# ATFE: A Two-dimensional Feature Encoding-based Sentence-level Attention Model for Distant Supervised Relation Extraction

Shiyang Li[1], Qianqian Ren*[1], Zechao Liu[2]
[1]Department of Computer Science and Technology, Heilongjiang University, Harbin 150080, China
[2]School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
*Email: renqianqian@hlju.edu.cn

***Abstract*—Distant supervised relation extraction has recently attracted researchers attention in the knowledge graph. However, the current feature encoding model of sentences can not fully represent the features in sentences, which poses a challenge. To solve this problem, we propose a two-dimensional feature encoding-based sentence-level attention model for relation extraction. In this model, we first employ bidirectional long short-term memory networks(BiLSTM) to capture the temporal dependency of the words in the sentence. Then we employ multi-dilated convolution to obtain the higher-level semantic units hidden in the sentence. Afterwards, we combine the above two-dimensional features to embed the encoding of sentences, which is expected to enhance the model's ability to express sentence features. Finally we build sentence-level attention to complete the relation extraction task. Compared with other excellent methods, the proposed approach provides a significant performance improvement.**

***Index Terms*—Deep learning, Attention, Distant supervision, Relation extraction**

## I. Introduction

In recent years, the growing commercial interest in artificial intelligence related fields has spurred the development of knowledge graphs. Many knowledge graphs related techniques have been proposed and applied. Among them, the knowledge base widely used for natural language processing(e.g. Freebase, Wikidata) related issues such as search engines and question answering systems. Some existing large-scale knowledge bases are composed of a large number of triples (e.g. <Jack_Ma, founder, Alibaba>), which implement information storage in a structured manner etc[2, 4]. These triples concisely reflect the two objective entities and the relation between them. However, these existing knowledge bases are not sufficient to cover all the facts in the real world. We need to continuously expand our knowledge base to increase its integrity. Many researches focus on the study of relation extraction, which can automatically obtain unknown relations in knowledge bases from plain text.

Relation extraction is the process of obtaining two entities and the relation between them from an unknown text. This is meaningful during the expansion of the knowledge base. The initial extraction is performed in a supervised manner, requiring people to manually label the training data, which is a time-consuming and very expensive task[3, 5]. Until 2009, the concept of distant supervision is proposed. This method automatically generates relational training data by aligning entities in the text with the known triples in the knowledge base. Distant supervision can effectively avoid the tedious and time-consuming manual annotation process [13]. However, this method can also mislabel and generate a lot of noise during the experiment. To solve the problems caused by distant supervision, multi-instance learning is proposed and widely used [6]. Lin et al propose a sentence-level attention method to make full use of the relation information in all sentences [10]. In this method, the weighted sum of all sentences in the package is used to express the relation between entity pairs. Guo et al add entity recognition on the above method to further obtain entity background knowledge to improve relation extraction performance [9].

In recent years, various deep learning architectures have been proposed to replace traditional natural language to encode sentence features. Convolutional neural networks(CNN) are proposed to code sentence semantic [1, 11, 14, 18]. On this basis, deep convolutional neural networks(DCNN) and residual networks are proposed to enhance the model's ability to express sentence features [7]. Recurrent neural networks (RNN) and long short-term memory networks (LSTM) are generally adopted to model the temporal dependency of words in the sentence and achieved certain results [19, 20]. These techniques along with other tricks are usually combined to improve the effectiveness of models.

In this paper, we propose a two-dimensional feature encoding -based sentence-level attention model(ATFE) for distant supervised relation extraction. We first build a two-dimensional feature encoder to embed the encoding of sentences, which is expected to combine the temporal dependency of the words and the higher-level semantic units in sentences to enhance the model's ability to express sentence features. Afterwards, we use the result of two-dimensional feature encoding to built sentence-level attention to complete the relation extraction task. Experiments on real data sets show that compared with baseline models, our model can perform sentence feature encoding more precisely, moreover further improve the performance of relation extraction related algorithms.

The contributions of the paper are as follows:

DOI reference number: 10.18293/SEKE2021-107

- We propose ATFE, a new sentence feature encoding model to obtain the two-dimensional feature representation of the sentence.
- We incorporate the sentence-level attention mechanism with our model to calculate the extracted relation probability.
- We implement experiments on real data sets to validate the performance of our proposed model. The experimental results show that our model can encode sentences feature better. It can be more effectively used in the distant supervision relation extraction model to improve task accuracy.

## II. Related Work

Most existing relation extraction methods can be roughly divided into two categories: one is based on word sequence, and the other is based on the dependency tree. The method based on the dependency tree is to model the dependency tree of the sentence instance as the input data, which will not be described in detail here. The method based on the word sequence is to use the word sequence to build a model. The model is used to encode sentence features to obtain the semantic representation of sentences. Since deep learning was proposed, neural network models have brought tremendous changes to the research on feature extraction. Nowadays, using neural networks to automatically learn features in sentences for relation extraction tasks has been widely studied. Some classic models for feature encoding of sentences have been proposed, such as the piecewise convolutional neural network(PCNN) model [17]. The researchers applied piecewise max-pooling to the model to make it better encode sentence structure information.

Although the method effectively improves the effect of the relation extraction task. However, due to the influence of the CNN network structure, the small size convolution kernel cannot capture the temporal dependency of the words in the sentence and perform more accurate sentence feature coding. In subsequent works, the researchers introduced RNN to the task of relation extraction and obtained long-term dependency by capturing the time sequence information of the words in the sentence. Among them, long short-term memory networks(LSTM) is an excellent RNN model that is composed of computing units.

These works actively promote the improvement of the accuracy of the relation extraction model, and achieve great success. However, they all ignore the higher-level semantic unit information in the sentence. The semantic unit is shown in figure 1:

Different from word-level information, it is higher-level information hidden in phrases or sentences. They are combined to express the semantics of sentences. Therefore, we believe that accurately capturing the representation of semantic units in a sentence is the key to enhance the model's ability to express sentence features.

Encoding based on semantic information in sentences has attracted many researchers in natural language processing(nlp) recently. The diversity of models based on deep learning enables them to try different methods to improve the effectiveness of various nlp tasks, such as semantic segmentation, text emotion analysis and machine translation. To the best of our knowledge, this is the first effort to adopt the two-dimensional feature encoding model including BiLSTM and multi-dilated convolution in distant supervised relation extraction.

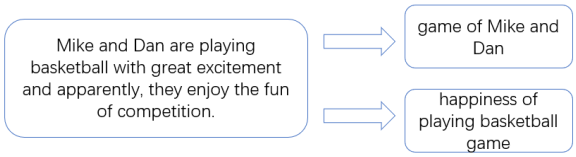


Fig. 1. The text "Mike and Dan are playing basketball with great excitement and apparently they enjoy the fun of competition" contains two important points, "game of Mike and Dan" and "happiness of playing basketball game". These information are called semantic units.

## III. Methodology

For $n$ sentences $\{x_1, x_2, x_3, \ldots, x_n\}$, each sentence consists of $m$ words, denoted as $x_i = \{a_1, a_2, a_3, \ldots, a_m\}$, which contains two entities (head_entity and tail_entity). The purpose of our model is to calculate the probability of each relation $r$. For the entire relation extraction model, we divide it into two parts:

- **Two-dimensional feature encoder:** Given $n$ sentences $\{x_1, x_2, x_3, \ldots, x_n\}$, use our proposed model to perform feature encoding on sentence vectors.
- **Sentence-level attention:** We make full use of the multi-instance learning idea, extract sentence information of the target entity pair through all contained relation to predicting the relation probability of the target entity pair.

### A. Two-dimensional Feature Encoder

The structure of the two-dimensional feature encoder is shown in figure 2. The model is composed of vector representation, network layer, piecewise max pooling. The following describes how the model is implemented:

**Vector representation.** Since the neural networks cannot directly recognize the words in the sentence, we should use the encoding tool to transform the words into low- dimensional vectors. Considering that the length of each sentence is different, and important information may be contained anywhere in the sentence, we pad zeros around sentences to make them equal in length, which is in order to facilitate the model encode the sentence vector. And we add the position information of the given entity pair.

**Word embedding and position embedding.** Word embedding is distributed representations of words. It can map words in texts to a low-dimensional vector that can capture syntactic and semantic meanings. Position embedding is an important part of the model. It is defined as the combination of the relative distances from each word in a sentence to two given entities, as shown in figure 3. The final embedding method is shown in Vector Representation in figure 2. If the specified

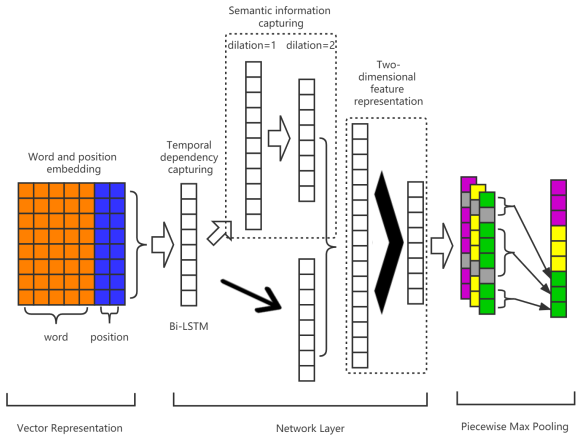


Fig. 2. Two-dimensional feature representation sentence encoder.

word embedding is $dw$ and the position embedding is $dp$ , given the vector sequence $x_i$={ $a_1$, $a_2$, $a_3$,..., $a_m$ }, the length of $a_i$, denoted as $da_i$=$dw$ + 2* $dp$. In our model, we set $dw$=5 and $dp$=1, then $da_i$=5 + 2* 1=7.

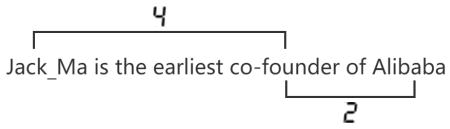


Fig. 3. The distance from "co-founder" to the head entity "Jack_Ma" in the sentence is 4, and the distance to the tail entity "Alibaba" is 2.

**Network layer.** The matrix containing sentence word embedding and position embedding is input into the network for feature encoding. First, BiLSTM is used to capture the timing information of the words in the sentence to obtain long-term dependencies. The essence of BiLSTM is a two-way LSTM structure.

**Temporal dependency capturing.** LSTM is a kind of recurrent neural network and has been widely used. By adding control gates (including input gates, output gates and forget gates) to the network, the network can eliminate unnecessary words in sentences and retain important words. However, due to the structural characteristics of LSTM, it is impossible to encode the information from back to front. It makes the network unable to carry out more fine-grained encoding. To solve this problem, BiLSTM first performs a LSTM from front to back, and then performs a LSTM from back to front, and next combines two results to obtain the final feature encoding. In the model, we input sentences into the network in word order and use the characteristics of the words that enter the network first, which is calculated together with the next word that enters the network. This process is repeated until the last word is processed. The process is shown in figure 4.

**Semantic information capturing.** Based on the representation generated by BiLSTM, we introduce multi-dilated convolution to capture the semantic unit representation in

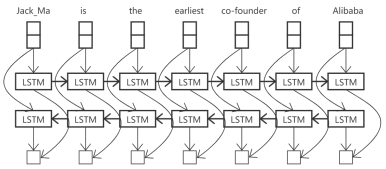


Fig. 4. First we input "Jack_Ma", "is", "the", "earliest", "co_founder", "of", "Alibaba" in turn to get three hidden vectors and then input "Alibaba", "of", "co_founder", "earliest", "the", "is", "Jack_Ma" in turn to get three hidden vectors. Finally, the hidden vectors are spliced to get the encoding of the sentence.

the sentence or phrase through the temporal dependency in the sentence. Dilated convolution is actually a special CNN design. By adding "holes" to the convolution, the receptive field can be expanded exponentially without adding additional parameters. To prevent the dilated segments of the convolutional kernel from causing the missing of vital local correlation, we design the network as a two-layer convolution with different dilation rates, as shown in figure 5.

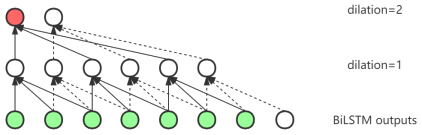


Fig. 5. Multi-dilated convolution structure.

In the proposed model, the scale of the convolution kernel is 3, and the dilation of the two-layer dilated convolution is 1 and 2, respectively. In this way, each convolution kernel in the highest layer can observe 7 inputs from BiLSTM from left to right. While expanding the receptive field, it also prevents the highest level from processing sentence information that is too long, and reduces the noise in the sentence caused by the influence of some irrelevant information. At the same time, to remain the sentence vector sequence dimension of the input and output of each network layer, we set the different padding size to make the same length of the convolution process. In this way, we can capture the semantic unit representation from phrase-level information with a smaller dilation rate and sentence-level information with a larger dilation rate.

**Two-dimensional feature representation.** Finally, we will perform one-dimensional splicing of the results obtained by two-layer convolution to obtain feature vectors containing sentence timing information and semantic units. The feature vector is passed through a convolutional layer with a convolution kernel 1 to output the two-dimensional features of the final sentence.

The convolution process with convolution kernel 1 is equivalent to the calculation process of full connection. It can

When setting the dilation rate of different layers of dilation convolution, you should note that the dilation rate of each layer has no other common factors other than 1. Otherwise, the grid effect will be generated by adding "holes" in the convolution.

increase the nonlinearity of the network by changing the vector dimension, thus the network can express more complex features.

**Piecewise max pooling.** According to the given entity pair, the sentence is divided into three parts ($p_{i1}$, $p_{i2}$, $p_{i3}$), and then the maximum pooling operation is performed on each part (the piecewise max pooling part is in figure 3, the gray block represents the position of the entity pair in the sentence). The process can be defined as:

$$[x]_{ij} = [max(p_{i1}), max(p_{i2}), max(p_{i3})] \tag{1}$$

### B. Sentence-level Attention

To make full use of the relation information expressed in each sentence in multi-instance learning, we use sentence-level attention to complete the relation extraction task[8]. Its structure is shown in the figure 6:

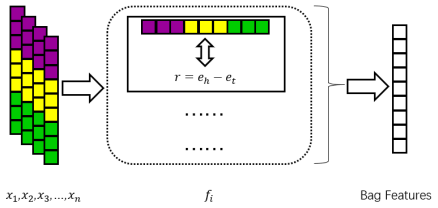


Fig. 6. Sentence-level Attention.

Given a set of entity pairs $< e_h, e_t >$, the set *S* consists of all sentences containing this entity pair, denoted as $S = \{x_1, x_2, x_3, \ldots, x_n\}$. We perform weighted summation on the feature encodes of all sentences in the set to get the feature encode of set S:

$$s = \sum_i w_i x_i \tag{2}$$

To measure the contribution of the sentence containing the entity pairs in predicting whether the entity pair has a relation *r*, we first match the vector representation of the relation *r* with $x_i$, we calculate their similarity:

$$f_i = ||x_i - r||_2 \tag{3}$$

Here, due to the nature of word embedding, we use entity vector difference to represent the relation characteristics between entity pairs[12, 16]:

$$r = e_h - e_t \tag{4}$$

Then $f_i$ represents the score using the sentence to predict whether the entity pair has a relation $r$, and we calculate $w_i$ according to the following formula:

$$w_i = \frac{exp(f_i)}{\Sigma_k exp(f_k)} \tag{5}$$

In this way, we get a vector representation of the set *s* of sentences containing a given entity pair. We use a linear function to represent the score of each possible relation *r*, it is the final output of the neural network.

$$O = RS + d \tag{6}$$

Where *R* is the representation matrix of the relation, and *d* is the deviation vector. Then we get the scores for *i* possible relations. We calculate the probability of each relation *r* accordingly:

$$p(r|s, \theta) = \frac{exp(o_r)}{\Sigma_{k=1}^{n_r} exp(o_k)} \tag{7}$$

Where $n_r$ is the total number of distant supervised alignments.

## IV. Training and Experiment

This section introduces the optimization process, data set, experimental environment, and parameter setting of our model respectively. Finally, the performance of the model is compared with some baseline methods.

### A. Model Optimization and Training

For the entire relation extraction model, we use cross entropy to define the objective function as:

$$J(\theta) = \sum_{i=1}^{s} \log_p(r_i|s_i, \theta) \tag{8}$$

We use Adam to optimize the objective function. Adam has a great advantage in non-convex optimization problems and is an extension of Stochastic Gradient Descent (SGD). It uses the first-order moment estimation and the second-order moment estimation of the gradient to dynamically adjust the learning rate of each parameter. It can maintain high-efficiency calculations while occupying a small amount of computer memory. When training the model, we introduce dropout to prevent the model from overfitting to get the best model training results[15].

### B. Datasets and Preprocessing

Here we use Nyt10 data set for the relation extraction task. The word2vec tool is used for sentence vectorization.

**Nyt10.** The nyt10 data set is maintained by Tsinghua University and is used for relation extraction experiments. It is generated by adjusting the New York Times Corpus and Freebase. A total of 53 kinds of relations are included, among which there is a relation NA, which indicates that the head entity and the tail entity in the sentence have no relation. We use the nyt10 data set for model training and verification. The training set includes 466876 sentence examples. The test set consists of 172448 sentence examples.

**Parameter settings.** In order to determine the optimal settings of each parameter of the model, based on the data provided by [10, 11, 18], we made more attempts on the sentence embedding size, window size, learning rate, and other parameters. In the process of multi-dilated convolution, try the window size and the number of convolution layers between {2,

TABLE I
EXPERIMENTAL PARAMETER SETTINGS.

| | | |
|---|---|---|
| Window size | | 3<br>1 |
| Hidden size | Bilstm<br>Dilated convolution<br>CNN | 60<br>230, 60<br>230 |
| Word embedding | | 50 |
| Position embedding | | 5 |
| Batch size | | 160 |
| Dropout probability | | 0.5 |

3}, the feature map size of different networks is selected in {30, 60, 120, 200, 230, 256}, and the batch size is selected in {100, 160, 180, 220}. Finally, the parameters used in our experimental model are listed in Table 1.

### C. Evaluation Metric

The purpose of our experiment is to prove that a more accurate sentence feature coding model can improve the performance of the relation extraction method based on remote supervision. Finally, the model is evaluated through the Precision -Recall curve and the AUC and F1 values of the model on the test set verification set.

We compare our model with several previous works, figure 7 clearly shows the Precision-Recall curve of all methods.

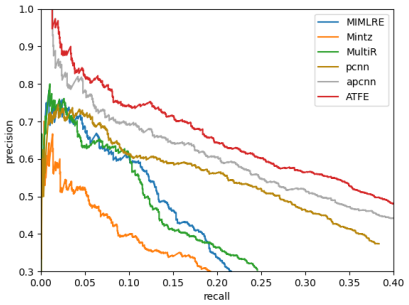


Fig. 7. Precision-Recall curve of each method in the comparison experiment.

As can be seen from figure 7, the neural network-based models (ATFE, APCNN, PCNN) are significantly better than other baseline models and have higher precision and recall. At the same time, among these three models, the performance of our model (ATFE) is significantly better than the other two neural network models. It is clear that improving the neural network's ability to express the characteristics of sentences can have a positive effect on the task of relation extraction. Furthermore, In the entire P-R curve range, ATFE achieves higher precision than all baseline models.

In addition, in order to verify the effectiveness of two-dimensional feature representation for the final performance improvement of relationship extraction. We conducted step-by-step experiments on our proposed model. When the input data and the sentence-level attention method are unchanged, the ATFE model is compared with the ABiLSTM model and the dilated convolution model (ADNN) to verify its effectiveness. The result is shown in figure 8. It can be seen from the figure that due to the limitations of the network structure, the final results of the ABiLSTM and ADNN models are not satisfactory, and the two-dimensional feature representation model (ATFE) that combines sentence time information and higher-level semantic representations produce a satisfactory result.

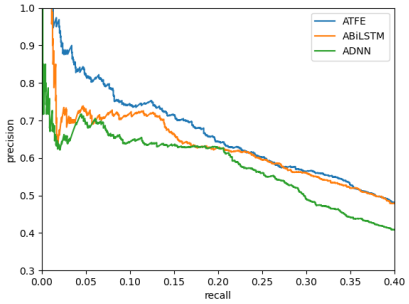


Fig. 8. Step-by-step experiment results.

Furthermore, the AUC and F1 values of our proposed ATFE model in the validation set and test set are further compared with the APCNN model. The result is shown in figure 9:

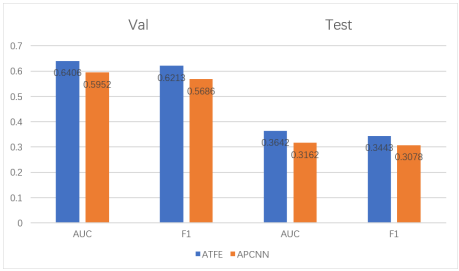


Fig. 9. Performance comparison of ATFE and APCNN on the validation set and test set, using different colors to distinguish the two methods.

Through the above evaluation, we can find that the performance of our proposed ATFE model on the test set and verification set is significantly higher than the APCNN model. In the validation set, the highest AUC value of the ATFE model can reach 0.6404, and the highest F1 value can reach 0.6213. The APCNN model only has an AUC value of 0.5952 and an F1 value of 0.5686. Similarly in the test set, the final AUC and F1 values of the ATFE model are 0.3642 and 0.3443, respectively. The AUC and F1 values of APCNN are 0.3162 and 0.3078, which are lower than the ATFE model.

The above several experiments show that our model has better performance than other baseline models. It shows that capturing higher-level semantic representations through the temporal dependency of words in sentences, obtain the two-dimensional feature encoding of the sentence can improve the model's ability to encode sentence features, and accordingly enhance the performance of relation extraction tasks based on remote supervision. It played a positive role in promoting the development of this research work.

Experiment summary and questions. It can be seen from figure 7 that when the recall is low (less than 0.05), the precision of our model has a rapid decline. It is because the heldout evaluation suffers from false negative in Freebase. Although this problem can be eliminated by manual evaluation, it will inevitably lead to huge labor costs as the size of the data set increases. So whether or this false negative label can be eliminated or corrected through the autonomous learning of the model will become a direction for future researchers to improve the performance of the model.

## V. Conclusion

In this paper, we propose an ATFE model to solve the problem that the feature encoding model of sentences in the previous distant supervised relation extraction method cannot adequately represent the features in sentences. We first use BiLSTM to capture the time dependence of words in sentences by bidirectional encoding. On this basis, we design multi-dilated convolution to further acquire the higher-level semantic units hidden in sentences. Finally, the feature encoding ability of the model is maximized by fusing the captured two-dimensional features, We build sentence-level attention to complete the relation extraction task. Compared with similar methods, our proposed method has significant performance improvement.


## Acknowledgement

This work was supported by the Basic Research Program (no. JCKY2019604B002).


## References


[1] Adel, H., Schütze, H., 2017. Global normalization of convolutional neural networks for joint entity and relation classification. arXiv preprint arXiv:1707.07719 .

[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data, in: The semantic web. Springer, pp. 722–735.

[3] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research 3, 1137–1155.

[4] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 1247–1250.

[5] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2011. Natural language processing (almost) from scratch. Journal of machine learning research 12, 2493–2537.

[6] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, Luke Weld, D.S., 2011. Knowledge-based weak supervision for information extraction of overlapping relations, in: Proceedings of the 49th annual meeting of ACL: human language technologies, pp. 541–550.

[7] Huang, Y.Y., Wang, W.Y., 2017. Deep residual learning for weakly-supervised relation extraction. arXiv preprint arXiv:1707.08866 .

[8] Ji, G., He, S., Xu, L., Liu, K., Zhao, J., 2015. Knowledge graph embedding via dynamic mapping matrix, in: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pp. 687–696.

[9] Ji, G., Liu, K., He, S., Zhao, J., et al., 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions., in: AAAI.

[10] Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M., 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), pp. 2124–2133.

[11] Liu, C., Sun, W., Chao, W., Che, W., 2013. Convolution neural network for relation extraction, in: International Conference on Advanced Data Mining and Applications, Springer. pp. 231–242.

[12] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26, 3111–3119.

[13] Mintz, M., Bills, S., Snow, R., Jurafsky, D., 2009. Distant supervision for relation extraction without labeled data, in: ACL, pp. 1003–1011.

[14] Santos, C.N.d., Xiang, B., Zhou, B., 2015. Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:1504.06580 .

[15] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

[16] Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014. Knowledge graph and text jointly embedding, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1591–1601.

[17] Zeng, D., Liu, K., Chen, Y., Zhao, J., 2015. 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.

[18] Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., 2014. Relation classification via convolutional deep neural network, in: COLING 2014, pp. 2335–2344.

[19] Zhang, D., Wang, D., 2015. Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006 .

[20] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B., 2016. 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), pp. 207–212.