Denoising Distant Supervision for Relation Extraction with Entropy Weight Method

Mengyi Lu[1] and Pengyuan Liu[2]

Beijing language and culture university, Beijing,China
lmy0722@foxmail.com
liupengyuan@pku.edu.cn

Abstract. Distant supervision for relation extraction has been widely used to construct training set by aligning the triples of the knowledge base, which is an efficient method to reduce human efforts. However, this method inevitably suffers from wrong labeling problems leading too much noise that will severely hurt the performance of relation extraction. To tackle this problem, in this paper, we propose a denosing model based on Entropy Weight Method (EWM) to filter the noise and select most relevant sentences. First, in a pretraining stage, we develop a sentence-level relation aware attention mechanism to distinguish several most relevant sentence, increasing the attention weights for those critical sentences. Second, we filter the noisy sentences by calculating the entropy weight using the above attention matrix, and then we employ intra-bag and inter-bag attentions to aggregate these selected sentence representations. Experiments on the NYT dataset show that our method can significantly reduce the noisy instance and achieve the state-of-the-art model performance.

Keywords: Relation Extraction · Distant Supervision · Noise Filtering.

1 Introduction

Relation Extraction (RE) is defined as a task of generating relation triple facts from plain texts, which is widely used to facilitate a lot of Natural Language Processing (NLP) tasks including knowledge base construction [1] and question answering [2]. As the fully supervised RE approaches are limited by the consuming and labour intensive labeled training set, distant supervision strategy [1] is proposed as a promising approach to automatically create training data via aligning Knowledge Bases (KBs) with texts. The basic assumption of distant supervision is that if two entities e_1 and e_2 have a relation r in KBs, then all the sentences in corpus that contain these two entities will express this specific relation and will be labeled as the training instances of r. Although distant supervision is effective to label data automatically, it suffers from the noisy labeling problem.

To address the issue of noisy labeling, previous studies adopt multi-instance learning to consider the noises of instances(Riedel[1], Yao[1], and McCallum

2010; Hoffmann et al. 2011; Surdeanu et al. 2012; Zeng et al[11]. 2015; Lin et al. 2016;Ji et al. 2017). For example, Zeng et al.propose to combine multi-instance learning [15] with Piecewise Convolutional Neural Networks (PCNNs) to choose the most likely valid sentence and predict relations. In these studies, the training and test process is proceeded at the bag level, where a bag contains noisy sentences mentioning the same entity pair but possibly not describing the same relation. However, these methods unable to handle the sentence-level prediction and are sensitive to the bags with all noisy sentences which do not describe a relation at all.

In this paper, we proposed a novel approach of filtering sentences called a entroy weighted method (EWM) to distinguish the relevant sentence and alleviate negative effect of noisy labeling problem. The overall idea of the model is as follows. First, We use a pretrain strategy, In this stage, we extracts all sentence features using PCNNs and learns the weights of sentences by the relation aware attention module. We hope that the attention mechanism is able to selectively focus on the relevant sentences through assigning higher weights for valid sentences and lower weights for the invalid ones. In this way, the attention matrix can recognize multiple valid sentences in a bag, and we retrain the model to filter the noisy sentences by calculating the entropy weight using the above attention matrix(we will show more calculation details later). More specifically, For a bag, we first use PCNNs to extract each sentences feature vector, then compute the attention weight for each sentence through multiplying each possible relation which is utilized as the query, and according the attention weight we calculate the entropy weight setting a threshold to filter the noisy sentence, and then the bags represention compute by the weighted sum of the selected vaild sentence feature vectors, Furthermore, the representation of a group of bags in the training set which share the same relation label is calculated by weighting bag representations using a similarity-based inter-bag attention module. Finally, a bag group is utilized as a training sample when building our relation extractor.

Our contributions of this paper include:

- We introduce a denoise approach named the Entropy Weight Method which can select the most relevant sentences and filter those wrong labeling sentences, this strategy can effiectively reduce the noise and improve the model performance.
- We propose a two-step model to better capture different levels of structural information and fuse them for classification.
- Our method achieve the-state-of-art performance on the widely used New York Times (NYT) dataset [4] .

2 Related Work

Distant supervision for relation extraction, first introduced by Mintz et al.[3], automatically generates training data through heuristic alignment between a knowledge base and plain texts. Although distant supervision is an efficient way to scale relation extraction to a large number of relations, the basic assumption

used in the alignment is so strong that it will inevitably bring wrong labeling problem.

To alleviate noise, Hoffmann et al.[29] , Riedel et al. [4] and Surdeanu et al. [19] build multi-instance learning paradigms. Specifically, Riedel et al. [4] uses at-least-one assumption to resolve the problem. Hoffmann et al.[29] builds a probabilistic graphic model and intends to resolve multi-instance with over- lapping relations in distant supervision. Surdeanu et al.[19] trains a Bayesian framework by expectation maximization (EM) algorithm. In addition, researchers notice that the incompleteness of the knowledge base (i.e., Freebase) will result in the false negative problem and design a latent-variable approach (Ritter et al.[19]). Later, considering auto- matic feature engineering, Lin et al.[5] and Zeng et al. [11] integrate multi-instance learning model with PCNNs to extract relations on distantly supervised data. Although proved effective, MIL suffers from information loss problem because it ignored the presence of more than one valid instances in most bags. Recently attention mechanism attracted a lot of interests of researchers [26–28].Considering the flaw of MIL, Lin et al. [5] and Ji et al. [12] introduced bilinear and non-linear attention respectively into this task to make full use of supervision information by assigning higher weights to valid instances and lower weights to invalid ones. The two attention models significantly outperform MIL method. However, they suffer from noise residue problem because noisy sentences have harmful information but still have positive weights. The residue weights of noisy data mean that attention methods cannot fully eliminate the negative effects of noise.

3 Methodology

In this section, we present an overview of our model for distant supervised RE, as illustrated in Figure 1. Our model follows three steps. First, We train the attention matrix in pretraining stage, then we use the etropy weight method to filter sentences, Finally, we adopt the Ye et al. [23] intra-bag and inter-bag attention mechanism to classify relationships.

3.1 Sentence Encoder

Sentence encoder transforms the sentence into its distributed representation. First, words in a sentence are transformed into dense real-valued vectors. For word token w, we use pre-trained word embeddings as low dimension vector representation. Following Zeng et al. [8], we use position embeddings as extra position feature. We compute the relative distances between each word and two entity words, and transform them to real-valued vectors by looking up randomly-initialized embedding matrices. We denote the word embedding of word w by $w_w \in \mathbb{R}^{d^w}$ and two position embeddings by p_{w_1}, p_{w_2} The word representation is then composed by horizontal concatenating word embeddings and position embeddings:

$$s_w = [w_w; p_w^1; p_w^2].s_w \in \mathbb{R}^{(d^w+2\times d_p)} \tag{1}$$

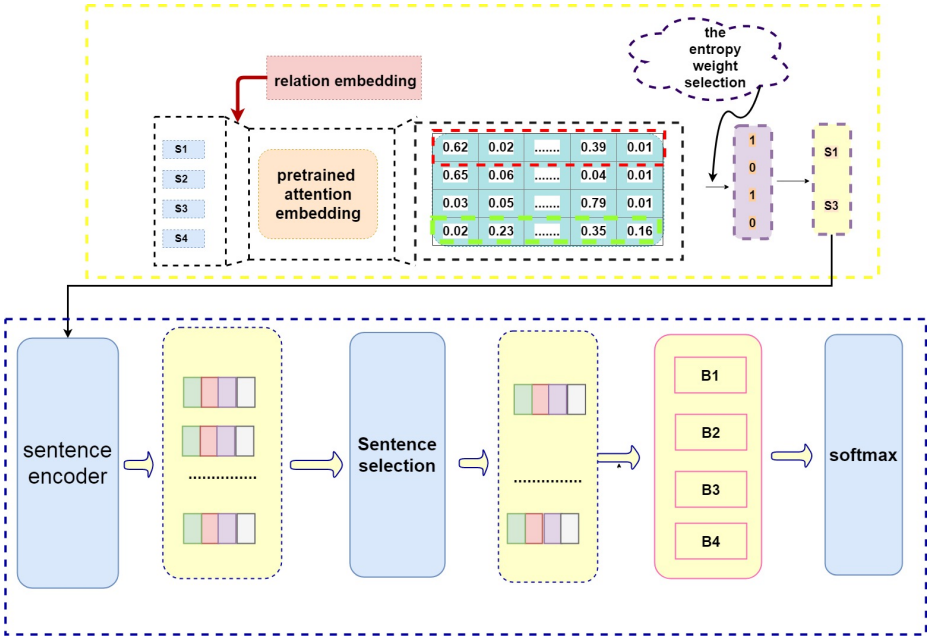

Fig. 1. The architecture of our model. It has two parts: (a) pretraining module and (b) overall framework.

Then, given a sentence and corresponding entity pair, we apply PCNN to construct a distributed representation of the sentence.

3.2 Pre-trained Attention Embedding

In our implementation, a pre-training strategy is adopted. We first train the model with relation-aware attention mechanism until convergence.we use PCNNs to extract each sentences feature vector, then compute the attention weight for each sentence through multiplying each possible relation which is utilized as the query,Therefore we pre-train the attention embedding in relation extraction, and use the pre-trained attention embedding to calculate the Entropy Weight. In this way, the Entropy Weight will more accurately reflect the information contained in a sentence bag for its relations.

3.3 Sentence Selection

Although previous work yields high performance, there still exists some drawbacks. MIL suffers from information loss problem because it ignored multiple valid sentences and used only one sentence for representing a bag and training. Attention-based methods have noise residue problem because they assigned

small but still positive weights to harmful noisy sentences, which means noise effects werent completely removed.

In our work, we narrow our sentence space and focus on select the sentences that are likely to be relevant to the relation r. We use the following formulas as our selection mechanism.

The Entroy Weight Method: In information theory, entropy is a measure of uncertainty. The more information, the less uncertainty, the less entropy; The smaller the amount of information, the greater the uncertainty, and the greater the entropy. According to the characteristics of entropy, the randomness and dispersion degree of sentences can be judged by calculating entropy value. We use this idea to simulate the correlation degree between sentences and relationships. The irrelevance with sentence between relationship is defined as a degree of dispersion. The greater the dispersion degree of the is, the greater its influence (weight) on evaluation is, and the smaller its entropy is. vice versa. We calculate it according to the following formula.

We adopt the weight matrix obtained by pre-training and take the maximum and minimum values of each row:

$$s_{ij} = \frac{a_{ij} - min(a_i)}{max(a_i) - min(a_i)} \tag{2}$$

where s_{ij} is the embedding normalization by attention value.

$$E_i = -\frac{1}{ln(n)} \sum_{i=1}^{n} p_{ij} ln(p_{ij}) \tag{3}$$

where $p_{ij} = \frac{s_{ij}}{\sum_{i=1}^{n} s_{ij}}$. E_i means the sentence information entropy.

$$M = E_w = \frac{e^{E_i}}{\sum_{i=1}^{n} e^{E_i}} \tag{4}$$

we take a threshold M, if the entropy value is greater than the threshold meaning that the sentence dispersion degree is smaller, the degree of irrelevance of sentences to relationships is smaller, and it will be reserved. If it is less than the threshold, the sentence dispersion degree is greater, it is considered as noise to discard it. We alleviate noise residue problem by only assigning attention weights to selected sentences. The unselected noisy data will not be assigned weights, and will not participate in training process.

3.4 Intra-Bag attention

We hope that the attention model can learn higher weights for valid instances and lower weights for the invalid ones. In experiments, we will show the weights of an example. we use $S_i \in \mathbb{R}^{m_i \times 3d_c}$ represent the sentences represention within bag b_i,where $R \in \mathbb{R}^h \times 3d_c$ denote relation embedding matrix where h is the

number of relations. $S_1, S_2, ..., S_q$ are feature vectors (computed by PCNNs) of all instances in a bag,we propose the following formulas to compute the attention weight. For each sentence j in the bag i with respect to each relation k, and aggregates to bag represention as b_k^i , by the following equations:

$$b_k^i = \sum_{j=1}^{m} a_{kj}^i s_j^i \tag{5}$$

where k is the relation index and a_{kj}^i is the attention weight between the k-th relation and the j-th sentence in bag b_i.and the a_{kj}^i can be calculated as:

$$a_{kj}^i = \frac{exp(e_{kj}^i)}{\sum_{j=1}^{m_i} exp(e_{kj}^i)} \tag{6}$$

where $e_{kj}^i$is the matching degree between the k-th relation query and the j-th sentence in bag b_i,and it is defined as:

$$e_{kj}^i = r_k s_j^i \top \tag{7}$$

where r_k is the k-th row of the relation embedding matrix R.

3.5 Inter-Bag attention

Inspired by the works of Ye et al. [23] and Yuan et al. [24] which utilizes bag-level attention mechanism to deal with the noisy bag problem. At the sentence level, noise cannot be completely removed because it assumes that at least one correct sentence exists. However, in the process of distant supervising the construction of the data set, there may be not exist a correct sentence in a bag, furthermore, the knowledge base cannot be covered all the relationships. The entity expressing the truely relationships may not exist in a given corpus, so there will still exist noise. To solve the problem, we combine several sentence bags of the same relation type and to get more attention to the more relevant bags. we obtain the superbag representation as follow equation:

$$f = \sum_{i=1}^{m} \gamma_{ik} b_k^i \tag{8}$$

$$\gamma_i = \frac{e^{(S(r_k, b_{ik})}}{\sum_{j=1}^{m} e^{(S(r_k, b_{ik}))}} \tag{9}$$

$$S_{i,j,k} = \frac{x_{i,j} r_k}{|x_{i,j}| \, |r_k|} \tag{10}$$

where b_k^i is the bag representation w.r.t. B_i for the k-th relation and r_k is the attention parameter corresponding to the j-th relation.

4 Experiment

Our experiments are intended to show that our model can capture high weight sentences and take full advantage of informative sentences for distant supervised relation extraction. In the experiments, we first introduce the dataset and evaluation metrics used. Next, we determine some parameters of our model. And then we evaluate the effects of our model performance, and we also compare our method to some classical methods. Finally,we do some experimental analysis and case study.

4.1 Dataset and Evaluation Metrics

We evaluate our model on a widely used dataset which is developed by Riedel et al.[2] . This dataset was generated by aligning Freebase with the New York Times (NYT) corpus. The dataset contains 53 relations (including no relation NA) and 39,528 entities. The training data contains 522,611 sentences, 281,270 entity pairs and 18,252 relational facts. The test dataset contains 172,448 sentences, 96,678 entity pairs and 1,950 relational facts. We use word2vec to train word embedding on the NYT corpus and use the embeddings as initial values. Following previous work [1, 2, 5, 8], we evaluate our model in the held-out evaluation, which evaluates our model by comparing the extracted relation facts with those in Freebase, and report both the precision/recall curves and Precision@N (P@N) of the experiments.

4.2 Experimental Settings

In this section, we study the influence of one parameter on our model: the threshold value M is defined in Equation (4). We tune our models using three-fold validation on the training set. We use a grid search to determine the optional parameter: $M \in \{0, 0.3, 0.5, 0.7, 1\}$. Most of them followed the hyperparameter settings in Ye et al. [23].

4.3 Performance Evaluation

We compare our method with these previous works: PCNN+ONE [11] selects the sentence with the highest right probability as bag representation; PCNN+ATT [5] use non-linear attention to assign weights to all sentences in a bag. Ye et al. [23] adopt the intra-bag and inter-bag attention modules achieve the state-of-the-art performance denoted PCNN+ATT_RA+BAG_ATT.ATT_RA means the relation-aware intra-bag attention method, and BAG_ATT means the inter-bag attention method.

In order to prove the superiority of our module, we propose a more intuitive and simpler way for instance selection: we set a threshold on attention weights calculated by entropy weight and filter sentences with lower weights than threshold. We denote this method as Entropy Weight Method (EWM). We

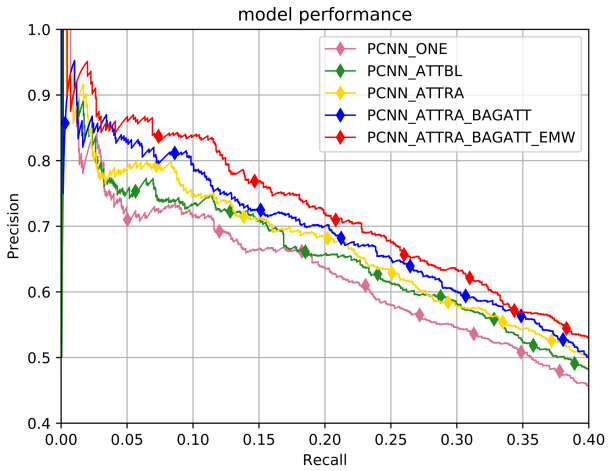

Fig. 2. The precision/recall curves for the combined model and the baselines.

adopt EWM to PCNN+ATT_RA+BAG_ATT to demonstrate the effectiveness of instance selectors, denoted as PCNN+ATT_RA+BAG_ATT+EWM. Figure2 shows the aggregated precision/recall curves, and Table1 shows the Precision@N with N = {100,200,300,ALL} of our approaches and all the baselines.

We can see our proposed methods achieved highest P@N values than all previous work. Furthermore,we have the following observations: (1) Similar to the results of Ye et al. [23], ATT_RA outperformed ATT_BL. It can be attributed to that the ATT _BL method only considered the target relation when deriving bag representations at training time, while the ATT_RA method calculated intra-bag attention weights using all relation embeddings as queries, which improved the flexibility of bag representations. (2) The BAG_ATT method performs better than the ones without BAG_ATT, it verified the effectiveness of the method. (3) The EWM method is efficient for this task. Especially when the test data set is all, the effect is more obvious. The results show that the EWM method can capture high weight sentences and take full advantage of informative sentences.

4.4 Analysis of Entropy Weight Threshold

In this section, we will investigate the effectiveness of the entropy weight threshold as denoted EWM. We fine-tune the hyperparameter threshold to achieve its best performance. Higher thresholds bring back information loss problem because more informative sentences are neglected. Lower thresholds bring back noise residue problem because more noisy sentences are selected and assigned weights. We conduct experiments on EWM with different thresholds. For clarify, we use a histogram to approximate precision/recall curves of different thresholds, shown in Fig.2. In our experiment, we found EWM value is set by 0.3 achiev-

Table 1. Top-N precision (P@N) for relation extraction in the entity pairs with different number of sentences. Following (Lin et al., 2016 [5]), One, Two and All test settings random select one/two/all sentences on the bags of entity pairs from the testing set which have more than one sentence to predict relation.

Method	one				two				all			
	100	200	300	mean	100	200	300	mean	100	200	300	mean
PCNN+ONE(Zeng 2015)	66.7	62.8	54.8	61.4	71.3	68.3	62.2	67.2	70.2	68.1	61.4	66.5
PCNN+ATT((Lin 2016))	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
PCNN+ATT_BL	78.6	73.5	68.1	73.4	77.8	75.1	70.3	74.4	80.8	77.5	72.3	76.9
PCNN+ATT_RA	79.2	73.9	68.3	73.8	81.5	77.5	72.7	77.5	83.6	79.9	72.3	78.6
PCNN+ATT_BL+BAG_ATT	84.8	**78.9**	70.7	78.1	84.5	79.5	74.2	79.4	88.8	83.9	77.3	83.3
PCNN+ATT_RA+BAG_ATT	86.8	77.6	73.8	**79.4**	90.8	**79.1**	74.4	81.4	91.8	83.9	77.6	84.4
PCNN+ATT_BL+EWM	78.8	72.6	67.9	73.1	77.2	72.4	66.9	72.1	81.2	77.0	71.7	76.6
PCNN+ATT_RA+EWM	79.2	73.6	68.2	73.6	81.6	77.7	70.3	76.5	83.8	84.0	78.7	84.8
PCNN+ATT_RA+BAG_ATT+EWM	**86.9**	77.6	**73.9**	79.3	**91.1**	78.9	**74.5**	**81.5**	**91.8**	**84.6**	**78.9**	**85.1**

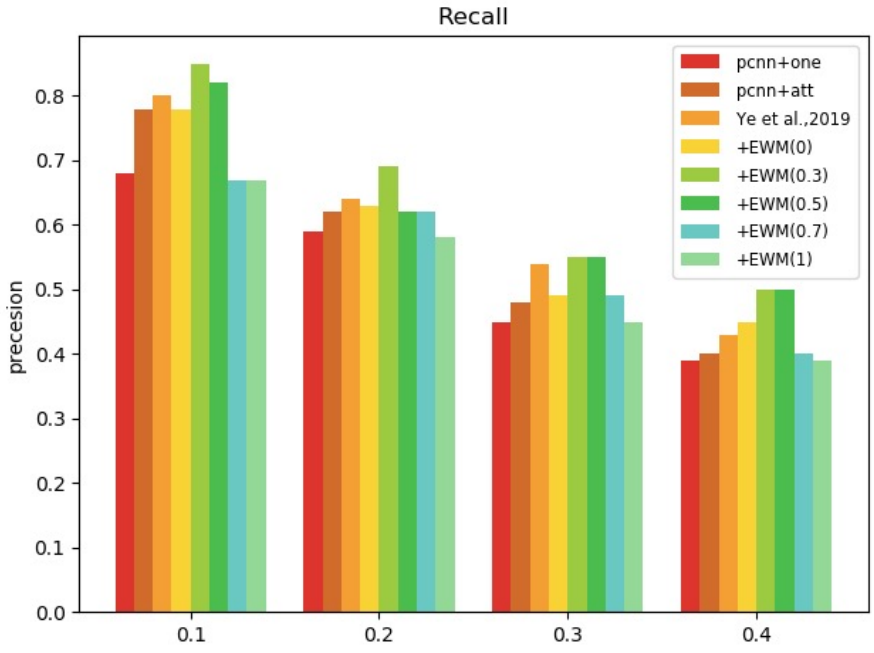

Fig. 3. Aggregate precision/recall histogram of EWM with different thresholds .

ing the best performance, with higher threshold the performance will decline, In this way, some relevant sentences are filtered out and there is less effective information to be utilized. EWM(1.0) selects only the sentence with maximum attention weight to train. Similar to MIL select strategy, EWM(1.0) also has similar performance to PCNN+ONE. when the threshold set lower, more noisy sentences get involved, so the performance behave dissatisfactory. The result also close to PCNN+ATT model (equivalent to EWM(0)).

4.5 Case study

Table 2 shows an example of our method selection result. The bag contains 4 sentences which the 4-th instance are invalid sentence. With the help of EWM,

Table 2. An example of Entropy Weight.

tuple	instance	select	EMW
/location/location /contains (New Orleans, Dillard University)	1. She graduated from [Dillard University] in [New Orleans] and received a masters degree in marine science from the College of William and Mary.	1	0.23
	2. Jinx Broussard, a communications professor at [Dillard University] in [New Orleans], said four members of her family had lost their houses to the hurricanes.	1	0.37
	3. I was grieving from the death when I graduated from high school, but I decided to go to [Dillard University] in [New Orleans]	1	0.31
	4.4. When he came here in May 2003 to pick up an honorary degree from [Dillard University], his dense schedule didnt stop him from calling Dooky Chases, the Creole restaurant he sang about in Early in the Morning Blues, where hed eaten his favorite dish ever since he lived in [New Orleans] in the 1950s.	0	0.09

attention mechanism only assigns high weights to selected sentences. The fourth sentence was a noisy sentence because the sentence in this bag didnt express the relation */location/location/contains* between the two entities $NewOrleans$, and $DillardUniversity$. Therefore, the attention mechanism can select the valid instances and is useful in our task.

5 Conclusion

In this paper, we proposed a novel approach of filtering sentences called a entroy weighted method (EWM) to distinguish the relevant sentence and alleviate negative effect of noisy labeling problem in distant supervision relation extraction.

Experimental results show our method is able to selectively focus on the relevant sentences through assigning higher weights for valid sentences and lower weights for the invalid ones.

Acknowledgements

This work is supported by Beijing Natural Science Foundation(4192057).

References

1. Han, X., Liu, Z., Sun, M.: Neural knowledge acquisition via mutual attention between knowledge graph and text (2018)
2. Lee, C., Hwang, Y.G., Jang, M.G.: Fine-grained named entity recognition and relation extraction for question answering. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 799800. ACM (2007)
3. M. Mintz, S. Bills, R. Snow, D. Jurafsky, "Distant supervision for relation extraction without labeled data", Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. (AFNLP) Assoc. Comput. Linguistics, pp. 1003-1011, Aug. 2009.
4. S. Riedel, L. Yao, A. McCallum, "Modeling relations and their mentions without labeled text", Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, pp. 148-163, 2010.
5. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: 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), vol. 1, pp. 21242133 (2016)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
7. Santos, C.N.d., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:1504.06580 (2015)
8. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 23352344 (2014)
9. Nguyen, T.H., Grishman, R.: Relation extraction: perspective from convolutional neural networks. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp. 3948 (2015)
10. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
11. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 17531762 (2015)
12. Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In AAAI, pages 30603066.

13. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 207212 (2016)
14. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Pro- cessing Systems, pages 59986008.
15. Dietterich, T.G., Lathrop, R.H., Lozano-P erez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. intell. 89(12), 3171 (1997)
16. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balc azar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8 10
17. R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations", Proc. 49th Annu. Meeting Assoc. Comput. Linguistics Hum. Lang. Technol., pp. 541-550, Jun. 2011.
18. A. Ritter, L. Zettlemoyer, O. Etzioni, "Modeling missing data in distant supervision for information extraction", Trans. Assoc. Comput. Linguistics, vol. 1, pp. 367-378, Oct. 2013.
19. M. Surdeanu, J. Tibshirani, R. Nallapati, C. D. Manning, "Multi-instance multi-label learning for relation extraction", Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. Assoc. Comput. Linguistics, pp. 455-465, Jul. 2012.
20. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 12011211. Association for Computational Linguistics (2012)
21. Sorokin, D., Gurevych, I.: Context-aware representations for knowledge base relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 17841789 (2017)
22. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 207212 (2016)
23. Ye, Zhixiu, and Zhenhua Ling. "Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions.." arXiv: Computation and Language (2019).
24. Yuan, Yujin, et al. "Cross-relation Cross-bag Attention for Distantly-supervised Relation Extraction." national conference on artificial intelligence (2019).
25. Dietterich, T.G., Lathrop, R.H., Lozano-P erez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. intell. 89(12), 3171 (1997)
26. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
27. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
28. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 22042212 (2014)
29. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of ACL (pp. 541550).