Multi-level Attention-Based Neural Networks for Distant Supervised Relation Extraction

Linyi Yang, Tin Lok James Ng, Catherine Mooney, Ruihai Dong

Insight Centre for Data Analytics, University College Dublin, Ireland
{linyi.yang,james.ng,ruihai.dong}insight-centre.org
{catherine.mooney}ucd.ie

Abstract. We propose a multi-level attention-based neural network for relation extraction based on the work of Lin et al. to alleviate the problem of wrong labelling in distant supervision. In this paper, we first adopt gated recurrent units to represent the semantic information. Then, we introduce a customized multi-level attention mechanism, which is expected to reduce the weights of noisy words and sentences. Experimental results on a real-world dataset show that our model achieves significant improvement on relation extraction tasks compared to both traditional feature-based models and existing neural network-based methods.

1 Introduction

Relation Extraction (RE) aims to identify relations between entities from natural language text. It plays a key role in many natural language processing (NLP) tasks, including question answering, web search, and knowledge-based construction. The existing relation extraction approaches can be divided into supervised learning methods [6], semi-supervised learning methods, [19], and unsupervised learning methods, [3]. In the supervised approaches, sentences in a corpus are first hand-labelled by domain experts to produce labelled examples of specific relations. The identified examples are then used to induce rules for identifying additional instances of relations. In other words, relation extraction is considered to be a multi-class classification problem. While supervised methods can achieve high precision [6], labelling training data requires enormous amount of effort from domain experts which is very time-consuming.

To address this problem, Mintz et al. [14] applied distant supervision to automatically generate training data via aligning the New York Times (NYT) news text with the large-scale knowledge base Freebase [4], which contains more than 7300 relationships and more than 900 million entities. They assume that if two entities have a relationship in a known knowledge base, then all sentences that contain these entity pairs will express this relationship in some way. For example, (Ireland, *capital*, Dublin) is a relational triple fact stored in Freebase. All sentences with synonyms for both entities, Ireland and Dublin, are considered to be an expression of the fact that (Ireland, *capital*, Dublin) holds. All these sentences will be regarded as positive instances for relation extraction. Although

distant supervision is an effective strategy for automatically labelling training data and a sound solution to leverage the availability of big data on the web, it suffers from the wrong labelling problem as the assumption is too strong. For instance, the sentence "Modern Ireland also has Dublin, whose budding metropolitan area is home to about 1.5 million people of Ireland 's population of close to 4 million." does not express the relation *capital* between two entities, but will still be regarded as a positive instance. The multi-instance learning introduced by [15],[10],[20] can alleviate the wrong labelling problem but is still far from satisfactory. These feature-based methods highly rely on NLP toolkits, such as part-of-speech annotations and syntactic parsing and the output of pre-existing NLP systems often leads to error propagation which will hurt the performance of proposed models. To solve this problem, many scholars [17], [22], [12] attempt to apply deep learning techniques instead of feature-based methods to relation extraction tasks, and our work will also focus on that.

Our proposal is an extension of [12]. In this paper, we propose a novel Bidirectional Gated Recurrent Unit (BiGRU) network integrated with a multi-level attention mechanism to automatically extract features without manual intervention. The contribution of our work can be summarized as follows. First, to further alleviate the wrong labelling problem, we build a multi-level attention mechanism in addition to sentence-level attention mechanism, which is expected to dynamically reduce the weights of both noisy words and sentences. Second, we evaluate our model on a widely used dataset developed by [15]. Finally, we show that our model achieves significant improvement compared to the state-of-art methods.

2 Related Work

RE generally relates to the extraction of relational facts, or world knowledge from the Web [21]. It is one of the most important subtasks in information extraction. Distant supervision is an alternative learning paradigm which assumes that if two entities have a relationship in a known knowledge base, then all sentences that contain these two entities will express this relationship [14], [10], [20]. As a form of weak supervision, distant supervision exploits relation repositories including Freebase [4], and DBpedia [1] to define a set of relation types and identify the text in a corpus which associate with the relations to produce the training data. Although distant supervision has emerged as a popular choice for training relation extractors and shows promising results in the task of relation extraction, it is inevitably accompanied by the wrong labelling problem. Hence, [15], [10], [20] applied multi-instance learning to alleviate the wrong labelling problem. These conventional methods inherit the knowledge discovered by the NLP toolkits for the pre-processing tasks. Their performance was intensely affected by the quality of supervised NLP toolkits as the output of pre-existing NLP systems often leads to error propagation or accumulation.

With the recent revival of interests in neural networks, many researchers [17], [23], [22], [12], [24] have utilized deep neural networks in relation classifi-

cation without handcrafted features. Although neural network based methods provide an effective way of reducing the number of handcrafted features, these approaches which build classifiers based on sentence-level annotated data, cannot be applied to large-scale knowledge bases due to the lack of training data. Therefore, a novel model dubbed Piecewise Convolutional Neural Networks with multi-instance learning was proposed by [22]. In the work of [22], multi-instance learning is integrated into a deep neural network model, which assumes that if two entities participate in a relation, at least one sentence that mentions these two entities might express that relation. Their integrated models are trained by selecting the most likely instance for each entity pair, and it is apparent that the method will lose a large amount of information in those neglected instances. To make full use of instances, [12] proposes a sentence-level attention-based convolutional neural network (CNN). Their model aims to make full use of sentences by allocating different weights to different instances in terms of their contributions in expressing the semantic relation information. Their proposed method achieves better results compared to [22]. Besides, [24] employs the attention mechanism with Bidirectional Long Short Term Memory Networks (BiLSTM) to capture the most important semantic information in a sentence for relation classification. Our proposal is an extension of [12] by combining with the work of [24].

3 Methods

In this section, we describe our novel neural network architecture to fulfill distant supervision for relation extraction before delving deeper into how each of the components within this model work in greater detail.

3.1 Overview

The distant supervised relation extraction problem is considered as a multi-instance problem. First, we present our BiGRU-based network that incorporates a multi-level attention mechanism. Figure 1 shows our neural network architecture which demonstrates the process that handles one instance of a bag. As shown in Figure 1, our model contains six components:

1. *Input layer*: Original sentences input to this model;
2. *Embedding layer*: Each word is mapped into a 50-dimension vector;
3. *BiGRU layer*: Using a neural network to get features automatically;
4. *Word attention*: Produce a weight vector on word level, and merge the word-level features into a sentence-level representation;
5. *Sentence attention*: Allocate different weights to different sentences in terms of their contribution in expressing the semantic relation information;
6. *Output layer*: Extract relation with the relation vector weighted by sentence-level attention.

We now introduce these components in more detail.

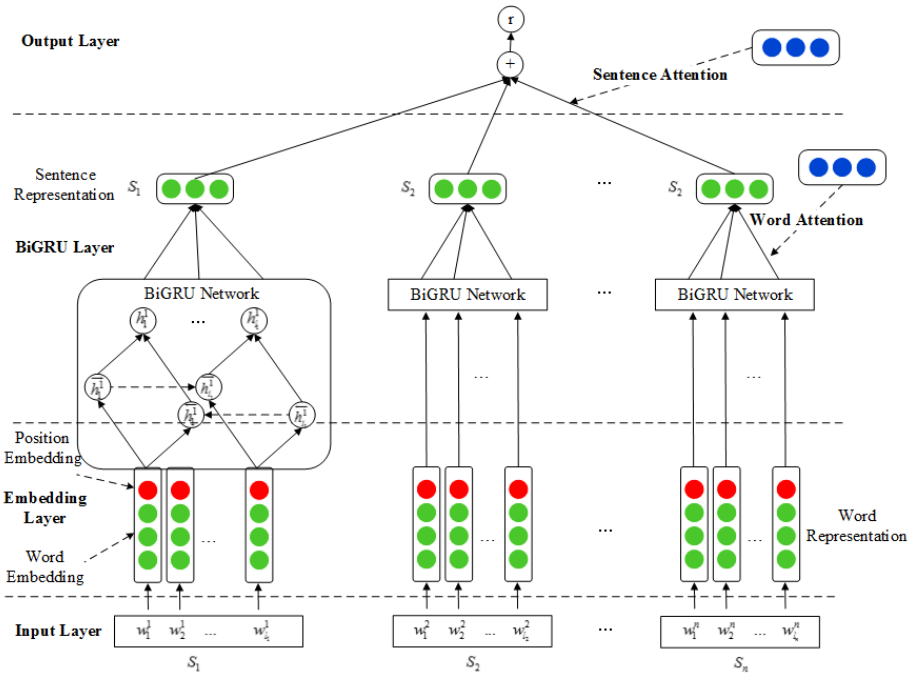

Fig. 1. The architecture of our Bidirectional GRU model with a multi-level attention mechanism

3.2 Vector Representation

Following [23], we transform each input word of a sentence into the concatenation of two kinds of representations:

1. *A word embedding*: Capture both semantic and syntactic informaion of the word.
2. *A position embedding*: Specify the position information of this word with respect to two target entities.

Word Embeddings. Word embeddings aim to transform words into distributed representations which capture both semantic and syntactic meanings of words. Each input word token is transformed into a low-dimensional real-valued vector by looking up pretrained word embeddings[13]. Specifically, given a sentence x consisting of m words $x = \{w_1, w_2, \ldots, w_{m-1}, w_m\}$, every word w_i is represented by a valued vector. Word representations are encoded by column vectors in an embedding matrix $V \in \mathbb{R}^{d^w \times |v|}$, where d^w is the dimension of the word vectors, and v is a fixed-sized vocabulary.

Position Embeddings. The main idea behind the use of word position embedding in a relation extraction task is to give some reference to the neural layer

of how close a word is to the target nouns, based on the assumption that closer words have more impact than distant words. The experimental result reported in [16] suggests that the use of word position embeddings is informative. Hence, in this paper, the position embedding of a word is further used as a vector concatenated with the word embedding.

3.3 Bidirectional Gated Recurrent Units in Neural Networks

In this paper we adopt a popular LSTM variant, Gated Recurrent Unit (GRU) network, which was first introduced by [7]. In addition, we employ a Bidirectional Gated Recurrent Unit (BiGRU) network based on the idea that the output at time t may not only depend on the past information, also the future information. Specifically, supposing the j-th hidden unit is computed in the neural network. First, it merges the cell state and hidden state then generates the reset gate q_j , which is computed by:

$$q_j = \sigma([W_r x]_j + [U_r h(t-1)]_j) \tag{1}$$

where σ represents the sigmoid function, $[.]_j$ is the j-th element of a vector, x and $h(t-1)$ are the input vector and previous hidden state respectively, and W_r and U_r are weight matrices. Second, it combines the forget and input gates into a single update gate. The update gate z_j is computed by:

$$z_j = \sigma([W_z x]_j + [U_z h(t-1)]_j) \tag{2}$$

Then, the actual activation of the proposed unit h_j is computed by:

$$h_j(t) = z_j h_j(t-1) + (1 - z_j)(\widetilde{h_j})(t) \tag{3}$$

where

$$\widetilde{h_j}(t) = \tanh([Wx]_j + [U(q \odot h(t-1))]_j) \tag{4}$$

Finally, we adopt an element-wise sum to add forward and backward states produced by Bi-GRU as the output of the j^{th} word.

$$h_j(t) = [\overrightarrow{(h_j(t)} \oplus \overleftarrow{(h_j(t)}] \tag{5}$$

3.4 Customized Attention Mechanism

Attention-based neural networks were first introduced by [2] for sequence to sequence learning in machine translation. Double attention mechanisms have also been previously developed in machine translation by [5]. In this section, we adopt a customized attention mechanism for relation extraction tasks. Our attention mechanism aims to use neural networks with word-level attention to obtain the representations of the sentences, then employ sentence-level attention to reduce the influence of false-negative sentences existing in each entity pair.

Word-level Attentions. Not all words contribute equally to the semantic relation information of an entity pair for relation extraction. For this reason, our word-level attention dynamically pay attention to the words in sentences that are more significant for semantic relation information. Suppose that given a sentence s containing n word embeddings, s = $\{w_1, w_2, \ldots, w_n\}$. The word embeddings are passed to GRU units repectively to get hidden states $\{h_1, h_2, \ldots, h_n\}$. The weights of the input columns at each time-step is called attention α. Inspired by [24], we can obtain the representation of sentences through the following equations:

$$s = \sum_{i=1}^{n} \alpha_i h_i \tag{6}$$

$$\alpha_i = \frac{exp(e(h_i))}{\sum_k exp(e(h_k))} \tag{7}$$

where $e(.)$ is a measure function that reflects the relevance between each word and relation of the entity pair in a sentence, and W is a weight matrix.

$$e(h_i) = W \tanh(h_i) \tag{8}$$

Sentence-level Attentions. Inspired by [12], the representation of S is computed as a weighted sum of these sentence vectors $\{s_1, s_2, \ldots, s_j\}$:

$$S = \sum_{i=1}^{j} \alpha_i s_i \tag{9}$$

where α_i is the weight of each sentence vector s_i.

The semantic information of set S would rely on the representations of all the sentences, each of which contains information that whether the entity pair expresses the relation. Like [12], we adopt a sentence-level attention to minimize the influence of the noisy sentences. It first measures the relevance between the instance embedding and the relation r. Then, it allocates more weight to true-positive instances and less weight to wrong labelling instances to reduce the influence of noisy sentences. α_i is calculated as:

$$\alpha_i = \frac{exp(\phi(s_i, r)))}{\sum_k exp(\phi(s_k, r))} \tag{10}$$

where $\phi(.)$ is a query-base function which scores how well the input sentence s_i and the relation r matches, $\phi(.)$ is defined as:

$$\phi(s_i, r) = s_i A r \tag{11}$$

where A denotes a weight matrix, and r is the representation of relation r.

3.5 Output

The output layer determines the relation label of an input set of sentences. In practice, we calculate the conditional probability through a softmax function as:

$$p(r|S) = \frac{exp(o_r)}{\sum_{k=1}^{n_r} exp(o_k)} \tag{12}$$

where n_r denotes the number of relations and o is the output of our model, which is defined as:

$$o = RS + b \tag{13}$$

where R is the representation matrix of relations and $b \in \mathbb{R}^{n_r}$ is a bias vector. Inspired by [22] and [12], we employ a loss function using cross-entropy at the entity-pair level. Then loss function is defined as follows:

$$L(\theta) = \sum_{i=1}^{N} \log p(r_i|S_i;\theta) \tag{14}$$

where N denotes the number of sentence sets for each entity pair and θ indicates all parameters of this model. For optimization problem, we adopt the Adaptive Moment Estimation (Adam) [18] update rule to learn parameters by minimizing the loss function.

Furthermore, in order to prevent overfitting, we apply dropout [11] on the output layer. The strategy of dropout aims to achieve better performance during the testing phase by randomly dropping out neural units during the training phase. Then, the output of our model is rewritten based on equation (13) as follows:

$$o = R(S \circ h) + b \tag{15}$$

where the vector h contains Bernoulli random variables with probability p.

4 Experiments

Our experiments aim to illustrate that our deep neural networks with sentence-level attention integrated with word-level attention can alleviate the wrong labelling problem benefiting from taking advantage of all informative words for relation extraction. In this section, we first specify our settings and describe the methods that we use for our evaluations. Next, we compare the performance of our model on a widely used dataset with several state-of-the-art methods, including traditional featured-based methods and neural network approaches. Finally, we show that our approach, BiGRU+2ATT, can consistently and effectively improve the previous best-performing model, PCNN+ATT.

4.1 Data

Entity mentions for both datasets are recognized using the Stanford open-source toolkit called entity tagger [9]. Pre-Trained Word Vectors are learned from New York Times Annotated Corpus (LDC Data LDC2008T19), which can be obtained from LDC [1].

We evaluate our model on a widely used dataset [2] which is generated by [15]. [15] presents a novel approach to extract relations from text without explicit training annotation. The Freebase relation instances are divided into two parts, one for training and one for testing. There are 53 possible relationships within this dataset including a special relation NA which represents that there is no relation between two entities. A total of 522,611 sentences, 281,270 entity pairs and 18,252 relational facts are stored in the training data, while the testing data includes 172,448 sentences, 96,678 entity pairs, and 1,950 relational facts.

4.2 Evaluation Metrics

Like [14], we adopt held-out evaluation to assess our model in distant supervised relation extraction. The held-out evaluation compares the relation facts between entity pairs discovered from the test set with those in knowledge base. However, the new relation instances that are not in knowledge base also could be discovered by the testing systems. We just assume that the testing systems have similar performance in relation facts inside and outside knowledge base so that we can provide an approximate measure of precision without manually evaluation. Here we report both the precision/recall curves and precision at n (P@N) [8] which considers only the topmost results returned by the model.

4.3 Experimental Setup

Word Embeddings The word embeddings used in this work are initialized by means of unsupervised pretraining. Similar to previous work [12], we use the Skip-gram neural network architecture available in the word2vec tool developed by [13]). For both datasets, we adopt the same NYT corpus to train word embeddings with word2vec. We first drop the words which appear less than 100 times in the corpus and keep the rest as our vocabulary set. Then, we generate the word embedding in 50 dimensions. Finally, we concatenate the words of an entity when it has multiple words.

Parameter Settings We keep the same value and size of parameter with the baseline [12] in order to highlight the increase of performance comes from method rather than the increase of parameter size. We illustrate hyperparameters used in the experiments in Table 1.

[1] `https://catalog.ldc.upenn.edu/LDC2008T19`
[2] `http://iesl.cs.umass.edu/riedel/ecml/`

Table 1. Parameter settings

Dataset	Freebase
Sentence embedding size	230
Word dimension	50
Position dimension	5
Batch size	50
Dropout probability	0.5

4.4 Performance on Riedel Dataset

To evaluate the proposed method, we compare our approach against three representative feature-based methods and the best-performing neural network approach to show that our model can improve the performance.

Following [12], Table 2 presents the P@N for the top 100, top 200, and top 300 extracted instances. For BiGRU, the multi-level attention method achieves the best performance in all test settings. The results show that our multi-level attention mechanism can consistently improve the performance compared to only using neural network with sentence-level attention.

Table 2. P@N for the top 100, top 200, and top 300 extracted relation instances. For each testing entity pair, randomly select one, two, and all sentences for relation extraction.

Test Settings	One			Two			All		
P@N (%)	100	200	300	100	200	300	100	200	300
BiGRU+ATT	76.0	72.0	66.0	77.0	73.5	67.7	76.0	74.0	69.7
BiGRU+2ATT	**82.0**	**76.5**	**72.0**	**85.0**	**80.5**	**71.3**	**84.0**	**82.0**	**77.0**

Baselines:

- Mintz [14]: An original model for distant supervised relation extraction.
- Hoffmann [10]: A distant supervision for a relation extraction model based on multi-instance learning which handles overlapping relations.
- Surdeanu [20]: Jointly models the latent assignment of labels to instances and dependencies between labels assigned to the same entity pair.
- PCNN+ATT [12]: The best-performing model on the Riedel dataset so far. Different from other three feature-based baselines, this is a neural approach.

Comparison with Feature-based and Neural Methods: From Figure 2, we can note that our model significantly and consistently outperforms the three feature-based methods over the entire range of recall. Moreover, compared with PCNN+ATT, BiGRU+2ATT obtains better performance over nearly the entire

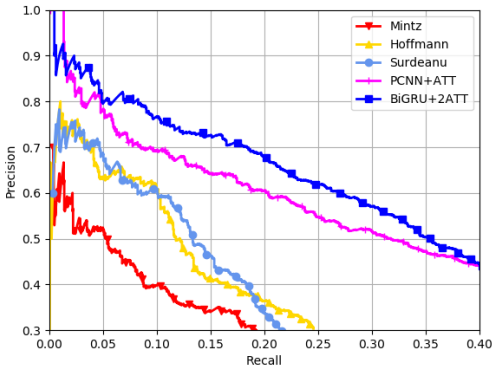

Fig. 2. Performance comparison of precision and recall curve for the proposed model with three feature-based and one neural baselines.

range of recall. It indicates that the proposed BiGRU integrated with word-level attention is beneficial. The reason is that the word-level attention would dynamically focus on the more informative words within sentences for the given relation.

4.5 Case Study

The selected three examples of our customized attention mechanism from the testing set are shown in the Table 3. For each sample, we display our attention weights of sentences, and we highlight the entity pairs with bold face.

From Table 3, we see that the first bag of sentences is composed of two sentences which are related to the triple (Robert L. Johnson, *founders*, Black Entertainment Television) which is stored in the knowledge base. The sentence with low attention weight does not express the relation Founders clearly. While the next sentence which has high attention weight demonstrates directly that Robert L. Johnson found Black Entertainment Television. The second example is related to the triple (Muhammad Yunus, *founders*, Grameen Bank). In this example, the relation fact also contains two sentences. The first sentence with low attention weight expresses the relation Founders implicitly, while the high one expresses directly what position Muhammad Yunus holds in the Grameen Bank. The last example is related to the triple (Ireland, *contains*, Cork). The result demonstrates that the two sentences have the equal attention weight expressing the relation that Cork is located in the Ireland.

5 Conclusion

In this paper, we develop BiGRU with multi-level attention, which automatically realizes learning features from data and makes full use of all informative words

Table 3. Three examples of sentence-level attention in NYT corpus

Relation	Founders
Low	Sports Sunday new act for a media mogul **Robert L. Johnson** sold **Black Entertainment Television** in 2000.
High	For, mrs. clinton, the strategy for reaching black voters at this early stage of, the campaign followed by phone calls to reinforce her candidacy from her, husband and supporters like **Robert L. Johnson**, who founded **Black Entertainment Television**.
Relation	Founders
Low	**Muhammad Yunus**, who won the Nobel peace prize last year, demonstrated with **Grameen Bank** the power of microfinancing.
High	On Sunday, though, there was a significant shift of the tectonic plates of Bangladeshi politics, as **Muhammad Yunus**, the founder of, a microfinance empire known as the **Grameen Bank** and the winner of, the 2006 Nobel peace prize ...
Relation	Contains
Equal	ConocoPhillips, the third-largest American oil company, began producing some diesel from, soybean oil last year at a plant in **Cork**, **Ireland**.
Equal	Zingerman's is unique in that it has a continental reach in the united states, said Peter Foynes, curator of the butter museum in **Cork**, **Ireland** ...

and sentences. We adopt word-level attention integrated with sentence-level attention to achieve better instance representation for the distant supervised relation extraction task. In practice, we evaluate our model on a widely used dataset to present the effect of multi-level attention mechanism. Experimental results show that our model outperforms not only state-of-the-art feature based methods but also neural network methods.

Acknowledgement. This research was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data (2007)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI. vol. 7, pp. 2670–2676 (2007)
4. Bollacker, K., Evans, C., Paritosh, P., Tim, S., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proc. 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250 (2008)

5. Calixto, I., Liu, Q., Campbell, N.: Doubly-attentive decoder for multi-modal neural machine translation. arXiv preprint arXiv:1702.01287 (2017)
6. Chan, Y.S., Roth, D.: Exploiting background knowledge for relation extraction. In: Proc. 23th ACL. pp. 152–160 (2010)
7. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
8. Craswell, N.: Precision at n. In: Encyclopedia of database systems, pp. 2127–2128 (2009)
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proc. 43rd ACL. pp. 363–370 (2005)
10. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proc. 49th ACL. pp. 541–550 (2011)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proc. 54th ACL (2016)
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: Proc. 47th ACL. pp. 1003–1011 (2009)
15. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. pp. 148–163 (2010)
16. Santos, C.N.d., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:1504.06580 (2015)
17. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proc. 2012 EMNLP. pp. 1201–1211 (2012)
18. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research 15(1) (2014)
19. Sun, A., Grishman, R., Sekine, S.: Semi-supervised relation extraction with large-scale word clustering. In: Proc. 49th ACL. pp. 521–529 (2011)
20. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proc. 2012 EMNLP. pp. 455–465 (2012)
21. Yates, A.: Extracting world knowledge from the web (2009)
22. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proc. 2015 EMNLP. pp. 1753–1762 (2015)
23. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proc. COLING. pp. 2335–2344 (2014)
24. 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: Proc. 54th ACL. vol. 2, pp. 207–212 (2016)