Relation Extraction with Weighted Contrastive Pre-training on Distant Supervision

Anonymous ACL submission

Abstract

Contrastive pre-training on distant supervision has shown remarkable effectiveness for improving supervised relation extraction tasks. However, the existing methods ignore the intrinsic noise of distant supervision during the pre-training stage. In this paper, we propose a weighted contrastive learning method by leveraging the supervised data to estimate the reliability of pre-training instances and explicitly reduce the effect of noise. Experimental results on three supervised datasets demonstrate the advantages of our proposed weighted contrastive learning approach, compared to two state-of-the-art non-weighted baselines.

1 Introduction

Relation extraction (RE) is the task of identifying the relationship between entities mentioned in the text, which can benefit many downstream tasks such as question answering and knowledge base population. Since most of the existing RE models (Zhang et al., 2020; Zeng et al., 2020; Lin et al., 2020; Wang and Lu, 2020; Zhong and Chen, 2021) are trained on the labeled data, the amount of training data limits the performance of supervised RE systems. To tackle this problem, recent work leverage a semi-supervised distant supervision (DS) (Mintz et al., 2009; Lin et al., 2016; Vashishth et al., 2018; Chen et al., 2021) approach to generate abundant training data by aligning knowledge bases (KBs) and raw corpora. However, distantly supervised relation extraction (DSRE) inevitably suffers from wrong labeling noise. Introducing a robust framework that utilize both the abundant but noisy data from DS and the scarce but accurate data from human annotations becomes a new research line to improve RE systems.

Recent works (Baldini Soares et al., 2019; Ormándi et al., 2021; Peng et al., 2020) propose a two-stage RE framework that they first design an RE oriented task to pre-train BERT on DS data and then fine-tune on human annotated (HA) datasets. Peng et al. (2020) use Wikipedia articles as the corpus and Wikidata as the KB in the pre-training stage to construct the DS data, and they introduce a contrastive learning based method to pre-train BERT on the generated DS data. Given an anchor instance with a specific relation in the DS data, their contrastive learning method randomly selects one positive sample holding the same relation and maximizes the similarity between the anchor and positive sample. Meanwhile, their method randomly selects multiple negative samples holding different relations from the anchor and minimizes the similarity between the anchor and negative samples. The results show that their RE oriented pre-training can effectively improve the final performance of the RE task on various target datasets.

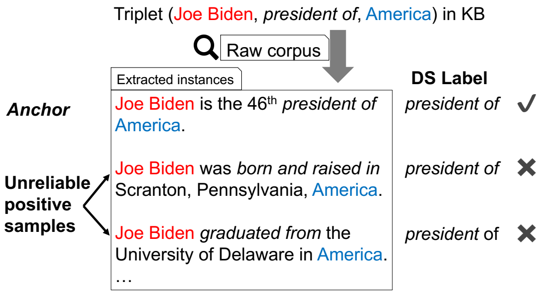
Triplet (Joe Biden, president of, America) in KB
Raw corpus
Extracted instances
DS Label
Anchor
Joe Biden is the 46th president of America.
president of ✓
Unreliable positive samples
Joe Biden was born and raised in Scranton, Pennsylvania, America.
president of ✗
Joe Biden graduated from the University of Delaware in America.
president of ✗
...

Figure 1: An example of unreliable positive samples caused by DS noise

However, in their pre-training stage, they ignore the intrinsic wrong labeling noise in the generated DS data. Since their method relies on the DS labeled relation types to sample positive and negative instances, the noisy labeling problem leads to unreliable samples in Figure 1, potentially limiting the pre-training stage's effectiveness. To better utilize DS data, we propose a novel weighted contrastive learning framework to both use the abundant DS data and tackle the inevitable DS noise. First, we train a relation classifier on the HA dataset and leverage the classifier to predict the relation type of instances in the DS data. Then for each DS instance, based on the output of the classifier, we can compute the confidence score to measure the reliability of its labeled relation type. Finally, we introduce weights based on computed confidence scores into the contrastive learning loss to focus more on reliable instances while less on noisy ones.

Besides, distant supervision relies on the existing KBs to align raw corpora. To alleviate the need for KBs, we propose a new strategy to extract a triplet set from the HA dataset for generating DS data. We also include a KB derived DS dataset in our experiments to show that our proposal can still work well for regular DS.

In conclusion, we propose a weighted contrastive pre-training approach for supervised relation extraction and introduce its details in Section 2. Then we perform the experiments on three datasets to compare our proposed method with existing baselines in Section 3.

2 Proposed Method

2.1 Overview

We show the overview of our proposal in the Figure 2. We start from generating the DS data relying on the HA dataset. Then in the first stage, we introduce a weighted contrastive learning method by leveraging the HA data to estimate the reliability of DS instances for contrastive pre-training. In the second stage, we further fine-tune our pre-trained model on the HA dataset.

2.2 Distantly Supervised Dataset Construction

Since DS uses existing knowledge bases to generate training data, in the case that we have no proper existing KBs in some domains but only the annotated dataset, we first extract all entities based on each sentence, and if any two of them are labeled a relation type, they will generate a triplet with a particular relation. Otherwise, they will still generate a triplet but labeled NA (no relation). After constructing the KB, we can extract sentences containing two entities of each triplet from raw corpora. To balance the number of sentences extracted by each triplet, we also add an upper bound 100 to the number of extracted sentences.

2.3 Two-stage RE Framework

Instance representation In our pre-training stage, we use BERT to obtain the representation for each input instance. For the input format, we follow PURE (Zhong and Chen, 2021) by adding extra special markers to mark the beginning and the end of two entities. For example, given an instance x: "***Joe Biden*** *is the president of* ***America***.", the input sequence is "[CLS] [H_CLS] *Joe Biden* [H_SEP] *is the president of* [T_CLS] *America* [T_SEP]. [SEP]". Denote the k-th output vector of the BERT encoder as h_k. Assuming i and j are the indices of two beginning entity markers [H_CLS] and [T_CLS], we define the instance representation as:

$$\mathbf{x} = h_i \oplus h_j \tag{1}$$

where $\oplus$ stands for concatenation. Then we use the instance representation for the further reliability estimation and the weighted contrastive learning in the pre-training stage.

Reliability estimation With the instance representation, we first fine-tune BERT on the HA dataset as a supervised RE task. Then with the trained relation classifier $\mathcal{F}$, we can make predictions on each instance in the DS data. Given an input instance x with DS labeled relation r, we can derive the confidence score c to estimate its reliability by:

$$c = \frac{\exp\left(\mathcal{F}(\mathbf{x},r)\right)}{\sum_{r'\in R}\exp\left(\mathcal{F}(\mathbf{x},r')\right)} \tag{2}$$

where R is the set of all relation classes, and $\mathcal{F}(\mathbf{x},r)$ computes the output of our relation classifier on the labeled class r. Through this approach, we can estimate the reliability of the labeled relation for each DS instance by its corresponding confidence score.

Stage 1: DS weighted contrastive pre-training Contrastive learning aims at maximizing the similarity between a given instance and its positive samples while minimizing the similarity between the given instance and its negative samples. As for existing work, Peng et al. (2020) focus on the relation level that DS instances labeled the same relation are positive samples while DS instances labeled different relations are negative samples. The latest DSRE work (Chen et al., 2021) augment the anchor as a positive sample to avoid the effect of DS noise. Both work do not explicitly address the problem of unreliable positive and negative samples.

In our work, we introduce a robust weighted contrastive learning (WCL) method with the help of

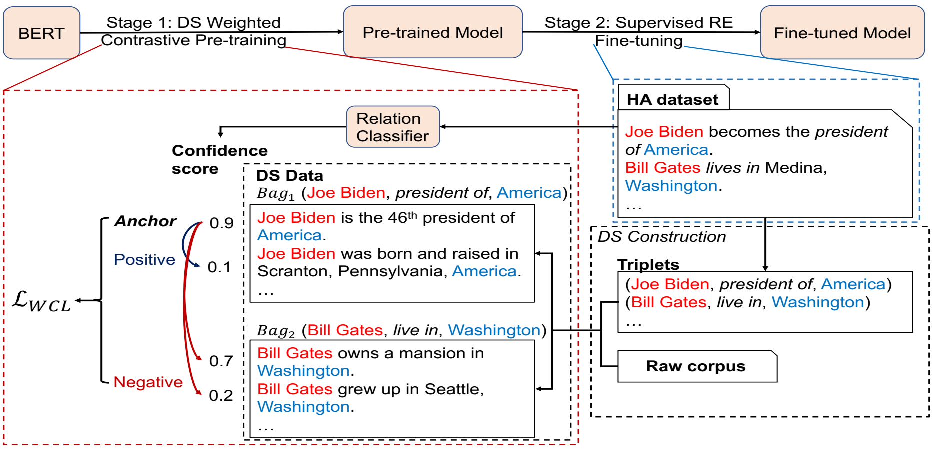
BERT
Stage 1: DS Weighted Contrastive Pre-training
Pre-trained Model
Stage 2: Supervised RE Fine-tuning
Fine-tuned Model
HA dataset
Joe Biden becomes the president of America.
Bill Gates lives in Medina, Washington.
…
Relation Classifier
Confidence score
DS Data
Bag1 (Joe Biden, president of, America)
Joe Biden is the 46th president of America.
Joe Biden was born and raised in Scranton, Pennsylvania, America.
…
Bag2 (Bill Gates, live in, Washington)
Bill Gates owns a mansion in Washington.
Bill Gates grew up in Seattle, Washington.
…
Anchor
Positive
Negative
0.9
0.1
0.7
0.2
$\mathcal{L}_{WCL}$
DS Construction
Triplets
(Joe Biden, president of, America)
(Bill Gates, live in, Washington)
…
Raw corpus

Figure 2: Overview of our proposed method.

reliability estimations for each instance. Given a batch input with multiple bags: $(Batch = \{B_i\}_{i=1}^{G})$ where G is the number of bags in one batch, and the labeled relational triplets are different from each other. Each bag B is constructed by a certain relational triplet (e_1, r, e_2) with all instances x inside satisfying this triplet. Moreover, each instance comes along with a confidence score c estimating its reliability: $B_i = \{x_j, c_j\}_{j=1}^{N_i}$, where N_i denotes the size of bag B_i. Then the weighted contrastive learning loss of an anchor instance x_j in the bag B_i is:

$$\mathcal{L}_{WCL}^{(ij)} = -\log\left\{\sum_{k=1}^{N_i} c_j c_k e^{cos(\mathbf{x}_j,\mathbf{x}_k)/\mathcal{T}} \Big/ \left(\sum_{k=1}^{N_i} c_j c_k e^{cos(\mathbf{x}_j,\mathbf{x}_k)/\mathcal{T}} + \sum_{r_m \neq r_j} c_j c_m e^{cos(\mathbf{x}_j,\mathbf{x}_m)/\mathcal{T}}\right)\right\} \tag{3}$$

where $cos(\cdot)$ denotes the cosine similarity between two instance representations, $\mathbf{x}_k$ denotes the representation of a positive instance sampled from the same bag, and $r_m \neq r_j$ denotes that negative samples x_m are selected from all instances in the batch that are labeled a different relation from x_j. We follow (Khosla et al., 2020) to incorporate multiple positive instances sampled from the same bag. $\mathcal{T}$ denotes a scaling temperature.

With the help of confidence scores, the model will focuses on more reliable instances while ignoring unreliable instances which keep paces with our goal to utilize reliable DS data.

Besides, to inherit the ability of language understanding from BERT and avoid catastrophic forgetting, we also adopt the masked language modeling (MLM) objective from BERT.

Eventually, we define our final pre-training loss:

$$\mathcal{L} = \mathcal{L}_{WCL} + \mathcal{L}_{MLM} \tag{4}$$

Stage 2: Supervised relation extraction We then fine-tune the pre-trained model on HA datasets with state-of-the-art (SOTA) methods. For i2b2 2010VA we follow BLUEBERT (Peng et al., 2019) by treating the relation extraction task as a sentence classification and replacing two named entities in the sentence with predefined tags. For the other two datasets, we follow the encoding method of PURE (Zhong and Chen, 2021) as introduced at the beginning of Section 2.3.

3 Experiments

3.1 Setup

HA and DS datasets We evaluate our approach on three HA relation extraction datasets: i2b2 2010VA, ACE05, and Wiki20m. Table 2 shows the statistics of each dataset. The i2b2 2010VA is a medical domain RE dataset while other two datasets are collected from general domains. We generate the DS data for i2b2 2010VA and ACE05 from corresponding raw corpora. Meanwhile, Wiki20m is a regular KB based distantly supervised RE dataset containing both DS data and HA data and it is worth noting that we intend to show that our method can also work well on existing DS datasets. Table 3 shows the statistics of DS data.

Baselines We have a naive baseline by directly fine-tuning (FT) on each dataset as a supervised RE task. We set two two-stage framework baselines: the first one is to use the SOTA method RE-Context-or-Names (RECN) (Peng et al., 2020)

Methods	**i2b2 2010VA**		**ACE05**		**Wiki20m**	
	25%	100%	25%	100%	25%	100%
FT	66.86	75.22	**62.81**	**70.41**	68.87	88.54
CIL + FT	67.92	75.39	59.72	69.69	89.67	91.64
RECN + FT	67.65	75.43	60.34	69.40	89.23	91.96
WCL + FT (ours)	**68.50**	**76.15**	61.30	69.47	**90.28**	**92.67**

Table 1: Evaluation results on various datasets. 25% denotes the low-resource setting, and 100% denotes the full-resource setting. We compute three-run average Micro-F1 for our proposed methods in all the results.

Dataset	# Rel.	# Train	# Dev	# Test
i2b2 2010VA	8	3,120	11	6,147
ACE05	6	10,051	2,424	2,050
Wiki20m	80	8,279	4,140	28,977
ACE05 (NP)	6	3,939	922	923

Table 2: Statistics of datasets. Rel. denotes relation types. NP denotes removing pronoun from ACE05.

Dataset	# Triplets	Corpora	# DS Ins. (NA)
i2b2 2010VA	2,777	MIMIC-III	36K (76K)
ACE05	3,883	Gigaword5	98K (461K)
Wiki20m	-	Wiki20m	286K (698K)
ACE05 (NP)	3,218	Gigaword5	60K (273K)

Table 3: Statistics of DS data. Triplets are extracted from the HA dataset. DS Ins. denotes relational instances generated by DS. NA denotes the no-relation instances. NP denotes removing pronoun from ACE05.

in pre-training, and the second one is to use the SOTA DSRE method Contrastive Instance Learning (CIL) (Chen et al., 2021) in pre-training.

Implementation details To further confirm the effectiveness of our proposal, we also conduct the experiments in the low-resource setting by randomly selecting 25% of the full HA data to construct the DS data for pre-training and finally finetune on this 25% HA data. Refer to Appendix A for other implementation details.

3.2 Main Results

Table 1 compares our model to other baselines. From the results, we can observe that: (1) For both the i2b2 2010VA and the Wiki20m, all two-stage models outperform the FT baseline, which indicates the effectiveness of our strategy to construct DS data from HA datasets, especially in the low-resource setting. (2) For both the i2b2 2010VA and the Wiki20m, our proposed model achieves the best F1 scores over all baselines. This improvement shows that it is worthy to estimate the reliability of each DS instance with the help of HA datasets in our weighted contrastive pre-training. (3) For the ACE05, the pre-training methods cannot outperform the FT baseline. To analyze this problem, we perform extra experiments on ACE05.

Methods	**ACE05 (no pronouns)**	
	25%	100%
FT	62.22	70.29
CIL + FT	63.31	69.76
RECN + FT	62.43	70.09
WCL + FT (ours)	**64.45**	**71.10**

Table 4: Evaluation on ACE05 after removing pronouns.

3.3 Further Analysis

We find that ACE05 contains many pronoun entities, for example, *"**He** lives in **America**."*. As pronoun entities such as *"**He**"* naturally come along with much more severe noise in DS, we also conduct extra experiments by removing sentences containing pronoun entities in ACE05 and the corresponding DS data to confirm the effect of pronouns.

After removing pronoun entities in ACE05, as shown in the Table 4, our model outperforms all baselines including FT, which indicates that pronoun entities bring mishandled noise in the pretraining stage and limit the effect of our DS data construction approach.

4 Conclusions

We introduce a weighted contrastive pre-training method by leveraging the HA dataset to estimate the reliability of instances in the abundant DS data. To alleviate the need for KBs, we also propose to construct DS data based on the triplets derived from the HA dataset for pre-training. Experimental results demonstrate that our proposed method outperforms SOTA work on target HA datasets.

One limitation of this work is that our method still needs a certain amount of HA data to achieve sufficient size of DS data and accuracy of reliability estimation. We leave it for future work to reduce reliance on HA data.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: Contrastive instance learning framework for distantly supervised relation extraction. 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 6191–6200, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel 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, Suntec, Singapore. Association for Computational Linguistics.

Róbert Ormándi, Mohammad Saleh, Erin Winter, and Vinay Rao. 2021. Webred: Effective pretraining and finetuning for relation extraction on the web. *CoRR*, abs/2102.09681.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online. Association for Computational Linguistics.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9507–9514.

Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. 2020. Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 236–246, Online. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Hyperparameter	Range	Best
Bag size	2-8	4
Batch size	8-32	16
Temperature	0.05-1.0	0.2

Table 5: Hyperparamter optimazition.

A Implementation Details

During the construction of DS data, we use the preprocessing tool NLTK to split raw corpora into sentences.

We use *bert-base-uncased* (Devlin et al., 2019) as the base encoders for ACE05, ACE05 (no pronouns) and Wiki20m, for a fair comparison with previous works. We also use *bluebert* (Peng et al., 2019) as the base encoder for i2b2 2010VA, since the SOTA performance is achieved based on this effective medical domain BERT.

For baseline models, we modify their official implementations to fit our experiments and follow the model settings in their papers. For our proposed method, the primary hyperparameters in the experiments are batch size, bag size, and contrastive learning temperature that directly influence the weighted contrastive learning loss, and we show our searching ranges and best values in Table 5.

We used 8 NVIDIA A100 for pre-training and 2 NVIDIA RTX3090 for fine-tuning.