Distant Supervision for Relation Extraction with Neural Instance Selector

Yubo Chen[1(✉)], Hongtao Liu[2], Chuhan Wu[1], Zhigang Yuan[1], Minyu Jiang[3], and Yongfeng Huang[1]

[1] Next Generation Network Lab, Department of Electronic Engineering, Tsinghua University, Beijing, China
{yb-ch14,wuch15,yuanzg14}@mails.tsinghua.edu.cn, yfhuang@tsinghua.edu.cn
[2] Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin, China
htliu@tju.edu.cn
[3] Fan Gongxiu Honor College, Beijing University of Technology, Beijing, China
ryancoper@emails.bjut.edu.cn

Abstract. Distant supervised relation extraction is an efficient method to find novel relational facts from very large corpora without expensive manual annotation. However, distant supervision will inevitably lead to wrong label problem, and these noisy labels will substantially hurt the performance of relation extraction. Existing methods usually use multi-instance learning and selective attention to reduce the influence of noise. However, they usually cannot fully utilize the supervision information and eliminate the effect of noise. In this paper, we propose a method called Neural Instance Selector (NIS) to solve these problems. Our approach contains three modules, a sentence encoder to encode input texts into hidden vector representations, an NIS module to filter the less informative sentences via multilayer perceptrons and logistic classification, and a selective attention module to select the important sentences. Experimental results show that our method can effectively filter noisy data and achieve better performance than several baseline methods.

Keywords: Relation extraction · Distant supervision
Neural Instance Selector

1 Introduction

Relation extraction is defined as finding relational sentences and specify relation categories from plain text. It is an important task in the natural language processing field, particularly for knowledge graph completion [7] and question answering [10]. Distant supervision for relation extraction aims to automatically label large scale data with knowledge bases (KBs) [14]. The labeling procedure is as follows: for a triplet (e_{head}, e_{tail}, r) in KB, all sentences (instances) that simultaneously mention head entity e_{head} and tail entity e_{tail} constitute a *bag* and are labeled as relation r.

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11108, pp. 209–220, 2018.
https://doi.org/10.1007/978-3-319-99495-6_18

However, distant supervised relation extraction is challenging because the labeling method usually suffers from the *noisy labeling problem* [17]. A sentence may not express the relation in the KB when mentioning two entities. Table 1 shows an example of the noisy labeling problem. Sentence 2 mentions *New Orleans* and *Dillard University* without expressing the relation */location/location/contains*. Usually, the existence of such noisy labels will hurt the performance of relation extraction methods. Thus, it's important to eliminate such noise when constructing relation extraction models.

Several approaches have been proposed to eliminate negative effects of noise instances. For example, Riedel et al. [17] proposed to use graphical model to predict which sentences express the relation based on the *at-least-once* assumption. Zeng et al. [24] propose to combine multi-instance learning [4] with Piecewise Convolutional Neural Networks (PCNNs) to choose the most likely valid sentence and predict relations. However, these methods ignore multiple informative sentences and only select one sentence from each bag for training. Therefore, they cannot fully exploit the supervision information.

In recent years, attention mechanism is introduced to this task to select information more effectively. Lin et al. [11] and Ji et al. [9] used bilinear and non-linear form attention respectively to assign higher weights to valid sentences and lower weights to invalid ones. Then the bag is represented as a weighted sum of all sentences' representations. However in these methods, the softmax formula of attention weights will assign positive weights to noisy data. These positive weights of noisy sentences violated the intuition that noisy sentences cannot provide relational information. Thus such attention based models can't fully eliminate the negative effect of noise.

In this paper we proposed a method called Neural Instance Selector (NIS) to further utilize rich supervision information and alleviate negative effect of noisy labeling problem. Our approach contains three modules. A sentence encoder transforms input texts into distributed vector representations with PCNN. A Neural Instance Selector filters less informative sentences with multilayer perceptrons and logistic classification. The NIS module can select multiple valid sentences, and exploit more information than MIL method. A selective attention module selects more important sentences with higher weights. In order to further

Table 1. An example of noisy labeling problem. The bold words are head/tail entities.

Triplet	Instances	Noisy?
(New Orleans, Dillard University, /location/location/contains)	1. Jinx Broussard, a communications professor at **Dillard University** in **New Orleans**, said ...	No
	2. When he came here in May 2003 to pick up an honorary degree from **Dillard University**, his dense schedule didn't stop him ... ever since he lived in **New Orleans** in the 1950's	Yes

eliminate noise effects than attention-based methods, we only assign attention weights to selected sentences. Experimental results on the benchmark dataset validate the effectiveness of our model.

2 Related Work

Early works focused on *feature-based* methods for relation extraction. GuoDong et al. [6] explored lexical and syntactic features with textual analysis and feed them into a SVM classifier. Bunescu et al. [3] connected weak supervision with multi-instance learning [4] and extend it to relation extraction. Riedel et al. [17] proposed *at-least-once* assumption to alleviate the wrong label problem. However, these methods lack the ability of fully utilizing supervision information and suppress noise. Besides, these methods cannot effectively use the contextual information.

Recent works attempt to use neural networks for *supervised* relation extraction. Socher et al. [19] represented words with vectors and matrices and use recursive neural networks to compose sentence representation. Zeng et al. [25], Nguyen et al. [16] and dos Santos et al. [18] extracted sentence level vector representation with CNNs. Other work adopted recurrent neural networks to this task [20,26]. However, these methods need sentence-annotated data, which cannot be applied to large scale corpus without human annotation.

In order to apply neural networks to *distant supervision*, Zeng et al. [24] proposed PCNN to capture sentence structure information, and combined it with Multi-Instance Learning [4] (MIL) to select the sentence with the highest right probability as bag representation. 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 [1,12,15,22]. Considering the flaw of MIL, Lin et al. [11] and Ji et al. [9] 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.

Different from MIL and attention methods, we propose a method named NIS to further solve the information loss and noise residue problem. First, We use PCNNs [24] to learn sentence representations. Second, an NIS module takes all sentences' representations in a bag as input, and uses a MLP to capture the information of noise. Third, a logistic classifier takes MLP output to select valid sentences and filter noisy ones. The NIS module can alleviate information loss problem by retaining more than one valid sentences. Finally, we assign attention weights to selected sentences and use them to compute bag representation. In this way noise residue problem is reduced by avoiding assigning weights to unselected sentences. Experimental results show that the NIS module can alleviate these two problems and bring better performance to baseline models.

3 Methodology

In this section, we will introduce our method. Our framework contains three parts, which is shown in Fig. 1. We will introduce these parts each by each.

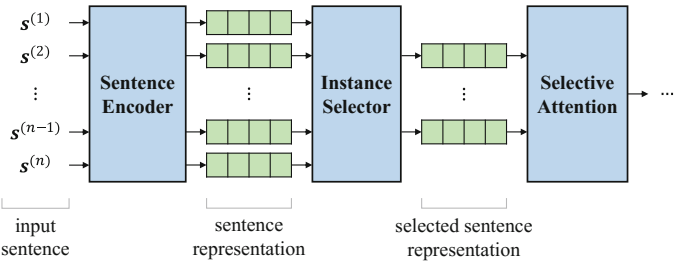

Fig. 1. Details of Instance Selector framework.

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. [24], 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 $\mathbf{w}_w \in \mathbb{R}^{d_w}$ and two position embeddings by $\mathbf{p}_w^{(1)}, \mathbf{p}_w^{(2)} \in \mathbb{R}^{d_p}$. The word representation $\mathbf{s}_w$ is then composed by horizontal concatenating word embeddings and position embeddings:

$$\mathbf{s}_w = [\mathbf{w}_w; \mathbf{p}_w^{(1)}; \mathbf{p}_w^{(2)}]. \quad \mathbf{s}_w \in \mathbb{R}^{(d_w + 2 \times d_p)} \tag{1}$$

Then, given a sentence and corresponding entity pair, we apply PCNN to construct a distributed representation of the sentence. Compared with common CNN, PCNN uses a piecewise max-pooling layer to capture sentence structure information. A sentence is divided into three segments by two entity words, then max-pooling is executed on each segment respectively. Following Zeng et al. [24], we apply tanh as activation function. We denote convolution kernel channels by c, and the output of PCNN by $\mathbf{f}^{(i)} \in \mathbb{R}^{3c}$.

3.2 Instance Selector

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

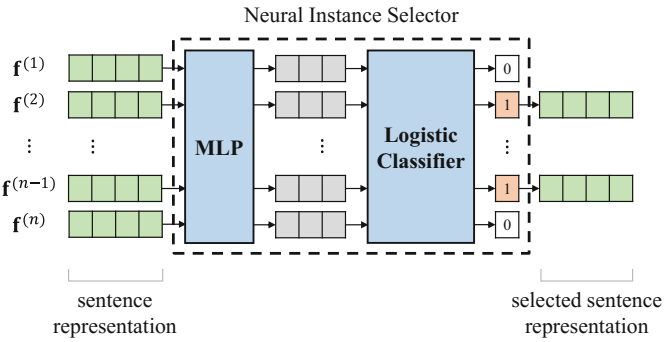

Fig. 2. The structure of NIS module. The 1s among the outputs of Logistic Noise Classifier indicate the corresponding sentence is valid, and 0s indicates invalid.

small but still positive weights to harmful noisy sentences, which means noise effects weren't completely removed.

In order to alleviate these two negative effects, we propose a method called **Neural Instance Selector (NIS)** to pick out more informative sentences. Figure 2 shows the structure of NIS. We use a small neural network to classify valid and invalid sentences. The core component of NIS is a Multilayer Perceptron (MLP) used for capturing information of noise. Then MLP output vectors are fed into a logistic noise classifier to produce sentence-level selection results. As shown in Fig. 2, NIS module has the ability of retaining multiple valid sentences, naturally reducing information loss problem. Then 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.

One alternative way for selecting instances is removing MLP and directly feeding PCNN output into logistic classifier [5]. However, we discover that this choice performs worse than NIS. This is because noise information is more complex than relation information, therefore requires deeper structure to be captured. The MLP can improve the non-linear fitting ability of instance selector. Also, the sentence level classifier has many alternatives. We conduct experiments on logistic classifier and two-class softmax classifier, and choose logistic classifier because of its better performance.

3.3 Selective Attention

The object of attention mechanism is to learn higher weights for more explicit instances and lower weights for less relevant ones. Attention-based models represent the ith bag M_i (with label y_i) as a real-valued vector $\mathbf{r}_i$. We denote the jth sentence's representation in the ith bag as $\mathbf{f}_i^{(j)}$. Previous work used bilinear [11] and non-linear form [9] attention. Considering computational efficiency and effectiveness, we choose the non-linear form in our method, denoted as **APCNN**.

Intuitively, relation information is useful when recognizing informative sentences. So we introduce relation representation and concatenate it with sentence representation to compute attention weights. Inspired by translation-based knowledge graph methods [2,21], the relation r is represented as the difference vector of entity word embeddings: $\mathbf{v}_{rel} = \mathbf{w}_{e_1} - \mathbf{w}_{e_2}$. Then $\alpha^{(j)}$ is computed through a hidden layer:

$$\alpha^{(j)} = \frac{\exp(e^{(j)})}{\sum_k \exp(e^{(k)})}, \tag{2}$$

$$e^{(j)} = \mathbf{W_a}^T(\tanh[\mathbf{f}_i^{(j)}; \mathbf{v}_{rel}]) + b_a. \tag{3}$$

With attention weights $\alpha^{(j)}$ computed, M_i is represented by $\mathbf{r}_i = \sum_{j=1}^{|M_i|} \alpha^{(j)} \mathbf{f}_i^{(j)}$. The bag representation is fed into a softmax classifier to predict relations and compute cross-entropy objective function $J(\theta) = -\sum_{i=1}^{T} \log p(y_i|\mathbf{o}_i;\theta)$:

4 Experiments

4.1 Dataset and Evaluation Metrics

We evaluate our NIS mechanism on the dataset developed by Riedel et al. [17] by aligning Freebase triplets with the New York Times (NYT) corpus. The training data is aligned to the years of 2005–2006 of the NYT corpus, and the testing to year of 2007. The dataset contains 53 relations (including 'NA' for no relation) and 39,528 entity pairs. Training data includes 522,611 sentences, the test set includes 172,448 sentences.

Following Lin et al. [11], we evaluate our method in the held-out evaluation. It provides an approximate measure of precision without time-consuming manual evaluation. We report the aggregate precision/recall curve and Precision@N in our experiments.

4.2 Parameter Settings

In our experiments, we use *word2vec*, proposed by Mikolov et al. [13], to pre-train word embeddings on NYT corpora. We select the dimension of word embedding d_w among {50, 100, 200, 300}, the dimension of position embedding d_p among {5, 10, 20}, the number of feature maps c among {100, 200, 230}, batch size among {50, 100, 150, 160, 200}. The best configurations are: $d_w = 50$, $d_p = 5$, $c = 230$, the batch size is 100. We choose MLP hidden layers' dimensions as [512, 256, 128, 64]. We use dropout strategy [8] and Adadelta [23] to train our models. For training, we set the iteration number over the training set as 20, and decay the learning rate every 10 epochs.

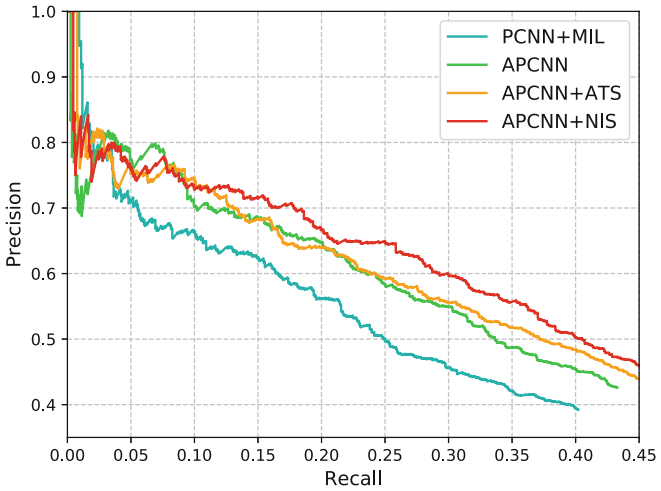

Fig. 3. Aggregate precision/recall curves for PCNN+MIL, APCNN, APCNN+ATS, APCNN+NIS. APCNN denotes the non-linear attention method proposed by Ji et al. [9]. We choose $0.8 \times \max(\textit{attention weight})$ as APCNN+ATS threshold with the highest performance in our experiments.

4.3 Performance Evaluation

We compare our method with two previous works: **PCNN+MIL** [24] selects the sentence with the highest right probability as bag representation; **APCNN** [9] use non-linear attention to assign weights to all sentences in a bag. In order to prove the superiority of our NIS module, we propose a more intuitive and simpler way for instance selection: we set a threshold on attention weights and filter sentences with lower weights than threshold. We denote this method as **Attention Threshold Selector (ATS)**. We adopt both ATS and NIS to APCNN to demonstrate the effectiveness of instance selectors, denoted as **APCNN+ATS** and **APCNN+NIS** respectively. Figure 3 shows the aggregated precision/recall curves, and Table 2 shows the Precision@N with $N = \{100, 200, 500\}$ of our approaches and all the baselines. From Fig. 3 and Table 2 we have the following observations:

Table 2. Precision@N of PCNN+MIL, APCNN, APCNN+ATS, APCNN+NIS.

Precision@N (%)	Top 100	Top 200	Top 500	Average
PCNN+MIL	71.72	67.84	61.62	66.89
APCNN	78.79	76.38	66.33	73.83
APCNN+ATS	73.74	76.38	65.53	71.88
APCNN+NIS	**78.79**	**76.38**	**69.94**	**75.04**

1. For both ATS and NIS, the instance selector methods outperform PCNN+MIL. It indicates that instance selectors can alleviate information loss problem because it can pick out more than one valid sentences in a bag.
2. Figure 3 shows that the instance selectors bring better performance compared with APCNN for both ATS and NIS method on high recall range. This is because the attention weights are assigned only to selected sentences. It indicates that our method can reduce noise residue problem because the weights of unselected sentences are masked as zero.
3. The NIS method achieves the highest precision over most of the entire recall range compared to other methods including the ATS. It indicates that NIS method can effectively eliminate negative effects of insufficient information utilization and residue noisy weights. It also proves that NIS is better than ATS at filtering noise because the MLP provides deeper structure to handle the complexity of noise information.

4.4 Effectiveness of NIS Module

The NIS module we propose has independent parameters, thus can be adapted to various kinds of neural relation extraction methods. To further demonstrate its effectiveness on different methods, we (1) replace MIL module of PCNN+MIL with our NIS module (denoted as **PCNN+NIS**). Note that this method is a **sentence-level** extraction, different from all other settings; (2) replace APCNN and APCNN+NIS's non-linear attention with bilinear form attention [11] (denoted as **PCNN+ATT** and **PCNN+ATT+NIS** respectively). We report the aggregated precision/recall curves of all the NIS methods in Fig. 4. From Fig. 4 we can see that:

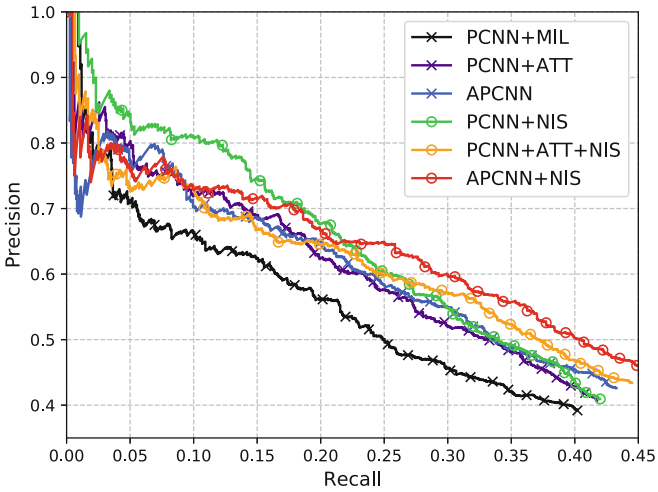

Fig. 4. Aggregate precision/recall curves of PCNN+MIL, PCNN+ATT, APCNN, PCNN+NIS, PCNN+ATT+NIS, APCNN+NIS.

1. Models with NIS module outperform all the corresponding baseline methods, proving its effectiveness and robustness on different structures.
2. PCNN+NIS model has better performance than PCNN+MIL. It indicates the role of NIS module in filtering noise. Lin et al. [11] have proved that sentence-level PCNN has worse performance than PCNN+MIL because it ignores the effect of noise. But our sentence-level PCNN+NIS method defeats PCNN+MIL, which means NIS module can filter out most noise sentences and improve sentence-level performance.
3. PCNN+NIS model also outperforms attention-based models. This is actually a comparison between hard selection and soft selection strategy. The result demonstrates that NIS's hard selection strategy can effectively reduce the negative effects of residue weights brought by soft attention strategy.

4.5 Analysis of ATS Threshold

Although not so powerful as NIS method, ATS method still brings improvement to APCNN model. However, ATS method needs a fine-tuned 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 ATS with different thresholds. For clarify, we use a histogram to approximate precision/recall curves of different thresholds, shown in Fig. 5.

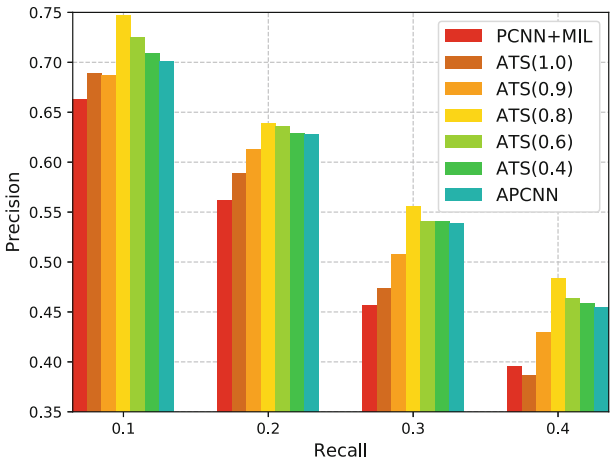

Fig. 5. Aggregate precision/recall histogram of ATS with different thresholds. ATS(α) means the threshold is $\alpha \times \max(\textit{attention weights})$. APCNN is equivalent to ATS(0). ATS(1.0) means only select the sentence with maximum attention weight as bag representation, similar to PCNN+MIL. We use histogram for clarify because some of the curves are too close.

In our experiments, the best model is ATS(0.8). With higher thresholds (ATS(1.0) and ATS(0.9)), the precisions decline significantly because less informative sentences are utilized. ATS(1.0) selects only the sentence with maximum attention weight to train. Similar to MIL select strategy, ATS(1.0) also has similar performance to PCNN+MIL. With lower thresholds (ATS(0.6) and ATS(0.4)), the performance decreases slightly, close to APCNN model (equivalent to ATS(0)). The reason is that when threshold is lower, more invalid sentences are involved in training, which means the noise effects cannot be fully eliminated. The change of performance with the threshold perfectly shows the impact of information loss and noise residue on relation extraction. It also proves the superiority of NIS because it provides deeper structure to capture the complex information of noise and doesn't require fine-tuned threshold.

4.6 Case Study

Table 3 shows an example of selection result and attention weights of a bag. The bag contains three instances in which the 1st instance is invalid. The remaining instances are informative because they all contain significantly keywords that express the corresponding relation */people/person/place_lived*. APCNN assigns bigger weight to the 1st sentence (invalid) than the 2nd sentence (valid). Therefore, the big noise residue will substantially hurt performance. Our NIS module correctly selects last 2 sentences as valid ones. The selection results shows NIS's ability of filtering noise sentences.

With the help of NIS, attention mechanism only assigns weights to selected sentences. APCNN+NIS assigns a very high weight to the 3rd sentence because the appearance of *lived* strongly indicates the *place_lived* relation. The 2nd

Table 3. An example of selection result and attention weight. The bold strings are head/tail entities, and the red strings are keywords to predict the relation. The relation */place_lived* corresponds the */people/person/place_lived* in Freebase.

Triplet	Instances	APCNN+NIS		APCNN
		Select.	Att.	
(Jane Jacobs, Toronto, /place_lived)	1. Alice Sparberg Alexiou, the author of the biography "**Jane Jacobs**: Urban Visionary" ... in a panel discussion based on the work of Ms. Jacobs, the urban planner who died in April in **Toronto**	0	-	0.3503
	2. Dovercourt has a penchant for arriving at rock clubs and bars with books by the famed urban critic **Jane Jacobs**, who has made **Toronto** her home for nearly 40 years	1	0.2890	0.0875
	3. **Jane Jacobs**, the activist who took him on, now lives in **Toronto**	1	0.7110	0.5623

sentence has *home*, but the semantic is not strong enough. The attention weights demonstrates that our attention module is able to selectively focus on more relevant sentences.

5 Conclusion

Distant supervision for relation extraction is an efficient method to find relational sentences in very large corpus without manual annotation. Existing methods suffers from information loss and noise residue problem. We proposed a method named NIS to alleviate these two negative effects simultaneously. We use a sentence encoder to transform input texts to vector representation, an NIS module to select multiple valid sentences and an attention module to assign weights to selected sentences. We conduct experiments on a widely used dataset and experimental results validate the effectiveness of our method.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, pp. 2787–2795 (2013)
3. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in natural Language Processing, pp. 724–731. Association for Computational Linguistics (2005)
4. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. intell. **89**(1–2), 31–71 (1997)
5. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu, X.: Reinforcement learning for relation classification from noisy data (2018)
6. GuoDong, Z., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 427–434. Association for Computational Linguistics (2005)
7. Han, X., Liu, Z., Sun, M.: Neural knowledge acquisition via mutual attention between knowledge graph and text (2018)
8. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
9. Ji, G., Liu, K., He, S., Zhao, J., et al.: Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: AAAI, pp. 3060–3066 (2017)
10. 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. 799–800. ACM (2007)
11. 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. 2124–2133 (2016)

12. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. 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: Volume 2, vol. 2, pp. 1003–1011. Association for Computational Linguistics (2009)
15. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 2204–2212 (2014)
16. 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. 39–48 (2015)
17. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
18. Santos, C.N.d., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:1504.06580 (2015)
19. 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. 1201–1211. Association for Computational Linguistics (2012)
20. 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. 1784–1789 (2017)
21. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph and text jointly embedding. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1591–1601 (2014)
22. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
23. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
24. 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. 1753–1762 (2015)
25. 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. 2335–2344 (2014)
26. 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. 207–212 (2016)