A Simple Model for Distantly Supervised Relation Extraction

Ziqin Rao, Fangxiang Feng, Ruifan Li*, Xiaojie Wang
School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
{ziqin_rao, fxfeng, rfli, xjwang}@bupt.edu.cn

Abstract

Distantly supervised relation extraction is challenging due to the noise within data. Recent methods focus on exploiting bag representations based on deep neural networks with complex de-noising scheme to achieve remarkable performance. In this paper, we propose a simple but effective **B**ERT-based **G**raph convolutional network **M**odel (i.e., BGM). Our BGM comprises of an instance embedding module and a bag representation module. The instance embedding module uses a BERT-based pretrained language model to extract key information from each instance. The bag representaion module constructs the corresponding bag graph then apply a convolutional operation to obtain the bag representation. Our BGM model achieves a considerable improvement on two benchmark datasets, i.e., NYT10 and GDS[1].

1 Introduction

In the distant supervision relation extraction (DS-RE) setting, handling the noisy training data is a major challenge for downstream applications. To alleviate the severe noise problem in DS-RE, Riedel et al. (2010) incorporate the multi-instance learning (MIL) framework. Under this framework, the instances (i.e., sentences) for an identical entity pair are regarded as a bag. However, learning effective bag representations from the noisy data is a challenge resulting in unsatisfactory performance.

Recently, to obtain effective bag representations, various neural models are incorporated. Lin et al. (2016) propose a selective attention mechanism to capture relatively informative instances forming the bag representation. Vashishth et al. (2018) use graph convolution network (GCN) (Kipf and Welling, 2017) to encode syntactic information obtained from a dependency parser. However, the GCN's capacity in capturing the correlation among instances is not sufficiently explored.

Meanwhile, a prevalent trend is to use pretrained language models (PLMs) for various NLP tasks (Alberti et al., 2019; Tang et al., 2021; Wang et al., 2021; Li et al., 2021). PLMs work without explicit linguistic features and side-information like POS tags and entity types. Alt et al. (2019) use a PLM to incorporate more linguistic and semantic information. The correlations among instances are implicitly represented through a naïve attention mechanism. Very recently, Chen et al. (2021) propose PLMs combined with the contrastive instance learning (CIL) to build bag representations. CIL captures the correlation among instances through data augmentation with positive and negative pairs. These PLMs-based methods achieve a fabulous performance. Hence, a question arises, how about combining PLMs and GCNs to learn the instance correlations for bag representations?

Towards this goal, we propose an impressively simple model, i.e, BGM. Our BGM comprises of a PLM (e.g., BERT) and a concise GCN. The PLM brings accurate contextual representations for instances. The GCN whose nodes are instances encodes the correlation of instances in a bag through mutually aggregation. Thus, we finetune an off-the-shelf BERT with a GCN in an end-to-end fashion. With the obtained bag representation, the entity relation in a given sentence is then predicted.

Our contributions are twofold. **1**) We propose BGM model for DS-RE. Our BGM only comprises a PLM and a GCN but without any prior knowledge or data augmentation. **2**) Experimental results on two benchmark datasets show that BGM achieves a consistent improvement on performance.

2 Related Work

The noisy data is a major challenge in DS-RE for downstream applications. Previous works can be divided into two categories: PCNN-based methods

*Corresponding author.

[1]Code and datasets are available at https://github.com/ziqinrao.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2651–2657
October 12–17, 2022.

and PLMs-based methods. The former methods use Piecewise Convolutional Neural Networks (i.e., PCNN) as the backbone to encode sentences. Thus, various methods are proposed to acquire effective bag representations. Lin et al. (2016) propose the selective attention mechanism over the bag's instances to use more informative instances. This method implicitly models the instance correlations. Subsequently, Liu et al. (2017) propose a model by providing a better supervision with soft labels as golden labels. Han et al. (2018) exploit a hierarchical attention paradigm to better capture valid instances. Combined with previous intra-bag attention, Ye and Ling (2019) design an inter-bag attention to obtain bag-group representations. Cao et al. (2021) build a co-occurrence graph to learn embeddings to enhance bag representations. Shang et al. (2022) employ a pattern-aware self-attention network to automatically discover relational patterns for pre-trained transformers.

Recently, PLMs-based methods achieve remarkable performance in various NLP tasks, including DS-RE. Alt et al. (2019) adopt PLMs to incorporate a great deal of commonsense knowledge. Christopoulou et al. (2021) focus on relational tokens using the sub-tree parsing and capture informative instances with fine-tuning BERT. Chen et al. (2021) combine PLMs with contrastive learning with data augmentation to improve the overall performance. Note that the three PLMs-based methods follow the soft attention mechanism adopted in (Lin et al., 2016). The attention mechanism uses the bag's target relation to emphasize the instances which better express the bag relation.

In this paper, we propose a simple but effective model BGM. The BGM uses a PLM and a GCN for DS-RE. To the best of our knowledge, we are the first to use GCN to learn the bag representation *directly over the instances*. Note that Vashishth et al. (2018) also use GCNs but for encoding the syntactic tree of instances. In addition, compared with (Lin et al., 2016), we do not follow their selective attention with bag's target relation. Our BGM uses a self-attention mechanism to capture the correlation among instances.

3 Method

Suppose that an instance bag $\mathcal{B}^{(e^h,e^t)} = \{s_1, s_2, ..., s_{n^s}\}$ contains n^s instances which all include the entity-pair (e^h, e^t). Each instance $s = \{x_1, x_2, ..., x_{n^w}\}$ contains n^w words. The DS-RE task aims to extract the specific relation between the entity-pair (e^h, e^t) (i.e., the head and tail entities). To this aim, we design the BGM model shown in Figure 1, which includes an embedding layer and a bag representation layer.

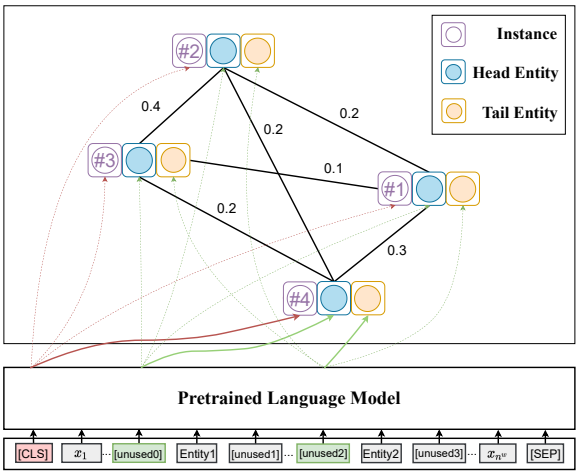

Figure 1: The overview of BGM of the embedding layer and the bag representation layer. Here, the BGM encodes a bag with four instances. For each instance, $[CLS]$, $[ununsed0]$ and $[unused2]$ are used.

3.1 Embedding Layer

To represent each instance in a bag, we use a BERT-based PLM as our embedding layer. Inspired by the role of $[CLS]$ in BERT, we adopt the $[unused*]$ to represent two involved entities. This effectively addresses the multi-word entities representation. Specifically, for each instance, the token sequence $\{[CLS], x_1, x_2, [unused0], e_h, [unused1], \cdots, [unused2], e_t, [unused3], x_{n^w}, [SEP]\}$ is fed into the BERT encoder. Then, three hidden states, i.e., $\{H_c, H_{u0}, H_{u2}\} \in \mathbb{R}^d$ which correspond to the tokens $[CLS]$, $[unused0]$, and $[unused2]$ are obtained for representing each instance and the corresponding two entities.

3.2 Bag Representation Layer

With the embedding layer, we then construct the bag graph $\mathcal{G} = \{\mathcal{V}, \mathcal{A}\}$ for each bag. The set of nodes $\mathcal{V}$ are initialized by concatenating the representations of instance and two entities, i.e., $[H_c; H_{u0}; H_{u2}]$. Thus, the dimension of node d_h equals $3d$. Moreover, the adjacency matrix $\mathcal{A} \in \mathbb{R}^{n^s \times n^s}$ is generated as,

$$\mathcal{A} = \text{softmax}\left(\frac{QW^Q \times \left(KW^K\right)^{\mathrm{T}}}{\sqrt{d_h}}\right) \quad (1)$$

where Q and K are both the concatenated representations of instances. Matrices W^Q and $W^K \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters.

Our GCN is updated by applying the classical method (Kipf and Welling, 2017). Suppose that $H^{(l)}$ denotes the input matrix of nodes of the l-th layer. The computation of next layer is given as,

$$H^{(l+1)} = \rho\left(\widetilde{\mathcal{A}} H^{(l)} W^{(l)} + b^{(l)}\right) \quad (2)$$

where $W^{(l)} \in R^{d_h \times d_h}$ is the trainable weight matrix, $b^{(l)}$ is the bias vector and ρ is an activation function (e.g., ReLU). To maintain the instance's original semantics, a self-connection is added to each node, i.e., $\widetilde{\mathcal{A}} = \mathcal{A} + I$; I is an identity matrix. Thus, for the last graph layer L, the node representation $H^{(L)}$ is obtained, which is used for constructing a bag representation.

3.3 Relation Prediction

We apply an average pooling on the bag representation $H^{(L)}$. A linear layer followed by a softmax layer is used to predict the relation $\hat{r}$ as follows,

$$\hat{r} = \text{softmax}\left(\text{MLP}(\text{AvgPooling}(H^{(L)}))\right). \quad (3)$$

The BGM is trained using a classical cross-entropy loss with gradient descent optimization.

4 Experiment

4.1 Dataset and Metric

Two benchmark datasets[2] are used. **NYT10** (Riedel et al., 2010) is generated by aligning the FreeBase's instances with NYT News Corpus. This dataset contains 39,528 entities and supports 53 types of relations (NA for no relation is included). Following previous works, NYT10 is split into 466,876/55,167/172,448 instances for training/validation/testing.

GDS (Jat et al., 2017) is recently built using Google RE corpus[3]. To meet MIL's *expressed-at-least-once* assumption, at least one sentence for each bag is correctly labelled which makes automatically evaluation more credible. The GDS dataset is officially split into 11,297/1,864/5,663 for training/validation/testing.

Metrics. We adopt four metrics, including precision-recall curve (**PR**), area under curve (**AUC**), Precision@N (**P@N**) values, and Micro-F1 score (**F1**) for evaluation.

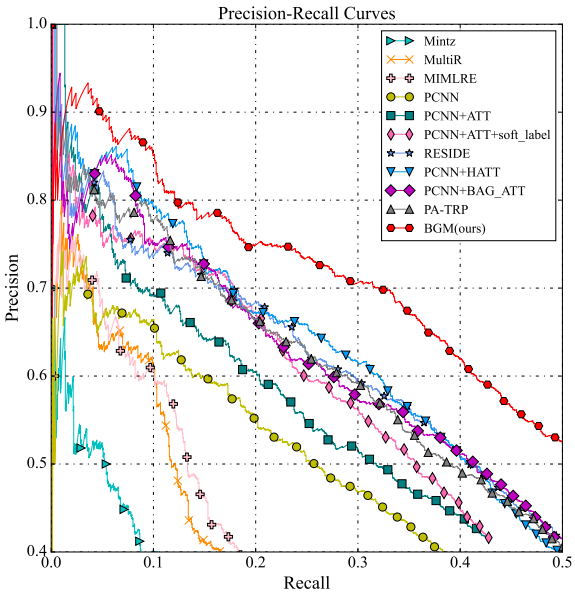

Figure 2: PR curves comparison on NYT10.

4.2 Implementations Detail

In our experiments, the BERT-base-uncased English version model[4] is used. The maximum of input sequence length of BERT is set to 120 and the hidden size is 768. The GCN has two layers. In addition, we apply drop-out rate p of 0.3 to GCN and 0.5 to all linear layers. The Adam optimizer (Kingma and Ba, 2015) is adopted to train the model with a learning rate of 2×10^{-5} and a batch size of 32 for up to 3 epochs. All experiments are conducted on an NVIDIA V100 GPU.

4.3 Baseline

We compare our BGM with the following baseline methods, including Mintz (Mintz et al., 2009), MultiR (Hoffmann et al., 2011), MIMLRE (Surdeanu et al., 2012), PCNN (Zeng et al., 2015), PCNN+ATT (Lin et al., 2016), PCNN+ATT+soft_label (Liu et al., 2017), BGWA (Jat et al., 2017), CNN+RL (Feng et al., 2018), DSGAN (Qin et al., 2018), RESIDE (Vashishth et al., 2018), PCNN+HATT (Han et al., 2018), PCNN+BAG_ATT (Ye and Ling, 2019), DISTRE (Alt et al., 2019), PA-TMR (Kuang et al., 2020), ToHRE (Yu et al., 2020), PA-TRP (Cao et al., 2021), SRKBP (Christopoulou et al., 2021), REDSandT (Christou and Tsoumakas,

[2]https://github.com/thunlp/OpenNRE

[3]https://research.googleblog.com/2013/04/50000-lessons-on-how-to-read-relation.html

[4]https://github.com/huggingface/transformers

Method	P@N							AUC
	100	200	300	500	1000	2000	MEAN	
PCNN+ATT (Lin et al., 2016)	73.0	68.0	67.3	63.6	53.3	40.0	60.9	34.1
BGWA (Jat et al., 2017)	76.0	74.0	-	-	-	-	-	36.7
CNN+RL (Feng et al., 2018)	79.0	73.0	-	-	-	-	-	37.4
DSGAN (Qin et al., 2018)	80.0	78.0	-	-	-	-	-	38.0
RESIDE (Vashishth et al., 2018)	81.8	75.4	74.3	69.7	59.3	45.0	67.6	41.5
PCNN+HATT (Han et al., 2018)	82.0	79.5	75.3	67.0	57.7	41.9	67.2	42.0
PCNN+BAG_ATT (Ye and Ling, 2019)	**91.8**	83.0	76.3	70.2	52.0	34.2	67.9	42.2
PA-TMR (Kuang et al., 2020)	83.0	79.0	-	-	-	-	-	43.7
ToHRE (Yu et al., 2020)	91.5	82.9	79.6	**74.8**	63.3	48.9	73.5	-
PA-TRP (Cao et al., 2021)	87.0	79.5	77.3	68.6	59.0	44.6	67.9	41.5
SRKBP (Christopoulou et al., 2021)	83.0	75.5	73.0	-	-	-	-	42.9
PSAN-RE (Shang et al., 2022)	79.2	71.1	66.8	65.9	60.4	48.1	65.2	43.8
DISTRE (Alt et al., 2019) ♯	68.0	67.0	65.3	65.0	60.2	47.9	62.2	42.2
REDSandT (Christou and Tsoumakas, 2021) ♯	78.0	-	73.0	67.6	-	-	-	42.9
CIL (Chen et al., 2021) ♯	90.1	86.1	**81.8**	-	-	-	-	50.8
Our BGM	90.3	**86.5**	80.0	74.6	**67.5**	**50.7**	**74.9**	**51.5**

Table 1: Comparison results on NYT10. The symbol ♯ denotes the PLMs-based methods.

2021), CIL (Chen et al., 2021), and PSAN-RE (Shang et al., 2022). Note that the three methods, including DISTRE, REDSandT, and CIL are PLMs-based.

4.4 Results and Analysis

On NYT10. Figure 2 plots the **PR curves** on the NYT10 dataset. We observe that compared with baseline models, our proposed BGM achieves better performance by a large margin. It means that our model could make full use of the training data and capture the critical information in noisy data.

Table 1 reports the **P@N** and **AUC** values on NYT10 dataset. We notice that our model achieves the best performance and outperforms the other baseline models in almost all metrics. Compared with the strong competitor CIL, our proposed model improves P@100 and P@200 by 0.2% and 0.5% respectively and improves AUC score by 1.4%. In addition, compared with ToHRE on P@MEAN, our model improves the score by 1.9%, i.e., 73.5 → 74.9 .

On GDS. Table 2 reports the results on GDS dataset. Our model achieves a comparable performance on the GDS dataset, in terms of P@100, P@200 and AUC, i.e., 100.0, 98.0 and 89.2. Our BGM achieves better performance compared with these baseline models.

Method	P@100	P@200	AUC
PCNN+ATT (Lin et al., 2016)	94.0	93.0	80.3
BGWA (Jat et al., 2017)	99.0	98.0	81.5
CNN+RL (Feng et al., 2018)	**100.0**	96.0	85.5
DSGAN (Qin et al., 2018)	99.0	97.0	84.5
RESIDE (Vashishth et al., 2018)	**100.0**	97.5	89.1
PCNN+HATT (Han et al., 2018)	99.0	97.0	85.4
PA-TMR (Kuang et al., 2020)	**100.0**	98.0	86.5
PA-TRP (Cao et al., 2021)	**100.0**	98.0	87.3
PSAN-RE (Shang et al., 2022)	97.0	**98.5**	**91.1**
Our BGM	**100.0**	98.0	89.2

Table 2: Comparison results on GDS.

Method	AUC	P@M	F1
Our BGM	**51.5**	**74.9**	**52.4**
BGM *w/o* GCN	46.7 (4.8↓)	68.3 (6.6↓)	51.6 (0.8↓)
BGM *w/o* EntCon	46.9 (4.6↓)	68.0 (6.9↓)	51.9 (0.5↓)

Table 3: Ablation study of our BGM on NYT10.

PLM	AUC	P@M	F1
Bert-based-uncased	51.5	**74.9**	52.4
Bert-based-cased	49.7	71.0	53.7
Bert-large-uncased	52.9	72.4	56.3
Distilbert-base-uncased	49.5	71.5	50.9
Xlnet-base-cased	47.5	67.1	53.1
Albert-based-v2	48.3	72.1	50.3
Roberta-base	48.8	68.5	52.7
Roberta-large	**53.2**	74.5	**57.1**

Table 4: Comparison of PLMs in BGM on NYT10.

4.5 Ablation Study

In Table 3, **BGM *w/o* GCN** degenerates our full BGM to a naïve BERT. BGM *w/o* GCN has a comparable performance. However, it fails to interrelate the instances for encoding bag representations, the overall performance drops, i.e., AUC ($51.5 \rightarrow 46.7$), P@Mean ($74.9 \rightarrow 68.3$) and F1 ($52.4 \rightarrow 51.6$). **BGM *w/o* EntCon** only uses the [CLS] token but not entity-aware instance representations. The AUC, P@M and F1 decrease. This shows that the entity information is essential for relation extraction. Moreover, to investigate the impact of various **PLMs** in our BGM, we replace the basic BERT-based with other representative PLMs. In Table 4, we observe that BGM with various PLMs as an embedding layer can achieve competitive performance. Due to larger parameters, Roberta-large achieves the best performance.

4.6 Case Study

We use two bags shown in Table 5 for case study on three methods, including BGM, BGM *w/o* GCN and BGM *w/o* EntCon. For **#1** bag, our BGM gives the relation, */location/country/administrative_divisions*. The other variants give the wrong relation, */location/location/contains*. The reason is that BGM *w/o* EntCon could not use the representations of key phrases "in the state of" in S3. Besides, BGM *w/o* GCN could not utilize the instance correlations in the bag. The captured information in S3 is not shared well with the other instances.

For **#2** bag, it expresses the relation */people/person/place_of_birth*. Our BGM *w/o* EntCon predicts the wrong relation, */people/person/place_lived*. In contrast, BGM and BGM *w/o* GCN can identify the golden truth. With the entity-aware instance representation, they capture the contextual information of entities embedded in "was born in". This helps the model focus on entities and capture the relation more effectively.

Bag	Instance
# 1	S1: ...she gazed at the work before her: ... and the landscape of the [Jalisco] region of [Mexico] .
	S2: ..., left his small ranch in the [Jalisco] region of [Mexico] for work in the promised land of the united states .
	S3: ...Italian real estate magnates who relocated to [Mexico] and built a series of sumptuous properties in the state of [Jalisco] that made it a magnet for the super-rich .
# 2	S1: ..., like Freddy Rodriguez's tribute to [Sammy Sosa] , who was born in the [Dominican Republic] , with a glass....

Table 5: Two bags from NYT10 for case study.

5 Discussion

Our BGM performs its calculation on the entire graph. The GCN layer calculates weights adopting the method of self-attention. In other words, the node features of the entire graph are updated after one calculation, and the learned parameters are not heavily related to the graph structure. Compared with other attention-based methods, we do not follow their selective attention with the bag's target relation. Our BGM uses a self-attention mechanism to capture the correlation among instances which are taken as the nodes of the graph. Therefore, GCN combined with self-attention is one of effective ways for the setting of DS-RE.

6 Conclusion

In this paper, we propose a simple but effective model, a.k.a. BGM based on PLMs and GCN for DS-RE. Each instance is represented using a BERT-based pre-trained language model. To capture the instance correlations, GCN for multiple instances within a bag is used. With this type of bag representation, a cross-entropy loss is applied for predicting the relation between entities. Extensive experiments on two benchmark datasets show the superior performance. In our future, we will investigate the hidden theory in-depth for better explainability of our BGM model. In addition, extending the BGM to dealing with the case of a single-instance bag is an interesting problem.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFF0303300 and Subject II under Grant 2019YFF0303302, and in part by the National Natural Science Foundation of China under Grant 62076032. The authors would also like to thank the editor and anonymous reviewers for their valuable comments on improving the final version of this paper.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

Yixin Cao, Jun Kuang, Ming Gao, Aoying Zhou, Yonggang Wen, and Tat-Seng Chua. 2021. Learning relation prototype from unlabeled texts for long-tail relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

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.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2021. Distantly supervised relation extraction with sentence reconstruction and knowledge base priors. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–26, Online. Association for Computational Linguistics.

Despina Christou and Grigorios Tsoumakas. 2021. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, 9:62574–62582.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 32.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245, Brussels, Belgium. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Sharmistha Jat, Siddhesh Khandelwal, and Partha P. Talukdar. 2017. Improving distantly supervised relation extraction using word and entity based attention. In *6th Workshop on Automated Knowledge Base Construction, AKBC@NIPS 2017, Long Beach, California, USA, December 8, 2017*. OpenReview.net.

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*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jun Kuang, Yixin Cao, Jianbing Zheng, Xiangnan He, Ming Gao, and Aoying Zhou. 2020. Improving neural relation extraction with implicit mutual relations. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pages 1021–1032. IEEE.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795, Copenhagen, Denmark. 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.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. DSGAN: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 496–505, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Yu-Ming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. 2022. A pattern-aware self-attention network for distant supervised relation extraction. *Information Sciences*, 584:269–279.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Hongyin Tang, Xingwu Sun, Beihong Jin, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Improving document representations by generating pseudo query embeddings for dense retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5054–5064. 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.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. UniRE: A unified label space for entity relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. 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 2810–2819, Minneapolis, Minnesota. Association for Computational Linguistics.

Erxin Yu, Wenjuan Han, Yuan Tian, and Yi Chang. 2020. ToHRE: A top-down classification strategy with hierarchical bag representation for distantly supervised relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1665–1676, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.