity instances within annotation when computing loss function, in order to mitigate the misguidance from possible unlabeled entities. Li et al. (2022c) further propose a weighted and adaptive sampling distribution to introduce direction to real unlabeled entities when down-sampling. However, the inherent randomness of sampling strategy may still keep some unlabeled entities when computing loss then make the decision boundary biased (Li et al., 2022a). As shown in Figure 2, the learned decision boundary for training example containing unlabeled entity instances tends to shift from the expected boundary towards the entity side. The previous works do not consider this problem.

To deal with these challenges, this paper proposes the **S**pan-based **C**ontrastive **L**earning with **R**etrieval **A**ugmented **I**nference (**SCL-RAI**) to tackle *Unlabeled Entity Problem*, which mitigates
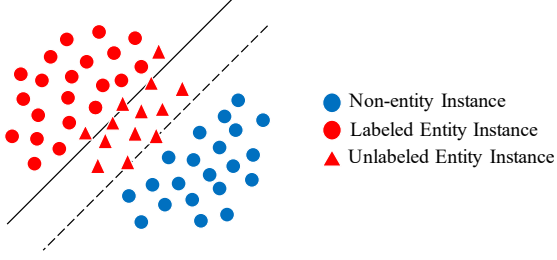
---

*Equal contribution.
†Corresponding author.

Figure 2: Illustration for decision boundary shifting phenomenon. The solid line is the learned boundary from datasets with unlabeled entities, the dashed line represents the expected boundary.

the limitations mentioned above, i.e., demanding additional corpus, ambiguity among neighboring entities and decision boundary shifting problem. Firstly, SCL-RAI tries to decrease the distance among span representations with the same labels while increasing it for different ones. Benefiting from our well-designed span-based contrastive learning, the ambiguity between entities is mitigated by the increased representation distance, so the model can capture the differences among different entity labels. We show in experiment that this contrastive learning objective also improves the model robustness under unlabeled entities. Furthermore, we propose Retrieval Augmented Inference to relieve decision boundary shifting phenomenon. It caches the center point representation for each entity type from the training set. Then, it computes a label distribution via cached representation and interpolates it with the distribution from the backbone NER model. Experiments on two real-world datasets show that SCL-RAI significantly outperforms previous SOTA methods.

2 Methodology

Our SCL-RAI consists of three modules: Span-based NER Model, Span-based Contrastive Learning, and Retrieval Augmented Inference.

2.1 Span-based NER Model

Span-based NER models have shown a strong ability to solve NER task, especially in flat NER and nested NER problem (Yu et al., 2020). For fair comparison, we follow Li et al. (2021, 2022c) on the design of Span-based NER model. Firstly, we use BERT (Devlin et al., 2019) as the text encoder to get the representations for words in sentence x:

$$[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n] = BERT(x) \qquad (1)$$

where h_i is the representation for word x_i. For each text span $s_{i,j}$ ranging from i-th word and j-th word in x, we get the span representation $\mathbf{s}_{i,j}$ as:

$$\mathbf{s}_{i,j} = \mathbf{h}_i \oplus \mathbf{h}_j \oplus (\mathbf{h}_i - \mathbf{h}_j) \oplus (\mathbf{h}_i \odot \mathbf{h}_j) \qquad (2)$$

where $\oplus$ is the concatenation operation and $\odot$ is the element-wise product operation. Finally, we use a two-layer non-linear projection to obtain the entity label distribution for every span $s_{i,j}$:

$$\mathbf{r}_{i,j} = tanh(W\mathbf{s}_{i,j}) \qquad (3)$$
$$\mathbf{o}_{i,j} = softmax(V\mathbf{r}_{i,j}) \qquad (4)$$

where W and V are trainable parameter matrices.

And the probability of l-th gold entity label for span instance $s_{i,j}$ is $\mathbf{o}_{i,j,l}$:

$$\mathbf{o}_{i,j,l} = \mathbf{v}_l^T \mathbf{r}_{i,j} \qquad (5)$$

We use cross entropy (CE) loss as our training objective:

$$loss_{ce} = \sum_{s_{i,j} \in D} -log(\mathbf{o}_{i,j,l}) \qquad (6)$$

where D is the collection of all training instances.

2.2 Span-based Contrastive Learning

To mitigate the ambiguity among entities, SCL-RAI tries to pull span belonging to the same class together in embedding space, while simultaneously pushing apart clusters of span from different classes.

This way, the clusters in entity representation space could better distinguish different types of entities. To this end, we propose a novel span-based contrastive learning objective to mitigate the ambiguity problem among entities. Meanwhile, we find in our experiment that this contrastive learning objective could also improve the robustness of SCL-RAI under unlabeled entity noises.

For span-based NER model, we conduct contrastive learning within a batch of span instances D; We use the cosine similarity to represent the distance between the span representations of two instances $s_{i,j}$ and $s_{\hat{i},\hat{j}}$:

$$d_{s_{i,j}, s_{\hat{i},\hat{j}}} = \frac{\mathbf{r}_{i,j} \cdot \mathbf{r}_{\hat{i},\hat{j}}}{|\mathbf{r}_{i,j}||\mathbf{r}_{\hat{i},\hat{j}}|} \qquad (7)$$

Then the span-based supervised contrastive learning loss function $loss_{scl}$ is defined as:

$$loss_{scl} = -\sum_{l \in L} \sum_{s_{i,j} \in D_l} \frac{1}{N_l - 1} \sum_{s_{\hat{i},\hat{j}} \in D_l} F(\mathbf{r}_{i,j}, \mathbf{r}_{\hat{i},\hat{j}})$$
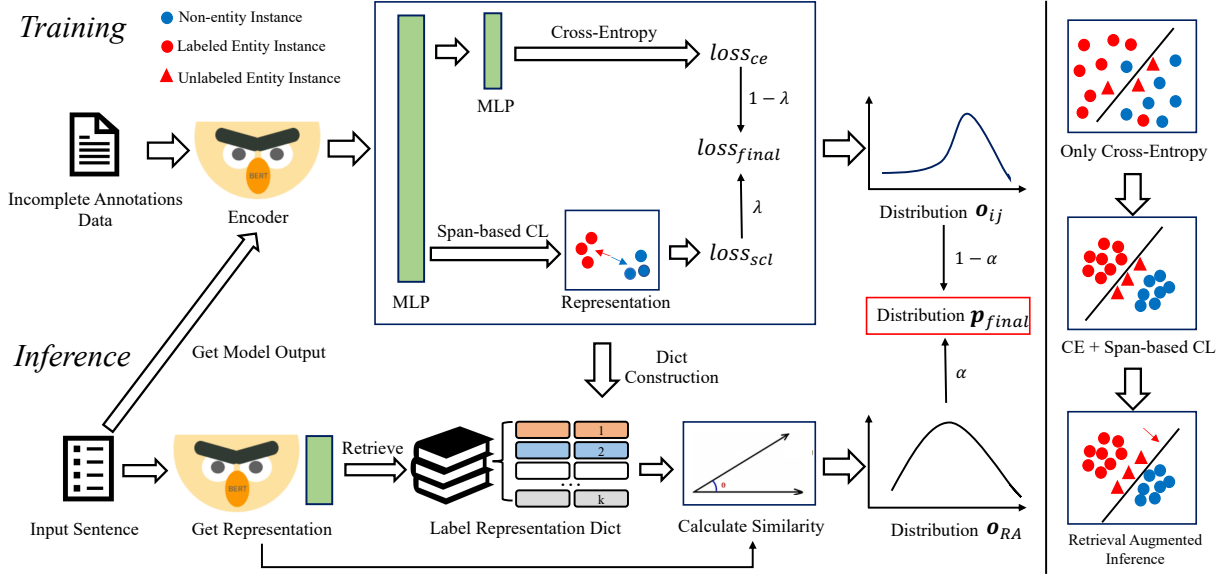$$(8)$$

Figure 3: General architecture of SCL-RAI.

where L is the size of the entity label set; $(i,j) \neq (\hat{i},\hat{j})$; N_l is the total number of span instances with the same entity label l in the batch; D_l is the collection of all training span instance with l-th entity label. $F(\mathbf{r}_{i,j}, \mathbf{r}_{\hat{i},\hat{j}})$ is:

$$F(\mathbf{r}_{i,j}, \mathbf{r}_{\hat{i},\hat{j}}) = log \frac{exp(d_{s_{i,j}, s_{\hat{i},\hat{j}}}/\tau)}{\sum_{s_{m,n} \in D_{\bar{l}}} exp(d_{s_{i,j}, s_{m,n}}/\tau)} \quad (9)$$

where τ is the temperature. $D_{\bar{l}}$ is the collection of span instances not with l-th entity label.

This span-based supervised contrastive learning loss pushes the span representations of instances with the same entity labels closer and pushes the span representations of instances with the different entity labels farther. We confirm in our experiment that this contrastive learning objective indeed improves the model robustness under unlabeled entities, compared with previous works.

Then we combine the cross entropy loss and span-based contrastive learning loss to get our final loss function:

$$loss_{final} = (1 - \lambda) * loss_{ce} + \lambda * loss_{scl} \quad (10)$$

where λ is a scalar hyperparameter.

2.3 Retrieval Augmented Inference

As we get the discriminative entity span representations via span-based contrastive Learning, we propose **R**etrieval **A**ugmented **I**nference (**RAI**) to facilitate the decoding process at the inference stage.

As shown in Figure 3, RAI can be split into two parts: (i) Firstly, it generates a central point representation for each entity type from the training set and stores them in a dictionary $Dict$. (ii) It calculates the similarity between the representation of span to be predicted and each entity type representation in $Dict$ to get the retrieval augmented label distribution $\mathbf{o}_{RA}$, then interpolates the distribution $\mathbf{o}_{i,j,l}$ from span-based NER model with $\mathbf{o}_{RA}$ to get the final label distribution. For example, the second "NBA" in Figure 1 will get high similarity value with the the central point representation of the entity type "ORG", due to the similar context with other "ORG" entities in training set. So it could decrease the possible high probability of non-entity label from Span-based NER model and increase it of "ORG" entity label. This way, we can shift the learned decision boundary toward the expected boundary in Figure 2.

Dictionary Construction: The dictionary $Dict$ used in SCL-RAI consists of a set of key-value pairs. Each key is an entity type and the corresponding value is the calculated central point representation from the training set. After training the model, we could get the dictionary for storing representations of all entity tags:

$$Dict = \{K, V\} = \{(l, \mathbf{r}_l) | \forall l \in L\} \quad (11)$$

$$\mathbf{r}_l = \sum_{s_{m,n} \in T_l} \frac{1}{N_l} \mathbf{r}_{m,n} \quad (12)$$

where T_l is the collection of all training span instances with l-th entity label; N_l is the total number

of span instances with the label l in the training set.
Label Distribution Interpolation: At the same time, Span-based NER model outputs representation $\mathbf{r}_{i,j}$ for the span to be predicted and its label distribution $\mathbf{o}_{i,j}$. Then we calculate the cosine similarity between $\mathbf{r}_{i,j}$ each cached representation from $Dict$ to obtain a new label distribution, i.e., retrieval augmented label distribution $\mathbf{o}_{RA}$:

$$\mathbf{sim}_{(i,j)} = concat(\frac{\mathbf{r}_{i,j} \cdot \mathbf{r}_{l_1}}{|\mathbf{r}_{i,j}||\mathbf{r}_{l_1}|}, ..., \frac{\mathbf{r}_{i,j} \cdot \mathbf{r}_{l_L}}{|\mathbf{r}_{i,j}||\mathbf{r}_{l_L}|}) \quad (13)$$

$$\mathbf{o}_{RA} = softmax(\mathbf{sim}_{(i,j)}) \quad (14)$$

where L is the number of entity labels.

We then set the value of non-entity label in $\mathbf{o}_{RA}$ to 0:

$$\mathbf{o}_{RA}[v] = 0 \quad (15)$$

where v is the index for the non-entity. This ensures the similarity of non-entity label does not participate in interpolation.

Finally, we interpolate the distribution $\mathbf{o}_{i,j,l}$ from span-based NER model with $\mathbf{o}_{RA}$ to get the final label distribution $\mathbf{p}_{final}$:

$$\mathbf{p}_{final} = (1 - \alpha) * \mathbf{o}_{i,j} + \alpha * \mathbf{o}_{RA} \quad (16)$$

where α is a hyperparameter to makes a balance between two distributions.

3 Experiments

3.1 Experimental Settings

Following (Yang et al., 2018; Li et al., 2021, 2022c), we adopt EC and NEWS as our datasets. The training set of EC and NEWS both consist of two parts: (1) the well-annotated set $\mathcal{A}$; (2) the distantly supervised set $\mathcal{DS}$, which contains a large amount of incompletely annotated sentences. Therefore, NER models trained on EC or NEWS suffer from *Unlabeled Entity Problem*. The dev/test set used in two datasets are well-annotated to evaluate the performance of model trained on datasets containing label noise.

EC In the e-commerce domain (EC), there are five types of entities: Brand, Product, Model, Material, and Specification. It contains 2,400 sentences labeled by annotators. The well-annotated set $\mathcal{A}$ is split into three sets: 1,200 sentences for training, 400 for dev, and 800 for testing. Then Yang et al. (2018) collect a list of entities to construct a dictionary from the training data and perform distant

Models	EC	NEWS
BERT-MRC	55.72	74.55
BERT-Biaffine Model	55.99	74.57
PU Learning	61.22	77.98
Partial CRF	60.08	78.38
Weighted Partial CRF	61.75	78.64
Vanilla Negative Sampling	66.17	85.39
Variant Negative Sampling	67.03	86.15
SCL-RAI	69.70	94.11
SCL-RAI+Vanilla Neg. Sampl.	**71.24**	**94.79**
- RAI	70.65 (-0.59)	93.71 (-1.08)
- SCL & RAI	66.17 (-5.07)	85.39 (-9.40)

Table 1: The F1-score results on two real-world datasets. "SCL" denotes Span-based Contrastive Learning and "RAI" denotes Retrieval Augmented Inference.

Models	$\mathcal{A}$	$\mathcal{A} + \mathcal{DS}$	Δ
Vanilla Negative Sampling	94.38	85.39	-8.99
SCL-RAI+Vanilla Neg. Sampl.	**95.33**	**94.79**	**-0.54**

Table 2: The robustness of SCL-RAI over unlabeled entities with different training set on NEWS dataset.

supervision on raw data to get the distantly supervised set $\mathcal{DS}$, which contains 2,500 sentences.

NEWS For news domain, Yang et al. (2018) use a NER data from MSRA (Levow, 2006). Yang et al. (2018) only keep entity type PERSON to get NEWS. Then (Yang et al., 2018) randomly select 3,000 sentences as training dataset, 3,328 as dev data, and 3,186 as testing data to get the well-annotated set $\mathcal{A}$. The rest set of MSRA is used as raw data, having 36,602 sentences. Yang et al. (2018) collect a list of person names from the training data. Then Yang et al. (2018) add additional names to the list. Finally, the list has 71,664 entries. Yang et al. (2018) perform distant supervision on raw data to obtain extra 3,722 sentences as the distantly supervised set $\mathcal{DS}$.

We adopt the same hyperparameter configurations for two datasets. We use Adam (Kingma and Ba, 2015) as optimizer with learning rate as 10^{-5} and bert-base as our encoder following Li et al. (2021, 2022c). The dimension of scoring layers W is set as 256. The scalar weighting hyperparameters λ and α are set as 0.1 and 0.5. The temperature parameter τ is set to 0.1. Since the label distribution is very unbalanced (most instances are non-entity), we also apply negative sampling and the same sampling rate as Li et al. (2021).

For evaluation, we use conlleval script[1] to compute the F1-score.

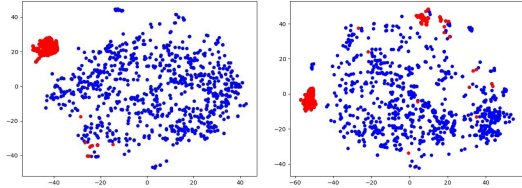[1] https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt.

Figure 4: t-SNE plots of the representations on NEWS test set. CE+Span-based CL (left), CE only (right). Red dot denote entities and blue dot denote the non-entities.

3.2 Results and Analysis

We report the results from: (1) **Traditional NER methods**: BERT-MRC (Yu et al., 2020) and BERT-Biaffine (Yu et al., 2020) ; (2) **Recent Attempts on Unlabeled Entity Problem**: PU Learning (Peng et al., 2019), Partial CRF (Yang et al., 2018), Weighted Partial CRF (Jie et al., 2019), Vanilla Negative Sampling (Li et al., 2021), Variant Negative Sampling (Li et al., 2022c) and our **SCL-RAI**. Since our method is orthogonal to that of (Li et al., 2021), we also report the results of SCL-RAI with their negative sampling strategy "**SCL-RAI**+Vanilla Neg. Sampl." to get better results.

We report our results in Table 1. Firstly, traditional NER models perform poorly on real-world datasets. So the SOTA NER models on well-annotated datasets are not robust over the *Unlabeled Entity Problem*. Then, our method has achieved new state-of-the-art results on the two datasets. Compared with SOTA model (Li et al., 2022c), we achieve the improvements of 2.67% F1 on EC and 7.96% on NEWS. With the negative sampling strategy, we further get the improvements of 4.21% F1 on EC and 8.64% F1 on NEWS. The improvements shows that our model has a stronger ability to mitigate the noise from unlabeled entities.

To verify the effectiveness of SCL-RAI, we show ablation studies in Table 1. It is clear that Span-based Contrastive Learning and Retrieval Augmented Inference are both important to cope with *Unlabeled Entity Problem*. In Table 2 and Table 4, we show the robustness of our model over unlabeled entities on NEWS dataset. Our SCL-RAI can obtain less F1 degradation when introducing dataset $\mathcal{DS}$ with unlabeled entities. In Figure 4, we show t-SNE plots of the learned representations of 2000 instances on NEWS test set, comparing Cross Entropy (CE) with and without the Span-based CL term. We can clearly see that the Span-based CL term enforces more compact clustering of entities.

For span-based NER model, we also conduct our

Batch Size	F1-score
8	70.78
16	71.24
32	70.87
64	70.79

Table 3: The Span-based Contrastive Learning robustness testing results on EC.

Variant	$\mathcal{A}$	$\mathcal{A} + \mathcal{DS}$	Δ
Vanilla Negative Sampling	76.82	66.17	-10.65
SCL-RAI	**76.44**	**71.24**	**-5.2**

Table 4: The robustness of SCL-RAI over unlabeled entities with different training set on EC dataset.

Span-based Contrastive Learning within a batch of span instances. Therefore, we test the robustness of Span-based Contrastive Learning for different batch sizes on EC. As shown in Table 3, we can clearly find that Span-based Contrastive Learning is robust for different batch sizes.

4 Conclusion

We propose the SCL-RAI to cope with *Unlabeled Entity Problem* in NER. Benefiting from our well-designed Span-based Contrastive Learning and Retrieval Augmented Inference, experiments on two real-world datasets show that SCL-RAI achieves more promising results than SOTA methods.

5 Acknowledgements

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. 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 4171–4186. Association for Computational Linguistics.

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

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.

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*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Changchun Li, Ximing Li, Lei Feng, and Jihong Ouyang. 2022a. Who is your right mixup partner in positive and unlabeled learning. In *International Conference on Learning Representations*.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022b. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Xiaoli Li and Bing Liu. 2005. Learning from positive and unlabeled examples with different data distributions. In *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*, pages 218–229. Springer.

Yangming Li, Lemao Liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yangming Li, Lemao Liu, and Shuming Shi. 2022c. Rethinking negative sampling for handling missing entity annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7188–7197, Dublin, Ireland. Association for Computational Linguistics.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*. AAAI Press.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2409–2419. Association for Computational Linguistics.

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 the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 995–1004. ACM.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Trans. Knowl. Data Eng.*, 30(10):1825–1837.

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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2159–2169. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.