A Brief Survey of Relation Extraction Based on Distant Supervision

Yong Shi[2,3,4,5], Yang Xiao[1], and Lingfeng Niu[2,3,4](✉)

[1] School of Computer and Control Engineering,
University of Chinese Academy of Sciences, Beijing 100190, China
ynynny@sina.com
[2] School of Economics and Management,
University of Chinese Academy of Sciences, Beijing 100190, China
{yshi,niulf}@ucas.ac.cn
[3] Key Laboratory of Big Data Mining and Knowledge Management,
Chinese Academy of Sciences, Beijing 100190, China
[4] Research Center on Fictitious Economy & Data Science,
Chinese Academy of Sciences, Beijing 100190, China
[5] College of Information Science and Technology,
University of Nebraska at Omaha, Omaha, NE 68182, USA

Abstract. As a core task and important part of Information ExtractionEntity Relation Extraction can realize the identification of the semantic relation between entity pairs. And it plays an important role in semantic understanding of sentences and the construction of entity knowledge base. It has the potential of employing distant supervision method, end-to-end model and other deep learning model with the creation of large datasets. In this review, we compare the contributions and defect of the various models that have been used for the task, to help guide the path ahead.

Keywords: Relation extraction · Deep learning · Distant supervision

1 Introduction

The fundamental purpose of Information Extraction (IE), which is one of the most important task of natural language processing (NLP), is extracting structured information from primitive unstructured text. Subsequently, the structured information can be used easily by people or program. With the development of the Internet, people create and share many contents. As a result, the Internet is filled with huge amounts of data in the form of texts, and it is possible analyzing these primitive unstructured text by hand scarcely. Therefore, IE systems are extremely important. They can extract meaningful facts from texts to build Knowledge Base, which can be used for applications like search, machine reading comprehension and text generation. IE can be done in unsupervised domain, in the form of OpenIE [6]. And unsupervised approaches don't need to predefine

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11538, pp. 293–303, 2019.
https://doi.org/10.1007/978-3-030-22744-9_23

any ontology or relation classes and the IE system should extract facts from the texts along with the relation phrases. Conversely, the supervised information extraction and classification methods specifically refer to the classification of an entity pair to a set of predefined relations or filling the predefined slots, which is trained by using documents containing mentions of the entity pair or structured data.

As one of the most important part of IE, the Relation Extraction (RE) is mainly responsible for identifying entities from text and extracting semantic relationships between entities [3,18,21]. RE system is able to predict whether a given document contains a relation or not for the pair. Further more, relation extraction system should predict which relation class out of a given ontology does that document point to, given that it does contain a relation, which can be regarded as a multi-class classification problem with an extra NoRelation class.

Supervised methods for relation extraction require large amount of training data for learning an desired model. Using hand annotated datasets for relation extraction takes tremendous time and effort to construct the datasets. However, there are already many knowledge bases built out such as DBpedia [1], Freebase [2], YAGO and Google Knowledge Graph. A large number of Entity-Relation-entity triplet has existed in these knowledge base, which contains useful semantic information can be used to promote the performance of relation extraction system. It needs to label the triples to the corresponding sentences in the primitive text only. Therefore, Mintz [12] proposed an assumption: if a sentence contains a pair of entities involved in a relation, then the sentence describes the relation of this pair. For example, all the sentences in the corpus that contain China and Beijing would be presumed have mentioned the relation that Beijing is the capital of China, as shown as Fig. 1. Then all these sentences are annotated as the training corpus data of the relation of the capital, and the pairs of entities are labeled simultaneously. Then put all the sentences corresponding to a relationship into a package, which is called a bag and all sentences in a bag have the same label. This work has been done later and is called multi-example learning. Such large datasets allow for learning more complex deep learning models for relation extraction. However, there are some false-positive sentences in the positive bags, which brings noise. The noise present in datasets generated through distant supervision also require special ways of modeling the problem like Multi-Instance Learning as discussed in the subsequent sections.

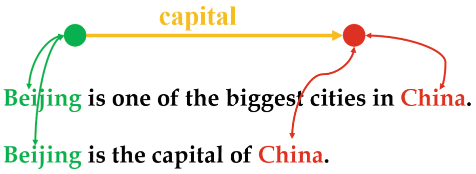

capital
Beijing is one of the biggest cities in China.
Beijing is the capital of China.

Fig. 1. A example of distant supervision.

In this review, we specifically focus on some different perspective of deep learning methods used for relation extraction.

2 Basic Concepts

In this section, we will introduce some basic concepts that are common across the models for relation extraction proposed recently.

2.1 Word Embeddings

Word Embeddings [10,11] are continuous distributional representations for the words in corpus, where each word is expressed as a continuous vector in a low dimensional latent space contrary to the high dimension of the one-hot vector. Word embeddings can capture the syntactic and semantic information about the word by predicting the context of words with unsupervised methods over large unlabeled corpus. Some work has been proposed to improve the word embeddings, such as Glove [13] and BERT [5]. After pre-training, all words are projected to an embedding matrix $E \in \mathbb{R}^{|V| \times d_w}$. Where d_w is the dimensionality of the embedding space and $|V|$ is the size of the vocabulary.

2.2 Position Embeddings

In natural language processing, the relative order of linguistic symbols has very important semantic information. Therefore, NLP system need to introduce this information into the model. In the relation extraction task, along with word embeddings, the input to the model also usually encodes the relative distance of each word from the entities in the sentence, In practice, the same continuous vectors as word embeddings are used instead of discrete form [20]. Position embeddings make neural network enable to keep track of the relative distance between words or entities in a sentence, which reserves the order information. The motivation is that words closer to the target entities probably imply more useful information reflecting the category of the relation between entities pair. The position embeddings comprise of the relative distance from current word to the entities. For example, in the sentence in Fig. 1 "Beijing is the capital of China." The relative distance between the word "capital" and head entity "Beijing" is 3 and tail entity "China" is -2. The distance are then encoded in a d_p dimensional embedding.

After obtaining word and position embeddings, the two embedding usually are Concatenated as the final representation of words and input the neural network. As the Fig. 2 shows.

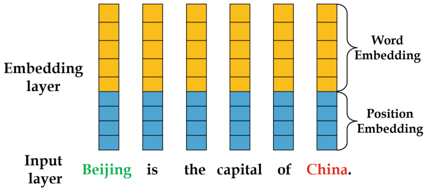


Fig. 2. A example of word representations with word and position embedding.

3 Datasets

In this section, We will introduce the commonly used data sets and evaluation metrics for relation extraction.

3.1 Supervised Dataset

The early works on relation extraction usually employed supervised training datasets. These datasets required intensive human annotation which meant that the data contained high quality tuples with little noise. But human annotation can be time-consuming, as a result of which these datasets were generally small. Both of the datasets mentioned below contain data samples in which the document sentence is already labeled with named entities of interest and the relation class expressed between the entity pair is to be predicted.

ACE 2005 Dataset: The Automatic Content Extraction dataset contains 599 documents related to news and emails. And the relations in the dataset are divided into 7 major types. 6 of the major relation types contain enough instances due to training and testing (average of 700 instances per relation type).

SemEval-2010 Task 8 Dataset: This dataset is a public dataset donated by Hendrickx et al. [7]. It contains 10717 samples which are divided as 8000 for training and 2717 for testing. It contains 9 relation types which are ordered relations. The directionality of the relations effectively doubles the number of relations, since an entity pair is believed to be correctly labeled only if the order is also correct. The final dataset thus has 19 relation classes (one for *Others* class).

3.2 Distant Supervision Datasets

To avoid the laborious task of manually building datasets for relation extraction, Mintz et al. [12] proposed the distant supervision to generate large number of relation extraction data automatically. They aligned sentences with known KBs, using the assumption that if a relation exists between an entity pair in the KB, then every sentence containing the mention of the entities pair would describe that relation. This distant supervision assumption is so strong that every sentence containing the entity pair mention may not express the relation between the pair and would bring some noise into the generated data. For example, For the tuple (*Beijing*, **Capital-of**, *China*) in the database and the sentence "Beijing is one of the biggest cities in China." At the distant supervision assumption, even this sentence do not describe the relation **Capital-of** between entity *Beijing* and entity *China*, it would also be labeled as the positive sample because of that it contains both the entities.

Riedel Dataset (also Called NYT): Riedel et al. [16] relaxed the distant supervision assumption by modeling the problem as a multi-instance learning problem, which can alleviate the problem mentioned above and reduce the noise. They released the Riedel dataset is the most popular dataset used in subsequent works building on distant supervision for relation extraction. This dataset was formed by aligning Freebase relations with the New York Times corpus (NYT). Entity mentions were found in the documents using the Stanford named entity tagger, and are further matched to the names of Freebase entities. There are 53 possible relation classes including a special relation NA which indicates there is no relation between the entity pair. The training data contains 522611 sentences, 281270 entity pairs and 18252 relational facts. The testing set contains 172448 sentences, 96678 entity pairs and 1950 relational facts.

GIDS: Jat et al. [8] created Google Distant Supervision (GIDS) dataset by extending the Google relation extraction corpus with additional instances for each entity pair. The dataset assures that the at-least-one assumption of multi-instance learning holds. This makes automatic evaluation more reliable and thus removes the need for manual verification. There are 5 possible relation classes between the entity pair. The training data contains 11297 sentences and 6498 entity pairs. The development set contains 1864 sentences and 1082 entity pairs. The testing set contains 5662 sentences and 3247 entity pairs.

4 Multi-instances Learning Models with Distant Supervision

Riedel et al. [16] regard the distant supervision relation extraction task as a multi-instances learning problem to relax the assumption, so that they could exploit the large training data created by distant supervision while being robust to the noise in the labels. Multi-instances learning is a form of weakly supervised learning problem where a label is given to a bag of instances, rather than a single instance. In multi-instances learning for relation extraction, Each entity pair in KB labels a bag of sentences. All the sentences in the bag contain the mention of the entity pair, but they do not contain the direct relation necessarily. Instead of giving a relation class label to every sentence, a label is instead given to each bag of the related entities. It assumes that if a relation exists between an entity pair, there is one document in the bag at least reflecting that relation of the given entity pair.

4.1 Piecewise Convolutional Neural Networks (PCNN)

The PCNN [19] model uses the multi-instance learning paradigm, with a neural network model to build a relation extractor using distant supervision data. The architecture of this model is similar to the model by Zeng et al. [20] proposed previously, with one important contribution of piecewise max-pooling over the sentence. The authors claim that the max-pooling layer can reduces the size of the latent feature remarkably. However, it is also losses important structure information between the entities in the sentence. This can be avoided by max-pooling in different segments of the sentence instead of the whole sentence. It is claimed that every sentence can naturally be divided into there segments based on the positions of the two entities in focus. By doing a piecewise max-pooling within each of the segments after convolution, the original sentence would be a more informative representation while still maintaining a vector that is independent of the input sentences length, which can alleviate the impact of long length sentences on relation extraction.

The disadvantage of this model is how multi-instance problem was set in the loss function. The paper defined the loss for training of the model as follows. Given T bags of sentences with each bag containing q_i sentences and having the label $y_i, i = 1, 2, ..., T$, the neural network gives the probability of extracting relation r from sentence j of bag i, d_i^j denoted as follows:

$$p(r|d_i^j, \theta); j = 1, 2, ..., q_i \tag{1}$$

where θ is the weight parameters of the neural network. Then the loss is given as follows:

$$J\left(\theta\right) = \sum_{i=1}^{T} \log\left(y_i|d_i^j, \theta\right) \tag{2}$$

$$j^* = \arg\max_j \; p\left(y_i|d_i^j, \theta\right); j = 1{,}2, \; ... \; {,}q_i \tag{3}$$

PCNN uses the one most-likely positive sentence only for the entity pair to reduce the noise during the training and prediction stage with Eq. 2. It means that the model ignore almost all other sentences in the bag. Even though not all the sentences in the bag express the true positive relation between the entity pair, information expressed by these sentences in the bag are useful.

The PCNN with Multi-instance learning for relation extraction is shown to outperform the traditional non-deep models such as the distant supervision based model proposed by Mintz et al. [12]. As a result, it is always chosen as the baseline model.

4.2 Selective Attention over Instances

To address the drawbacks of the PCNN model which only used the one most-relevant sentence in a bag as the positive sample. Lin et al. [9] used the attention mechanism over all the instances in the bag to handle noise problem. In this model, each sentence d_i^j of bag i is first encoded into a vector representation, r_i^j as PCNN did. Then the representation of the bag is gotten by taking attention-weighted average of all the sentence vectors $(r_i^j, j = 1, 2, ..., q_i)$ in the bag. The model computes a weight α_j for each instance d_i^j of bag i. These values are dynamic in the sense that they are different for each bag and depend on the relation category and the given sentence. The final representation of the bag is given as follows:

$$r_i = \sum_{j=1}^{q_i} \alpha_j r_i^j \tag{4}$$

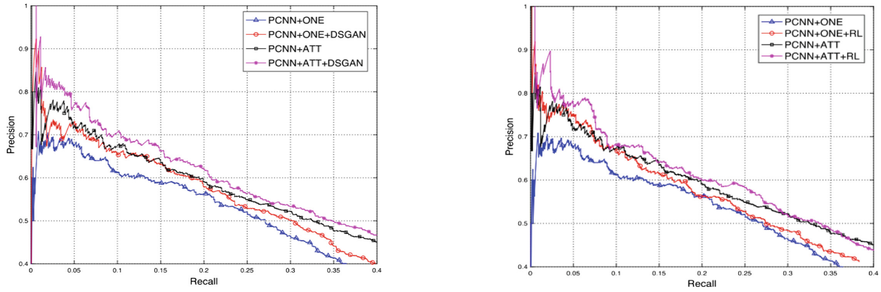
With the attention weighted representation of all the instances in the bag, the model is able to identify the importance of sentences from the noisy bag and all the information in the bag is utilized to predict the class of the relation.

4.3 Denoising Approach

Although the distant supervision relation extraction method can use the knowledge base to obtain a large amount of labeled data by annotating the texts automatically, much noise was introduce into the dataset, which would decline the performance of relation extraction system. Multi-instance learning models can be affected less, but it still fails to overcome the problem that all sentences in the bag is mislabeled. In order to reduce the noise of the bags, Qin et al. [14] proposed a method based on the Generative Adversarial Training to remove noise from the annotated-automatically data. The generator networks (G for short) estimates the probability distribution of the positive samples over a distant supervision bag. And then sampling positive sentences from noisy bag according to this probability distribution. The high-confidence samples generated by G are regarded as true positive samples. However, The discriminator (D for short) regards them as negative samples; conversely, the low-confidence samples are still treated as positive samples. For the generated samples, G maximizes the probability of being true positive; on the contrary, D minimizes this probability. The optimal G is obtained until the D has been greatest confused. As a result, the G is able to filter distant supervision training dataset and redistribute the false positive instances into the negative set, in which way to provide a cleaned dataset for relation classification.

Qin et al. [15] also proposed a denoise approach based on Deep reinforcement learning framework. The agent tries to remove the false positive samples from the distant supervision positive dataset P^{ori}. In order to get the reward, P^{ori} is split into the training set P_t^{ori} and the validation set P_v^{ori}; their corresponding negative part are represented as N_t^{ori} and N_v^{ori}. In each epoch i, the agent performs a series of actions to recognize the false positive samples from P_t^{ori} and treat them as negative samples. Then, a new relation classifier is trained under the new dataset $\{P_t^i, N_t^i\}$. With this relation classifier, F1 score is figured out from the new validation set $\{P_v^i, N_v^i\}$, where P_v^i is also filtered by the current agent. After that, the current reward is measured as the difference of F1 between the adjacent epochs. The above two algorithms are independent of the relation extraction model and are a plug-and-play technique that can be applied to any existing distant supervision relation extraction model.

The proposed methods on NYT dataset, the results are shown as Fig. 3. The experimental results show that the two methods can effectively remove noise and improve the extraction performance of distant supervision methods. Both method pipelines are independent of the relation prediction of entity pairs, so these models can be adopted as the true-positive indicator to filter the noisy distant supervision dataset before relation extraction. And the filter is more effective for selecting useful samples than the soft approach like [9,19] proposed.



(a) Aggregate PR curves of DS-GAN(source from [14]).

(b) Aggregate PR curves of Reinforcement Learning for Distant Supervision(source from [15]).

Fig. 3. The performance of the denoising approach based relation extraction methods.

4.4 Graph-Based Model

To the best of our knowledge, it is surprising to note that only a few works for relation extraction has tried to replace Convolutional Neural Networks with Recurrent Neural Networks for encoding the sentences. One important reason is that these methods hope to use the convolutional neural network to extract the combined semantic features between words or entities independent of position. Though RNNs intuitively fits more naturally to natural language tasks and can persevere the context information of the sentence.

In the distant supervision domain, Vashishth et al. [17] proposed RESIDE, a graph model based approach, which uses Bi-GRU over the concatenated positional and word embedding for encoding the local context of each word. For capturing long-range dependencies, the Graph Convolution Networks (GCN) over dependency tree is employed to get the syntactic information representation of each word. Then, attention over tokens is used to subdue irrelevant tokens and get an embedding for the entire sentence. Finally, This model use attention over sentences to obtain a representation for the entire bag, which is fed to a softmax classifier to get the probability distribution over the relations. The performance of RESIDE is evaluated in NYT dataset and the result is shown as Fig. 4. The results validate that GCNs are effective at encoding syntactic information, which is complementary to the context information captured by RNNs. This model could improve relation extraction with these information.

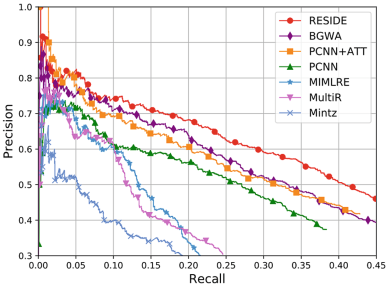


Fig. 4. The PR curves of RESIDE (source from [17])

In the supervised domain, Fenia Christopoulou et al. [4] proposed a walk based model for relation extraction. All the entities in a sentence are regarded as nodes in a fully-connected graph structure. The edges are on behalf of the position-aware contexts around the entity pairs. In order to capture different relation paths between two entities, The model construct up to a given length walks between each pair. The resulting walks are merged and iteratively used to update the edge representations. The model is evaluated on ACE 2005 for the task of relation extraction. And the model can achieve comparable performance compared with the state-of-the-art supervised relation extraction system without external syntactic tools. It shows that the dependencies between relations of entities can help extracting the final useful relations.

5 Conclusion

In general, the Entity relation extraction method with supervised learning has high accuracy, but it depends on the large labeled corpus and the construction of corpus is difficult. The unsupervised entity relation extraction does not need to define the entity relation type system in advance, which has domain independence. Scale open domain data has advantages that other methods can't match, but its clustering threshold is difficult to determine in advance. However, the distant supervision entity relation extraction only needs to label a small number of relation instances manually, which is suitable for entity relation extraction without labeling corpus, but its implementation process introduced noise into datasets, which makes the recall rate of the method lower. Many Successive works have tried to handle the noise and distant supervision assumption with mechanisms like selective attention and instance filter to improve the performance further by denoising. And the Graph-Based Model shows the huge potential to improve the relation extraction task by handling the dependency of entities. Future works for relation extraction can thus definitely try these approach to promote the RE system.

Acknowledgments. This work was supported by the National Natural Science Foundation of China [No. 11331012, Grant No. 91546201, No. 71331005, No. 71110107026, No. 11671379], UCAS Grant [No. Y55202LY00].

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: 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. ACM (2008)
3. Bunescu, R.C., Mooney, R.J.: Subsequence Kernels for relation extraction. In: NIPS (2005)
4. Christopoulou, F., Miwa, M., Ananiadou, S.: A walk-based model on entity graphs for relation extraction. In: ACL (2018)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018)
6. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam: Open information extraction: the second generation. In: IJCAI (2011)
7. Hendrickx, I., et al.: SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: SemEval@ACL (2010)
8. Jat, S., Khandelwal, S., Talukdar, P.: Improving distantly supervised relation extraction using word and entity based attention. CoRR abs/1804.06987 (2017)
9. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: ACL (2016)
10. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
12. 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, pp. 1003–1011. Association for Computational Linguistics (2009)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP (2014)

14. Qin, P., Xu, W., Wang, W.Y.: DSGAN: generative adversarial training for distant supervision relation extraction. In: ACL (2018)
15. Qin, P., Xu, W., Wang, W.Y.: Robust distant supervision relation extraction via deep reinforcement learning. In: ACL (2018)
16. 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
17. Vashishth, S., Yao, X., Gawade, S., Bhattacharyya, C., Talukdar, P.: Reside: improving distantly-supervised neural relation extraction using side information. In: EMNLP (2018)
18. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. J. Mach. Learn. Res. **3**, 1083–1106 (2002)
19. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: EMNLP (2015)
20. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: COLING (2014)
21. Zhou, G., Su, J., Zhang, J., Zhang, M.: Exploring various knowledge in relation extraction. In: ACL (2005)