NN-Denoising: A Low-Noise Distantly Supervised Document-Level Relation Extraction Scheme Using Natural Language Inference and Negative Sampling

Mengting Pan, Ye Wang, and Zhiyun Chen(✉)

East China Normal University, Shanghai 200000, China
{pmt,yewang}@stu.ecnu.edu.cn, chenzhy@cc.ecnu.edu.cn

Abstract. The task of document-level relation extraction (DocRE) is crucial in the field of natural language processing, as it aims to extract semantic relations between entities in a given document to facilitate deeper comprehension. Previous methods have primarily focused on fully supervised learning for DocRE, which requires a large amount of human-annotated training data, making it a tedious and laborious task. Recently, more and more attention has been paid to the incomplete labeling problem in human-annotated data, and it is believed to be the bottleneck of model performance. To address this limitation and mitigate annotation costs, we propose a low-noise distant supervision scheme for DocRE, called NN-Denoising, that combines natural language inference (NLI) models and negative sampling to filter out noise in the training data. The NLI model serves as a pre-filter for denoising the distant supervision (DS) labels, while negative sampling is employed to overcome the false negative problem in the filtered data. Our experimental results on a large-scale DocRE benchmark demonstrate the superiority of the proposed approach over existing baselines in distant supervision learning. Specifically, NN-Denoising achieves an improvement of 15.83 F1 points and 10.34 F1 points compared to the ATLOP and SSR-PU models, respectively.

Keywords: Document-level Relation Extraction · Distantly Supervised Learning · Low-Noise

1 Introduction

Relation extraction (RE) extracts semantic relationships among entities in text and has various applications such as sentiment analysis, information extraction,

Supported by Development of University Computer Fundamental Curriculum and Training Platform for Kunpeng Ecosystem, a Ministry of Education Industry-Academia-Research Project (No.201902146018).

L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14256, pp. 503–515, 2023.
https://doi.org/10.1007/978-3-031-44213-1_42

and knowledge graph construction. Most previous work has focused on sentence-level RE [12,34], which is limited as it cannot extract relationships between multiple sentences. To overcome this limitation, document-level relation extraction (DocRE) has been proposed [1,28,30,36], which extracts relationships within and between sentences.

Previous methods for DocRE have primarily focused on fully supervised learning, which can be time-consuming and labor-intensive in real-world scenarios due to the need for a large amount of human-annotated training data. Distantly supervised (DS) learning is a more efficient alternative, but it can introduce false positive (FP) problems [11]. In addition, the incomplete labeling problem, also known as the false negative (FN) problem, has received increasing attention in recent years [7,23]. To address the FN problem, previous work [25] proposed a unified positive-unlabeled learning framework - shift and squared ranking loss positive-unlabeled (SSR-PU) learning, to adapt DocRE with different levels of incomplete labeling. However, the SSR-PU method still faces challenges such as expensive labeling costs. To address these challenges, we propose a novel method for improving the FN and FP problems in distantly supervised learning for DocRE. Our approach, called NN-Denoising, aims to combine natural language inference (NLI) models and negative sampling to denoise the training data generated by DS. The NLI model serves as a pre-filter for denoising the DS labels, while negative sampling is employed to overcome the FN problem in the filtered data. This approach effectively improves the performance of the RE model by filtering out noisy training data and reducing the impact of incomplete labeling.

We conducted extensive experiments on the DS dataset provided in DocRED [30]. Recent work [23] has revealed the presence of severe incomplete labeling in the human-annotated data of DocRED. Through our rigorous experiments and analysis, we further identified prevalent FP and FN problems in the DS data. Subsequently, We compared the performance of our model with the state-of-the-art DocRE models under distantly supervised learning and observed a significant improvement in performance.

The contribution of this paper can be summarized as follows:

- We propose a novel approach to mitigate the FN problem in DocRE by introducing negative sampling, which, to the best of our knowledge, has not been applied before in this task.
- We present a low-noise distant supervision scheme, NN-Denoising, which utilizes NLI models and negative sampling to filter out noise in the training data, resulting in significantly improved performance of the DocRE model.
- Our proposed approach outperforms current state-of-the-art DocRE models under distant supervision, achieving substantial improvements of 15.83 F1 points and 10.34 F1 points over ATLOP [35] and SSR-PU [25], respectively.

2 Related Work

2.1 Document-Level Relation Extraction

Document-level relation extraction (DocRE) methods can be categorized into graph-based and transformers-based models. Graph-based models [8,15,29,32] use knowledge graphs to model and reason about entities and relations, while transformers-based models [22,27,33,35] leverage pre-trained language models and deep learning to achieve high-precision relation extraction. Recently, attention has been given to the incomplete labeling problem in DocRE datasets, with [7,23] pointing out that this is a bottleneck for model performance. Previous work, such as SSR-PU [25], was proposed to address this issue. However, [25] only addressed the FN problem in DocRE tasks under supervised learning and did not provide solutions for the FP problem that arises in distantly supervised learning.

2.2 Natural Language Inference

Natural language inference (NLI) is a crucial task in natural language processing that aims to determine the logical relationship between two statements. NLI models can be rule-based, logic-based [6,13,19], or deep learning-based [2,5]. Recently, there has been interest in using NLI models as independent RE models by formulating RE as an entailment task [21]. This approach has shown promising results in sentence-level RE tasks. Inspired by this, [24] conducted a study on the use of NLI as a pre-filter to improve distantly supervised DocRE and found that it can effectively enhance the task's performance.

2.3 Negative Sampling

Recently, a negative sampling method was proposed by [9] to address the FN problem in named entity recognition (NER) tasks. The method randomly samples a small subset of unlabeled spans as negative instances to induce the training loss, effectively eliminating the misleading effect of FN samples and improving the performance of NER models. Building upon this work, [23] also applied the negative sampling method to investigate the FN problem in RE tasks.

3 Methodology

3.1 NLI as Pre-filter

Here, we focus on the scenario where NLI is used as a pre-filter to filter the DS dataset, as presented in [24].

We start by taking a premise (p), which is an input text containing entity mentions of *head* and *tail*, and then construct a set of templates for each relation (r). Each template (t) in the set is specific to a particular relation and provides a structured representation of the relationship between *head* and *tail*. By combining a template and the premise, we can construct a hypothesis (h), which is a sentence that expresses the relationship between *head* and *tail* in a structured way. For instance, if the relation is "location", we might have templates like "The *head* is located in *tail*" and "The *tail* is near the *head*". Given a premise that mentions "Hawaii" as the *head* entity and "America" as the *tail* entity, we could construct the hypotheses "The Hawaii is located in America" and "The America is near the Hawaii" using these templates. To filter the DS dataset, we use an NLI model as a binary entailment task classifier. Given a premise p and a hypothesis h, the NLI model outputs a prediction score indicating whether the hypothesis is entailed by the premise or not.

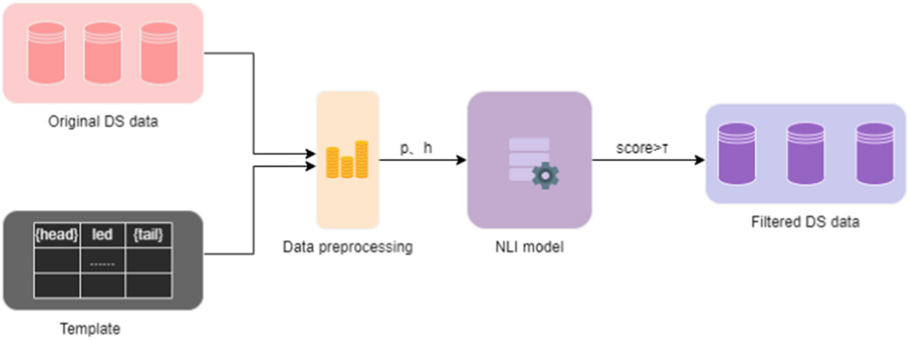

Fig. 1. The processing flow of NLI as a pre-filter.

As shown in Fig. 1, to apply this pre-filter, we integrate the original DS dataset with the template set and construct premise and hypothesis pairs for each labeled relation in the dataset. These pairs are then fed into the NLI model, which returns a prediction score. If the score is greater than a predefined threshold (τ), we retain the relation label for that hypothesis in the filtered DS dataset; otherwise, we discard it. We repeat this process for all hypotheses, resulting in a filtered DS dataset that is hopefully of higher quality.

Table 1. Percentages of triples left in the DS data after per-filtering with NLI.

Threshold τ	zero-shot
low (0.5)	73.4
med (0.95)	68.6
high (0.99)	59.0

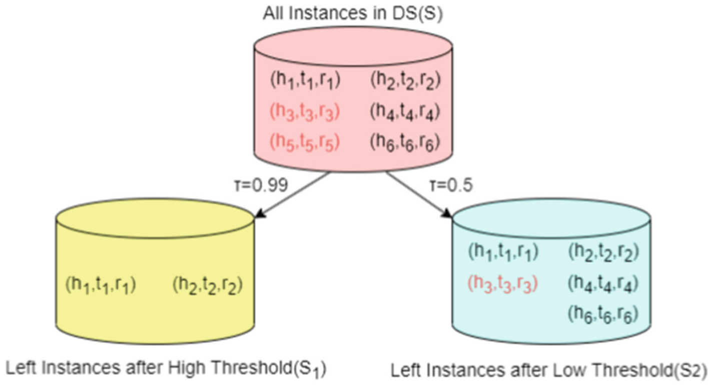

Fig. 2. Incomplete labeling problem in filtered data. FP samples are marked in red. (Color figure online)

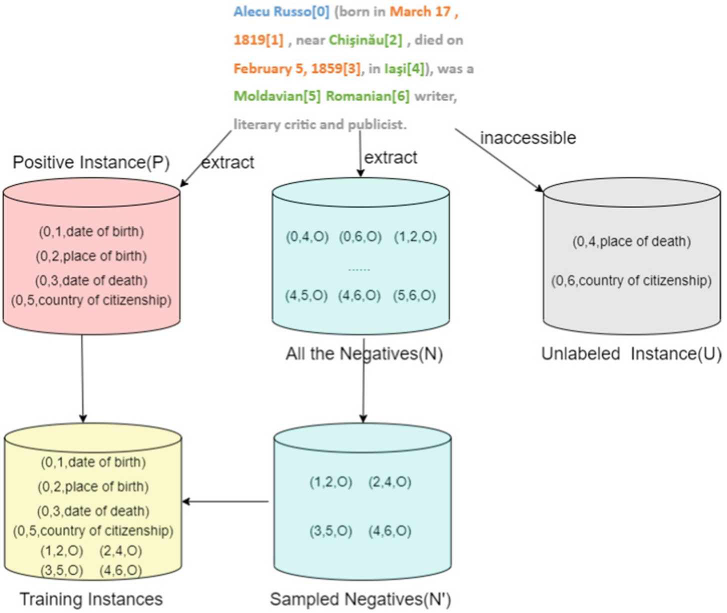
Alecu Russo[0] (born in March 17 , 1819[1] , near Chişinău[2] , died on February 5, 1859[3], in Iaşi[4]), was a Moldavian[5] Romanian[6] writer, literary critic and publicist.

Fig. 3. An example to depict how "whole-sample negative sampling" works.

Same as in [24], we set three threshold values (τ) of 0.5, 0.95, and 0.99 to evaluate the effectiveness of filtering. A higher threshold value leads to stricter filtering conditions and the removal of more FP samples. Table 1 displays the percentage of remaining triples in the filtered DS dataset under each threshold. However, our analysis shows that a threshold value of 0.99 may lead to filtering out some positive samples, exacerbating the incomplete labeling problem of the

DS dataset. This is why the RE model's performance was worse under high threshold conditions in previous work [24].

We provide a hypothetical example in Fig. 2, illustrating the impact of filtering with different threshold values. When the threshold is set to 0.99, only two positive samples remain, and two previously positive samples become false negatives. In contrast, a threshold of 0.5 retains more samples, including one false positive, but does not introduce any new false negatives. To overcome the FN problem during training with the low-noise dataset obtained through high threshold filtering, we introduce a negative sampling method explained in Sect. 3.2.

3.2 Training via Negative Sampling

To mitigate the issue of FN problem in the filtered DS dataset after high-threshold filtering, we introduce the negative sampling method proposed in [9]. This method has been successfully applied to NER tasks and is now applied to the DocRE task for the first time.

As illustrated in Fig. 3, given a document with labeled relation triples, let $\mathbf{P}$ = {(0, 1, "date of birth"), (0, 2, "place of birth"), (0, 3, "date of death"), (0,5,"country of citizenship")} be the set of labeled relation triples, and $\mathbf{U}$ ={(0, 4, "place of death"), (0, 6, "country of citizenship")} be the set of unlabeled relation triples. $\mathbf{P} \cup \mathbf{U}$ represents the ground-truth set of relation triples. Let $\mathbf{N}$ be the set of all negative samples, which includes all entity pairs in the document that are not part of the labeled relations, labeled as "O" to indicate that they are negative samples. The core idea of negative sampling is to randomly and uniformly select a small fraction of all negative samples for training to mitigate the misleading impact of false negatives. Let $\mathbf{N}'$ be a subset of $\mathbf{N}$, the final training data used in the model is $\mathbf{P} \cup \mathbf{N}'$.

Ultimately, a cross entropy loss used for training is incurred as:

$$\Big(\sum_{(i,j,l)\in\mathbf{P}} -\log(o_{i,j}[l])\Big) + \Big(\sum_{(i',j',l')\in\mathbf{N}'} -\log(o_{i',j'}[l'])\Big) \tag{1}$$

where the term $o_{i,j}[l]$ is the predicted score of a relation(i, j, l).

The negative sampling method based on [9] mentioned above is referred to as "whole-sample negative sampling" in this study. In addition, we propose a "label-specific negative sampling" method to better adapt to multi-label classification tasks. Instead of randomly and uniformly sampling negative samples from the entire dataset, we dynamically generate negative samples during training by considering each label individually. Specifically, we treat all samples other than those labeled with a particular label as negative examples for that label, and then perform uniform random sampling on these negative examples. This approach allows us to better capture the specific characteristics of each label and improve the model's ability to learn the nuances of the dataset. For example, in Fig. 3, when calculating the loss for the label "date of birth", we treat samples predicted as "date of birth" as positive samples and all other samples (including those predicted as "date of death", "place of birth", "place of death", "country of citizenship", and "*O*") as negative samples. Then, we randomly sample a subset of these negative samples for final loss calculation.

4 Experiments

Datasets and Evaluation Metric. This study uses the distantly supervised dataset from DocRED [30] as training data. For evaluation, the test set provided by Re-DocRED [23], a revised version of DocRED with more comprehensive annotations, is used. Evaluation metrics include micro F1(F1), micro ignore F1(Ign F1), precision(P), and recall(R), with Ign F1 measuring the F1 score excluding the relations shared by the training and test sets.

NLI Model. We used DeBERTaV3 [4], a pre-trained language model, as our NLI pre-filter. This model replaces masked language modeling (MLM) with replacement token detection (RTD) and achieves state-of-the-art performance. It was trained on 1.3 million hypothesis-premise pairs from 8 NLI datasets: MNLI [26], FEVER-NLI [16], NLI dataset from [18], and DocNLI [31](which is curated from ANLI [17], SQuAD [20], DUC20016[1], CNN/DailyMail [14], and Curation[2].

DocRE Model. For the DocRE model, we utilized ALTOP (Adaptive Thresholding and Localized Context Pooling) proposed in [35], which incorporates adaptive thresholding loss and localized context pooling techniques. This model addresses the issue of decision errors caused by using a global threshold in the original multi-label classification task and leverages pre-trained models to obtain better entity representations.

Implementation Details. We used pre-trained models from Huggingface[3] for the NLI pre-filter and considered only zero-shot scenarios. It is worth noting that the NLI pre-filter only filters the training data (DocRED) [30] and does not need to filter the test data (Re-DocRED) [23], as the test dataset is already a revised version. For the DocRE model, we implemented ATLOP model based on $BERT_{Base}$ [3] and $RoBERTa_{Large}$ [10], respectively. The learning rate was adjusted to 3e−5, and the training epochs were set to 10 when using BERT as the encoder. For RoBERTa, the hyperparameters remained unchanged. Negative sampling rates were adjusted from 0.01 to 0.1. We report the final model's performance rather than the best checkpoint, and all experiments were conducted on an NVIDIA A100-40GB GPU.

[1] https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html.

[2] Curation. 2020. Curation corpus base.

[3] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c.

Baseline. We re-implemented distantly supervised learning on ATLOP [35] and SSR-PU [25] and used their results as the baseline for our task. We used similar settings as our own method, with a learning rate of 3e−5 and 10 training epochs when using BERT as the encoder, while keeping other parameters unchanged. We reported the performance of the final model instead of the best checkpoint.

5 Results and Analysis

Table 2. Main result on DocRED. The number in the subscript represents the filtering threshold for NLI.

Model	F1	Ign F1	P	R
ATLOP+$BERT_{Base}$	42.35	39.36	75.55	29.42
SSR-PU+ATLOP+$BERT_{Base}$	47.84	41.61	50.29	45.62
NLI-filtering$_{0.99}$+ATLOP+$BERT_{Base}$	36.18	35.65	**92.40**	22.50
Negative-Sampling+$BERT_{Base}$	46.60	44.53	62.61	40.59
NN-Denoising$_{0.5}$+ATLOP+$BERT_{Base}$	56.02	51.84	63.40	50.17
NN-Denoising$_{0.95}$+ATLOP+$BERT_{Base}$	56.86	52.22	57.98	**55.78**
NN-Denoising$_{0.99}$+ATLOP+$BERT_{Base}$	**58.18**	**54.44**	64.32	53.10
ATLOP+$RoBERTa_{Large}$	43.47	40.47	76.06	30.43
SSR-PU+ATLOP+$RoBERTa_{Large}$	49.11	42.88	50.74	47.58
NLI-filtering$_{0.99}$+ATLOP+$RoBERTa_{Large}$	36.95	36.53	**93.81**	23.01
Negative-Sampling+$RoBERTa_{Large}$	46.50	39.57	37.88	**60.19**
NN-Denoising$_{0.5}$+ATLOP+$RoBERTa_{Large}$	55.49	50.16	51.82	59.72
NN-Denoising$_{0.95}$+ATLOP+$RoBERTa_{Large}$	57.54	52.66	55.92	59.25
NN-Denoising$_{0.99}$+ATLOP+$RoBERTa_{Large}$	**59.03**	**54.95**	60.72	57.43

Our experimental results, as shown in Table 2, emphasize the effectiveness of our proposed negative sampling method in addressing the FN problem in distant supervision DocRE models. Specifically, we set the filtering threshold for NLI and varied the negative sampling rate to achieve the best results. We implemented our models using both $BERT_{Base}$ and $RoBERTa_{Large}$ encoders.

From the results, we observe that: (1) our method outperforms the baseline models ATLOP and SSR-PU, achieving the best F1 score in both settings. With the BERT encoder, our model achieves 58.18 F1 points, surpassing the baseline models ATLOP (15.83 F1 points) and SSR-PU (10.34 F1 points), and with the RoBERTa encoder, our model achieves 59.03 F1 points, surpassing the baseline models ATLOP (15.56 F1 points) and SSR-PU (9.92 F1 points). This improvement in performance emphasizes the effectiveness of our method in mitigating the negative impact of FN and FP problem on DocRE models. (2) our results show that our model performs better with a high threshold for NLI filtering compared to a low threshold, which differs from the findings of previous studies [24]. This demonstrates the effectiveness of our negative sampling method in addressing the FN problem, which is the main bottleneck for RE model performance. Moreover, it confirms our hypothesis that the high threshold filtered DS dataset suffers from a more severe incomplete labeling problem than the low threshold filtered DS dataset. (3) when using only the NLI filtering method (NLI-filtering$_{0.99}$+ATLOP+$BERT_{Base}$), the performance of the DocRE model was the worst, even lower than that of directly training on the original DS dataset. This is because the incomplete labeling phenomenon in the high threshold filtered DS dataset is more severe, reducing the performance of the DocRE model. This finding further confirms that the FN problem, rather than the FP problem, is the performance bottleneck of distant supervision DocRE models. (4) our model's performance was improved compared to the baseline model ATLOP (+4.25 F1 points) when trained only with negative sampling (Negative-Sampling+ATLOP+$BERT_{Base}$). The performance was also comparable to the SSR-PU model that effectively addressed the FN problem. This further confirms the effectiveness of our negative sampling method in mitigating the FN problem in distant supervision DocRE models. (5) Our proposed method exhibits superior stability in terms of precision and recall compared to other methods, indicating that it does not suffer from overfitting and possesses excellent predictive capability.

Table 3. Effect of different sampling methods on model performance. “w” means “whole-document negative sampling” and “l” means “label-specific negative sampling”.

Model	F1	Ign F1	P	R
NN-Denoising$_{0.99}$+ATLOP+$BERT_{Base}$(w)	55.98	52.58	**66.22**	48.48
NN-Denoising$_{0.99}$+ATLOP+$BERT_{Base}$(l)	**58.18**	**54.44**	64.32	**53.10**

To comprehensively evaluate the effectiveness of our proposed sampling methods, we conducted a comparative analysis using the high threshold NLI filtering and the optimal negative sampling rate (0.01). As shown in Table 3, our proposed “label-specific negative sampling” method outperforms the “whole-document negative sampling” method by 2.2 F1 points. This improvement is mainly due to our method's ability to sample negative examples specific to each label, effectively addressing the label imbalance problem in DocRE datasets.

To investigate the impact of negative sampling rate on addressing the FN problem in the distant supervision DocRE task, we conducted experiments under both low and high threshold conditions in NLI filtering, with negative sampling rates ranging from 0.01 to 0.1. The results, as shown in Fig. 4, revealed that lower negative sampling rates led to better performance in both scenarios, indicating the effectiveness of negative sampling in reducing the risk of training the model with FN samples.

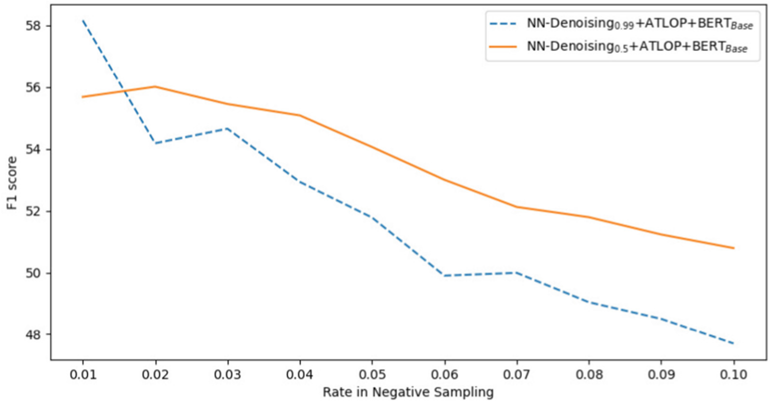

Fig. 4. Effect of different negative sampling rates on model performance.

We observed that the performance of the model was more sensitive to the negative sampling rate in the high threshold scenario compared to the low threshold scenario. This finding indicates that negative sampling is particularly effective in scenarios where the incomplete labeling problem is more severe, thus highlighting the potential of our proposed sampling method in addressing the FN problem in distant supervision DocRE models. This insight can guide future research in this area.

6 Conclusion and Future Work

The aim of this work was to address the issue of expensive human annotation in DocRE tasks by introducing DS learning. During the training of DocRE models, we noticed the potential bottleneck in performance caused by false negatives, as well as the severe FP problem introduced by DS. To address these issues, we proposed a method that improves the performance of DocRE models by combining NLI as a pre-filter to remove most false positive samples and negative sampling to solve the severe FN problem in the filtered DS dataset. Our experiments demonstrated that our model significantly outperforms baseline models under complete DS learning, with a performance improvement of 15.83 F1 points over the ATLOP and 10.34 F1 points over the SSR-PU.

In addition, we observed that different sampling methods have a significant impact on addressing incomplete labeling problem. In this work, we used two simple sampling methods based on random uniform sampling, there is still much room for improvement in this area. Future work can explore more effective sampling methods that can further enhance the training performance of our model. Furthermore, developing more robust and accurate NLI models is also a potential research avenue.

References

1. Christopoulou, F., Miwa, M., Ananiadou, S.: Connecting the dots: document-level neural relation extraction with edge-oriented graphs (2019). https://doi.org/10.48550/ARXIV.1909.00228. https://arxiv.org/abs/1909.00228
2. Das, R., Munkhdalai, T., Yuan, X., Trischler, A., McCallum, A.: Building dynamic knowledge graphs from text using machine reading comprehension (2018). https://doi.org/10.48550/ARXIV.1810.05682. https://arxiv.org/abs/1810.05682
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics (2018)
4. He, P., Gao, J., Chen, W.: DeBERTaV 3: improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing (2021). https://doi.org/10.48550/ARXIV.2111.09543. https://arxiv.org/abs/2111.09543
5. Henaff, M., Weston, J., Szlam, A., Bordes, A., LeCun, Y.: Tracking the world state with recurrent entity networks (2016). https://doi.org/10.48550/ARXIV.1612.03969. https://arxiv.org/abs/1612.03969
6. Hobbs, J.R., Stickel, M.E., Appelt, D.E., Martin, P.: Interpretation as abduction. Artif. Intell. **63**(1), 69–142 (1993). https://doi.org/10.1016/0004-3702(93)90015-4. https://www.sciencedirect.com/science/article/pii/0004370293900154
7. Huang, Q., Hao, S., Ye, Y., Zhu, S., Feng, Y., Zhao, D.: Does recommend-revise produce reliable annotations? An analysis on missing instances in docRED (2022). https://doi.org/10.48550/ARXIV.2204.07980. https://arxiv.org/abs/2204.07980
8. Li, B., Ye, W., Sheng, Z., Xie, R., Xi, X., Zhang, S.: Graph enhanced dual attention network for document-level relation extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 1551–1560. International Committee on Computational Linguistics (Online), December 2020. https://doi.org/10.18653/v1/2020.coling-main.136. https://aclanthology.org/2020.coling-main.136
9. Li, Y., Liu, L., Shi, S.: Empirical analysis of unlabeled entity problem in named entity recognition (2020). https://doi.org/10.48550/ARXIV.2012.05426. https://arxiv.org/abs/2012.05426
10. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach (2019). https://doi.org/10.48550/ARXIV.1907.11692. https://arxiv.org/abs/1907.11692
11. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: 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, Suntec, Singapore, pp. 1003–1011. Association for Computational Linguistics, August 2009. https://aclanthology.org/P09-1113
12. Miwa, M., Bansal, M.: End-to-end relation extraction using LSTMs on sequences and tree structures (2016). https://doi.org/10.48550/ARXIV.1601.00770. https://arxiv.org/abs/1601.00770
13. Moldovan, D.I., Clark, C., Harabagiu, S.M., Maiorano, S.J.: COGEX: a logic prover for question answering. In: North American Chapter of the Association for Computational Linguistics (2003)
14. Nallapati, R., Zhou, B., dos Santos, C.N., Gulcehre, C., Xiang, B.: Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv: Computation and Language (2016)

15. Nan, G., Guo, Z., Sekulić, I., Lu, W.: Reasoning with latent structure refinement for document-level relation extraction (2020). https://doi.org/10.48550/ARXIV.2005.06312. https://arxiv.org/abs/2005.06312
16. Nie, Y., Chen, H., Bansal, M.: Combining fact extraction and verification with neural semantic matching networks. arXiv: Computation and Language (2018)
17. Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., Kiela, D.: Adversarial NLI: a new benchmark for natural language understanding. arXiv: Computation and Language (2019)
18. Parrish, A., et al.: Does putting a linguist in the loop improve NLU data collection. Empirical Methods in Natural Language Processing (2021)
19. Raina, R., Ng, A.Y., Manning, C.D.: Robust textual inference via learning and abductive reasoning. In: Proceedings of the 20th National Conference on Artificial Intelligence, AAAI 2005, vol. 3, pp. 1099–1105. AAAI Press (2005)
20. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuaD: 100,000+ questions for machine comprehension of text (2016)
21. Sainz, O., de Lacalle, O.L., Labaka, G., Barrena, A., Agirre, E.: Label verbalization and entailment for effective zero and few-shot relation extraction. Empirical Methods in Natural Language Processing (2021)
22. Tan, Q., He, R., Bing, L., Ng, H., Academy, D., Group, A.: Document-level relation extraction with adaptive focal loss and knowledge distillation (2023)
23. Tan, Q., Xu, L., Bing, L., Ng, H.T., Aljunied, S.M.: Revisiting docRED - addressing the false negative problem in relation extraction (2022). https://doi.org/10.48550/ARXIV.2205.12696. https://arxiv.org/abs/2205.12696
24. Vania, C., Lee, G., Pierleoni, A.: Improving distantly supervised document-level relation extraction through natural language inference. In: Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pp. 14–20. Association for Computational Linguistics, Hybrid, July 2022. https://doi.org/10.18653/v1/2022.deeplo-1.2. https://aclanthology.org/2022.deeplo-1.2
25. Wang, Y., Liu, X., Hu, W., Zhang, T.: A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling (2022). https://doi.org/10.48550/ARXIV.2210.08709. https://arxiv.org/abs/2210.08709
26. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv: Computation and Language (2017)
27. Xu, B., Wang, Q., Lyu, Y., Yong, Z., Mao, Z.: Entity structure within and throughout: modeling mention dependencies for document-level relation extraction. In: Proceedings of the ... AAAI Conference on Artificial Intelligence (2021)
28. Xu, W., Chen, K., Mou, L., Zhao, T.: Document-level relation extraction with sentences importance estimation and focusing (2022). https://doi.org/10.48550/ARXIV.2204.12679. https://arxiv.org/abs/2204.12679
29. Xu, W., Chen, K., Zhao, T.: Document-level relation extraction with reconstruction. Cornell University - arXiv (2021)
30. Yao, Y., et al.: DocRED: a large-scale document-level relation extraction dataset (2019). https://doi.org/10.48550/ARXIV.1906.06127. https://arxiv.org/abs/1906.06127
31. Yin, W., Radev, D.R., Xiong, C.: DocNLI: a large-scale dataset for document-level natural language inference. Cornell University - arXiv (2021)
32. Zeng, S., Xu, R., Chang, B., Li, L.: Double graph based reasoning for document-level relation extraction. Cornell University - arXiv (2020)
33. Zhang, N., et al.: Document-level relation extraction as semantic segmentation. Cornell University - arXiv (2021)

34. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction (2018). https://doi.org/10.48550/ARXIV.1809.10185. https://arxiv.org/abs/1809.10185
35. Zhou, W., Huang, K., Ma, T., Huang, J.: Document-level relation extraction with adaptive thresholding and localized context pooling (2020). https://doi.org/10.48550/ARXIV.2010.11304. https://arxiv.org/abs/2010.11304
36. Zhou, Y., Lee, W.S.: None class ranking loss for document-level relation extraction (2022). https://doi.org/10.48550/ARXIV.2205.00476. https://arxiv.org/abs/2205.00476