Label-Free Distant Supervision for Relation Extraction via Knowledge Graph Embedding

Guanying Wang[1], Wen Zhang[1], Ruoxu Wang[1], Yalin Zhou[1]
Xi Chen[1], Wei Zhang[23], Hai Zhu[23], and Huajun Chen[1]*
[1]College of Computer Science and Technology, Zhejiang University, China
[2]Alibaba-Zhejiang University Frontier Technology Research Center, China
[3]Alibaba Group, China
21621253@zju.edu.cn

Abstract

Distant supervision is an effective method to generate large scale labeled data for relation extraction, which assumes that if a pair of entities appears in some relation of a Knowledge Graph (KG), all sentences containing those entities in a large unlabeled corpus are then labeled with that relation to train a relation classifier. However, when the pair of entities has multiple relationships in the KG, this assumption may produce noisy relation labels. This paper proposes a label-free distant supervision method, which makes no use of the relation labels under this inadequate assumption, but only uses the prior knowledge derived from the KG to supervise the learning of the classifier directly and softly. Specifically, we make use of the type information and the translation law derived from typical KG embedding model to learn embeddings for certain sentence patterns. As the supervision signal is only determined by the two aligned entities, neither hard relation labels nor extra noise-reduction model for the bag of sentences is needed in this way. The experiments show that the approach performs well in current distant supervision dataset.

1 Introduction

Distant Supervision was first proposed by Mintz (2009), which used seed triples in Freebase instead of manual annotation to supervise text. It marked text as relation r if (h, r, t) can be found in a known KG, where (h, t) is the pair of entities contained in the text. This method can generate large amounts of training data, therefore widely used in recent research. But it can also produce much noise when there are multiple relations between the entities. For instance in Figure 1, we may wrongly mark the sentence "*Donald Trump is the president of America*" as relation *born-in*, with the seed triple (*Donald Trump, born-in, America*).

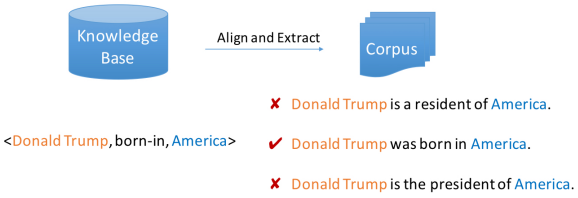

Figure 1: The mislabeled sentences produced by Distant Supervision.

Previous works have tried different ways to address this issue. One way named Multi-Instance Learning(MIL) divided the sentences into different bags by (h, t), and tried to select well-labeled sentences from each bag (Zeng et al., 2015) or reduced the weight of mislabeled data (Lin et al., 2016). Another way tended to capture the regular pattern of the translation from true label to noise label, and learned the true distribution by modeling the noisy data (Riedel et al., 2010; Luo et al., 2017). Some novel methods like (Feng et al., 2017) used reinforcement learning to train an instance-selector, which will choose true labeled sentences from the whole sentence set. These methods focus on adding an extra model to reduce the noisy label. However, stacking extra model does not fundamentally solve the problem of inadequate supervision signals of distant supervision, and will introduce expensive training costs.

Another solution is to exploit extra supervision signal contained in a KG. Weston (2013) added the confidence of (h, r, t) in the KG as extra supervision signal. Han (2018) used mutual attention of KG and text to calculate a weight distribution of train data. Both of them got a better performance by introducing more information from KG. However, they still used the hard relation label derived from distant supervision, which also brought

* Corresponding author.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2246–2255
Brussels, Belgium, October 31 - November 4, 2018.

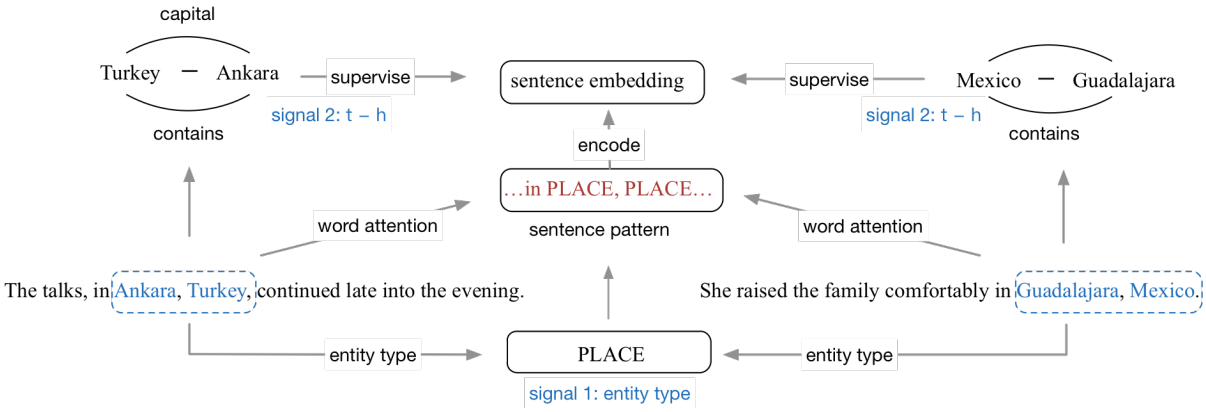

Figure 2: An instance of our label-free distant supervision method.

in much noise.

In this paper, we tend to avoid supervision by hard relation labels, and make full use of prior knowledge from a KG as soft supervision signal. We consider the TransE model proposed by Bordes (2013), which encodes entities and relations of a KG into a continuous low-dimensional space with the translation law $h + r \approx t$, where h, r, t describe the head entity, the relation and the tail entity respectively. Inspired by TransE model, we use $t - h$, instead of a concrete relation label r, as the supervision signal and make the sentence embedding close to $t - h$. Concrete relation labels may introduce mislabeled sentences, while $t - h$ is label-free, which is only determined by the two aligned entities and the the translation law.

Our assumption is that *each relation r in a KG has one or more sentence patterns that can describe the meaning of r*. For the example in Figure 2, we first replace the entity mentions in a sentence with the types of the aligned entities in the KG to form a sentence pattern. For example, "in Guadalajara, Mexico" will be replaced by "in PLACE, PLACE" to form a sentence pattern "in A, B" which conveys the meaning of "B contains A" and indicates the relation $contains$. For this sentence pattern, there may be a group of sentences sharing the same pattern but with different aligned entity pairs. In the first sentence "*The talks, in Ankara, Turkey, continued late into the evening*", $(Turkey - Ankara)$ implies both "*/location/country/capital*" and "*/location/location/contains*" as there are multiple relations between Ankara and Turkey in the KG. But in the similar sentence "*She raised the family comfortably in Guadalajara, Mexico.*", $(Mexico - Guadalajara)$ only implies "*/location/contains*" as there is no relation of "*/location/country/capital*" between Mexico and Guadalajara in the KG. As both $(Turkey - Ankara)$ and $(Mexico - Guadalajara)$ will be used to supervise the learning of the encoder for the pattern "in A, B", it makes the embedding of the sentence pattern closer to the correct relation "*/location/location/contains*" instead of the wrong relation "*/location/country/capital*". In this way, we do not need to label the sentences with the hard relation labels anymore.

The main contributions of this paper can be summarized as follows:

- As compared to existing distant supervision for relation extraction, our method makes better use of the prior knowledge derived from KG to address the wrong labeling problem.
- The proposed approach tends to supervise the learning process directly and softly by the type information and translation law, both derived from KG. Neither hard labels nor extra noise-reduction model for the bag of sentences is needed in this way.
- In the experiments, we show that the label-free approach performs well in current distant supervision dataset.

2 Related works

Relation extraction is intended to find the relationship between two entities given an unstructured text. Traditional methods use artificial characteristics or tree kernels to train a classification model (Culotta and Sorensen, 2004; Guodong et al., 2002). Recent works concentrate on deep neural

networks to avoid error propagation during generating features (Ebrahimi and Dou, 2010; Zeng et al., 2014; Zhou et al., 2016; Zheng et al., 2017). More complicated models were proposed to learn deeper semantic features, like PCNN (Zeng et al., 2015) and attention pooling CNN (Wang et al., 2016), graph LSTMs (Peng et al., 2017).

Most of the early works were trained on the standard dataset by manual annotation, such as SemEval-2010 Task 8. In actual scenarios, it will cost a lot of manual resources to generate labeled data. Distant supervision (Mintz et al., 2009) aimed to obtain large-scale training data automatically, which becomes the most versatile supervision method. However, it suffers from the noisy label problem. Many works concentrate on dealing with the noise of distant supervision. Multi-instance learning (Riedel et al., 2010; Surdeanu et al., 2012) addresses the problem in bag-level, which divides sentences into different bags by (h, t). Zeng (2015) selects the most correct sentence from each bag. Lin (2016) introduces attention mechanism by distributing different weight to each sentence in the same bag, which reduces the effect of noisy labels and increases utilization of train data. Luo (2017) uses a transition matrix to characterize the inherent noise, convert true distribution to noise distribution. The model is enhanced by curriculum learning. Feng (2017) trains an instance selector to select correct labeled sentences by reinforcement learning.

Most of the above methods introduce a complicated extra model to deal with the noisy label problem. Our work tends to avoid the noisy label from distant supervision, by using entity information and translation law in KG to introduce more supervision signal.

KG is composed of many triples like (*head, relation, tail*), which describe relationships between head entities and tail entities. TransE is first proposed by (Bordes et al., 2013) to encode triples into a continuous low-dimensional space, which based on the translation $h+r \approx t$. Many follow-up works like TransH (Wang et al., 2014), DistMult (Yang et al., 2014), and TransR (Lin et al., 2015), proposed advanced method of translation by introducing different embedding spaces. Some recent works attempt to jointly learn text and KG triples, including (Xie et al., 2016) and (Xiao et al., 2016). These models tend to strengthen the representation of entities and relationships for KG tasks, but not for text representation.

3 Methodology

Here we present LFDS (Label-Free Distant Supervision) that essentially avoids noisy labels introduced by traditional distant supervision. Figure 2 shows an instance of our method. First, we pre-train representations for entities and relations based on the translation law $h + r \approx t$ defined by typical KG embedding models such as TransE. Second, for each sentence in the train sets, we replace the entity mentions with the types of the entities in the KG. An attention mechanism is then applied to calculate the importance of words with regard to the sentence pattern. Third, we train the sentence encoder by the margin loss between $t-h$ and sentence embedding. Note we do not use the noisy relation labels to train the model. Finally, for prediction, we calculate the embedding of test sentences, then compare the sentence embedding with all relation embeddings learned by TransE, and choose the closest relation as our predicted result. We describe these four parts in details as below.

3.1 KG Embedding

We use typical KG embedding models such as TransE to pre-train the embedding of entities and relations. We intend to supervise the learning by $t - h$ instead of hard relation label r. Concretely speaking, given two entities, h and t, we regard the translation based upon TransE between h and t as the target relation representation. TransE interprets relationships as translations operating on low-dimensional embeddings of entities, with the formula $h+r \approx t$, where h, r, t represent head entity, relation, and tail entity separately. The model is proved to perform well in predicting the tail entity when given head entity and relation.

The problem is that there may be multiple relations between t and h. As the example in Figure 2, the vector calculated by $Turkey - Ankara$ contains information for both relations: "*/location/country/capital*" and "*/location/location/contains*". While supervising the learning of the sentence pattern "in PLACE, PLACE", it is difficult to distinguish the two relations by supervision signal from only one sentence. However, other sentences with the similar pattern but different aligned entity pairs can push the embedding of the pattern close to another

vector, such as $Mexico - Guadalajara$, which only represents "*/location/location/contains*" relation. As a result, the pattern will be closer to its correct relation "*/location/location/contains*".

Our work chooses TransE instead of other KG embedding models such as TransH or TransR, because TransE builds representations for h and t independent from fixed relation type r as the model assumes we do not know the specific relation r when training the encoder with supervision from $t - h$.

3.2 Sentence Embedding

In order to get a better representation of sentences, we had tried a variety of NRE models, such as Bi-LSTM(Zhou et al., 2016), SDP-LSTM (Yan et al., 2015), and typical CNN models. We chose PCNN (Zeng et al., 2015) to encode the sentence finally, which performs the best in our experiments. The encoder contains three parts as below.

Word Embeddings and Attentions. Instead of encoding sentences directly, we first replace the entity mentions e in the sentences with corresponding entity types $type_e$ in the KG, such as PERSON, PLACE, ORGANIZATION, etc. We then pre-train the word embedding by word2vec.

Attention mechanism is further applied to capture the importance of words with regards to the types information of entities as we assume the words close to the types information are more important.

First, we calculate the similarity between each word w^j and two entity types respectively:

$$A_1^j = f(type_{e_1}, w^j) \quad (1)$$

$$A_2^j = f(type_{e_2}, w^j) \quad (2)$$

$f(type_e, w^j)$ is the similarity function, which is defined as cosine similarity in this paper. $type_{e_1}$ and $type_{e_2}$ are the embeddings of the two entity types. Then the weight distribution for each word can be derived by exponential function:

$$\alpha_1^j = \frac{exp(A_1^j)}{\sum_{i=1}^{n} exp(A_1^i)} \quad (3)$$

$$\alpha_2^j = \frac{exp(A_2^j)}{\sum_{i=1}^{n} exp(A_2^i)} \quad (4)$$

We use the average weights of two entities as the attention of word w^j. Finally, the word embedding WF^j is derived as follows:

$$WF^j = \frac{\alpha_1^j + \alpha_2^j}{2} * w^j \quad (5)$$

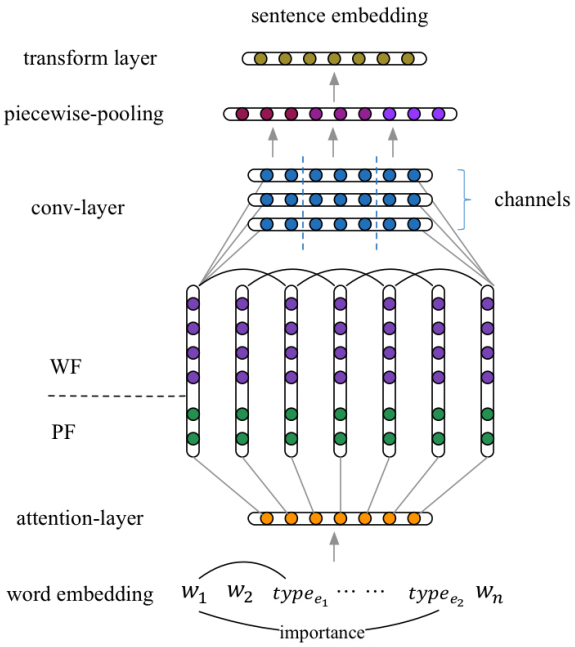

Figure 3: The sentence encoder with word attention and PCNN.

Position embedding. Zeng (2014) first proposed PFs to specify entity pairs. PF is a series of relative distances from current word to the two entities. For instance, for the sentence "*Damascus, the capital of Syria*", the distances from "capital" to the two entities are 3 and -2 respectively. The initial embedding matrix is randomly generated. Then we look up vector in the matrix by the two relative distances. The final position embedding will be the concatenation of $[PF_1, PF_2]$. As a result, we get a representation for each word:

$$w^j = [WF^j, PF_1^j, PF_2^j]$$

Then the input sentence representation will be:

$$x = w^1, w^2, ..., w^n$$

Piecewise-CNN. It was proved by (Zeng et al., 2015) that piecewise max pooling layer performs well in relation extraction, which tends to capture structural information between two entities. For each sentence, we use CNN to obtain a representation, then divide it into three parts by the two entities index. For each part, we perform a max pooling layer, thus we get 3-dimensional vector:

$$p_i = [p_{i_1}, p_{i_2}, p_{i_3}]$$

The shape of final vector will be $(bz, dc * 3)$, where bz represents batch size and dc is the number of channels.

The structure of whole model is shown in Figure 3.

3.3 Margin loss

In order to make the sentence embedding encoded by the PCNN model and relation embedding specified by $t-h$ based on the translation law as close as possible, we use margin loss with linear layer instead of cross-entropy loss with softmax layer. For the sentence embedding via PCNN layer, we perform a linear transformation to make its dimension equal to the relation representation.

$$s_e = \mathbf{W} * PCNN(x) + \mathbf{b} \tag{6}$$

Where $\mathbf{W}$ is the transformation matrix with shape $(dc*3, embedding\_dim)$. Then we define margin loss between $t-h$ and s_e as follows:

$$L = \sum_{s_e \in S} [(t-h-s_e+\gamma-(rand(t'-h, t-h')-s_e))]_+ \tag{7}$$

Where $rand(a, b)$ means choosing a or b. $t'-h$ is a negative instance of $t-h$, which is generated by randomly replacing t with other entities in KG, so does $t-h'$. For each sentence, we decrease the distance between $t-h$ and s_e, while increase the distance between the negative instance and s_e. γ is the reasonable margin between positive triple and negative triple. If the margin is already larger than γ, the loss of the sentence will be zero.

Another point to note is the special label NA in the dataset, which means there is no relationship between the two entities in the KG. In this case, $t-h$ is pointless and will confuse our encoder. To deal with this issue, we generate a fixed relation for NA, used as the negative relation for those sentences having some relationships. The minimum distance from NA to other relations is forced to be greater than $2*\gamma$, where γ is the margin in loss function. When the model is used for prediction, the NA is also included.

The training target of our model is shown as Figure 4, including the sentence encoder we introduced above.

3.4 Prediction

We build a sentence encoder which can output a sentence embedding with the same dimension as relation embedding from the KG. For a new test sentence, we first encode it with the model, then calculate the similarity between the sentence embedding and the embeddings of all candidate relations. The most similar relation to the sentence embedding is the predicted category.

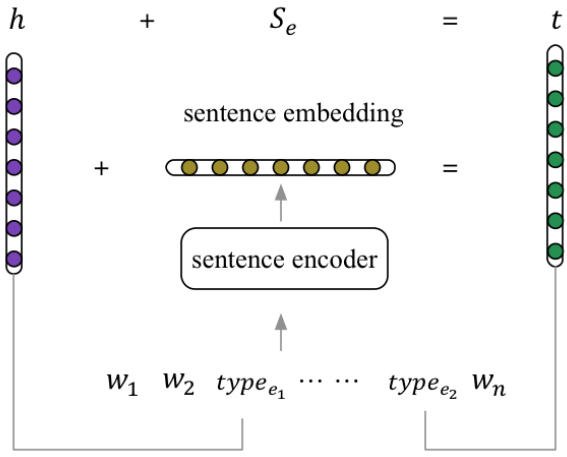

Figure 4: The training target.

$$r = arg\max_i(f(S_e, r_i)) \tag{8}$$

4 Experiments

Our experiments aim to provide positive evidence for the two main questions: (1) Whether or not the sentence pattern can express the essential part of the sentence? (2) Whether the abundant supervision signal in a KG is helpful to predict the true label for those mislabeled sentences?

To this end, we first introduce the widely used dataset for distant supervision, and evaluate our performance on the dataset. To further investigate the effectiveness of our model with noisy data, We divide the sentences in dataset into different categories, and show the study about some specific cases.

4.1 Datasets

The most widely used dataset was generated by Riedel (2010). It aligns the entities in Freebase with the New York Times (NYT) corpus, which contains all the news during 2005-2007. The sentences derived from news in 2005-2006 were used as the training data, while those from year 2007 were used as test data. After the alignment, there are 522,611 training sentences and 172,448 test sentences, labeled by 53 candidate relations in Freebase, and an extra label NA, which means there is no relation between the two entities in Freebase.

According to previous work (Mintz et al., 2009), we evaluate our model in the held-out evaluation and manual evaluation. The held-out evalu-

Parameter	Settings
Kernel size k	3
Sentence embedding size	100
Word embedding size	50
Position embedding size	5
Number of Channels	250
Margin	2
Learning rate	0.001
Dropout	0.5
Batch size	128

Table 1: Parameter settings.

ation calculates the precision-recall curves on the whole test set. For the false positives produced by the noisy labels in the test data, the precision will drop rapidly as the recall increases. In order to measure the precision, we need manual evaluation to check misclassified samples.

4.2 Experimental settings

4.2.1 Word Embeddings

In this paper, we use word2vec to train word embeddings on the NYT corpus. The window size of word2vec model is set as 5, and the embedding size is 50. We preserve those words appearing more than 10 times as vocabulary.

4.2.2 KG embeddings

We train the entities and relationships on FB40k[1] (Lin et al., 2015), which is generated for knowledge graph completion, with about 40,000 entities and 1318 relations. We set the embedding size as 100 instead of 50, which performs better in our experiment. Besides, we set the margin as 1 and train with learning rate 0.01. In order to test the performance of the vectors, we evaluate our model in KG completion tasks. The hit@10 of our final TransE model is 0.67, which is evaluated by predicting the closest 10 tail entities with specified head entities and relationships.

4.2.3 Parameter Settings

We use three-fold validation to determine the hyper-parameters. In the network layer, we try {3, 4, 5} for the kernel size, {100, 150, 200, 250, 300} for the number of channels, {5, 10, 15} for the position embedding size. In the update procedure, we use adaptive gradient descent with trying {0.1, 0.05, 0.01, 0.001} for the initial learning rate, and {64, 128, 256} for the mini-batch size. In the dropout operation, we set the probability as 0.5 referring to most of the classical experiments. Table 1 shows our final setting for all hyper-parameters.

[1]https://github.com/thunlp/KB2E.

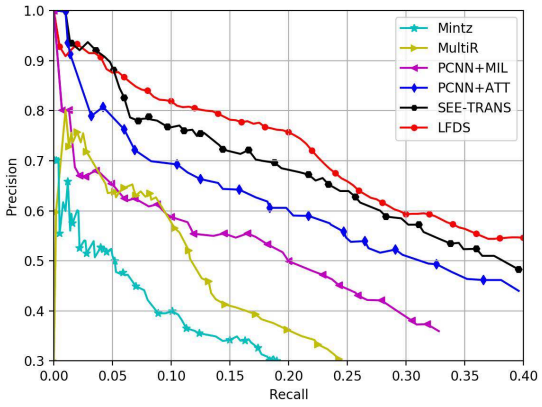

Figure 5: Performance comparison with Traditional methods.

4.3 Comparison with Traditional Methods

4.3.1 Held-out Evaluation

The held-out evaluation is performed directly on the test data. For the labels produced by distant supervision may not be precise, held-out evaluation is an approximate measure of our model, which is usually depicted by the precision-recall curve.

We select six representative models for comparison. *Mintz* (Mintz et al., 2009) proposed a feature-based model that first used distant supervision. *MultiR* (Hoffmann et al., 2011) is a multi-instance learning model under the at-least-one assumption. *PCNN+MIL* (Zeng et al., 2015) proposed the piece-wise pooling method, which is used as the encoder of our works. *PCNN+ATT* (Lin et al., 2016) performed selective attention over instances and got better results in the datasets. *SEE* (He et al., 2018) is a novel work that learned syntax-aware entity embedding for relation extraction and achieved state-of-the-art. The precision-recall curves are shown in Figure 5, where *LFDS* denotes our label-free distant supervision method.

We can observe from the figure that our LFDS method has an overall good performance compared to current works, especially with the growth of recall. It demonstrates that our model has a good classification ability in general, because the sentence pattern can capture the meaning of relations better than a sentence. The result can answer the first question we proposed at section 4.

Accuracy	Top 100	Top 200	Top 500	Average
Mintz	0.77	0.71	0.55	0.676
MultiR	0.83	0.74	0.59	0.720
PCNN+MIL	0.86	0.80	0.69	0.783
PCNN+ATT	0.86	0.83	0.73	0.807
SEE	0.91	0.87	0.77	0.850
LFDS	0.90	**0.88**	**0.83**	**0.869**

Table 2: Precision values for the top 100, 200 and 500 sentences.

4.3.2 Manual Evaluation

For the wrong labels produced by distant supervision, there will be many false positives in our evaluation inevitably, thus causing a sharp decline in the held-out precision-recall curves. Manual evaluation is necessary to evaluate the model more precisely. Following the previous works, we selected the top 100, top 200, and top 500 sentences, which is ranked by the predicted confidence, then evaluated the precision artificially. The result is shown in table 2.

We can see that the precision is higher than held-out evaluation, because manual evaluation avoid the effect of wrong labels. Our LFDS method achieved a consistently higher precision compared with current works, especially when recall increases. Compared to held-out evaluation, manual evaluation can show our model's ability in differentiating noisy sentence. Detail analysis will be shown in Section 4.4.

In the manual procedure, we found some wrong cases caused by entity types. The entity types in Freebase can be ambiguous, where "ORGANIZATION" may be confused with "PLACE". It causes error propagation in our experiments.

4.4 Case Study

To further prove the effectiveness of our model, especially in distinguishing noisy labels, we select some specific relationships for detail analysis. The noisy labels are produced by the entity pairs which have multiple relationships between them. In this case, different relationships will share the same entity pairs in knowledge graph. We defined this kind of relationships as "overlapping" relationships. The more entity pairs it shares with other relation, the overlapping degree of the relation is higher, which means the relation is harder to distinguish.

Case 1: Non-overlapping Relations. The first case is the non-overlapping relation. For triples of the non-overlapping relation r_1 as (h, r_1, t), there are few triples like (h, r_2, t) in KG, where r_2 is another relation in our candidate relations set. That means for this kind of relation, almost no noisy label will be produced. One of these relation is */business/person/company*. There are near 200 sentences in the test set, with our evaluation of precision achieving 0.98. It proves that our encoder with sentence pattern and label-free supervision is effective in basic classification, which is a convincing answer of the first question we proposed at section 4.

Case 2: Partly-overlapping Relations. The second case is the partly-overlapping relation, in which two relations may share a certain number of entity pairs in Freebase. For instance, the relation */location/country/capital* shares many entity pairs with */location/location/contains* but not all entity pairs in Freebase have both $capital$ and $contain$ relations.

For those entity pairs having both relations, traditional distant supervision would produce two labels for sentences such as:

> "The talks, in Ankara, Turkey, continued late into the evening."

The noisy labels in the train set are hard to differentiate. Recent noise reduction methods commit to improving the distinguishing ability of the model by adding extra models. Our experiment proves that our label-free supervision method not only achieves better differentiation performance but also does not need to train extra noise reduction models. Cases are shown in Table 3.

The prediction results indicate that the model is capable of learning the embedding of the sentence pattern we want. For instance, the model captures the pattern like "in PLACE, PLACE", and tends to predict the sentence with this pattern for */location/location/contains*, while the pattern "PLACE, the capital of PLACE" for */location/country/capital* respectively. When both two relations are labeled for the same sentence in the test set, our model can predict the correct label with the corresponding patterns.

Another similar but more interesting example is */people/person/nationality* and */people/person/place_lived*. In this case, the two relations share a certain number of entity pairs in Freebase like the previous example. But because of the incompleteness of Freebase, many sentences with only one label are actually wrongly labeled.

Sentence	Label with normal distant supervision	Prediction with LFDS	Pattern
The talks, in *Ankara*, *Turkey*, continued late into the evening.	/location/location/contains /location/country/capital	/location/location/contains	in PLACE, PLACE
..., said Mr.Cho, 25, who was born in *Seoul*, *South Korea*, and educated at a boarding school in Scotland.	/location/location/contains /location/country/capital	/location/location/contains	in PLACE, PLACE
On Wednesday, suicide bombings killed 33 people in *Algiers*, the capital of *Algeria*.	/location/location/contains /location/country/capital	/location/country/capital	PLACE, the capital of PLACE
Farah has lived in *India*, Europe and South Africa, and only started revisiting Mogadishu in 1996, after two decades away.	/people/person/nationality	/people/person/place_lived	PERSON lived in PLACE
He was George Mcgovern of South Dakota – not *Frank church* of *Idaho*, who was involved in other antiwar legislation.	/people/person/place_lived	/people/person/nationality	PERSON of PLACE

Table 3: The comparison between labels from normal distant supervision and our label-free relation prediction

For example,the sentence "Farah has lived in India, ..." is labeled with only one relation */people/person/nationality* because there is only one $nationality$ relation in Freebase. But the actual meaning of the sentence is to say $Farah$'s $place\_lived$ is $India$. This type of wrongly labeling problem is caused by incompleteness of Freebase which is very common for many other knowledge graphs.

However, our label-free method can correct this problem because it essentially learns the sentence patterns that are determined only by the sentence itself and the aligned entity pairs. As shown by the last two examples in Table 3, our model successfully learned the patterns "PERSON lived in PLACE" for */people/person/place_lived* and "PERSON of PLACE" for */people/person/nationality* respectively.

These instances show that our model is capable of learning some sentence patterns and mapping them to the corresponding relations in Freebase, which can distinguish noise sentences effectively. It indicates that our label-free supervision with prior knowledge introduced by the translation laws and entity types in KG is effective in avoiding noise, which can answer the second question we proposed at section 4 credibly.

Case 3: Mostly-overlapping Relations. The final case is mostly-overlapping relations, in which the two relations share most entity pairs in Freebase. One example is */people/person/place_of_birth*, which shares most of its entity pairs with */people/person/place_lived* in FB40k, because a person's birthplace and residence are likely to be the same. That means in the process of training with TransE, the two relations are updated by similar gradients, which will produce similar representations for $t - h$. In this case, the relations are really hard to differentiate, because there are not enough distinct supervision signals in the KG. We tend to resolve this situation in future work by utilizing prior knowledge derived from relation paths.

5 Conclusion

In this paper, we argue that the noise label problem in distant supervision is mainly caused by the incomplete use of KG information. Thus we propose a label-free distant supervision method, which supervises the learning of the embedding of sentence patterns by $t - h$ and entity types, instead of hard relation labels. We conducted experiments on the widely used relation extraction dataset and showed that with the recall increasing, our model performs better than state-of-the-art results. This demonstrates that our approach can effectively deal with the noise problem and encoding sentence pattern for relation extraction.

In the future, we plan to utilize more information in knowledge graphs to improve the distant supervision signal. For instance, the reasoning path can introduce new prior knowledge, which is a key direction in current works of KG. The path may produce new supervision signals for two entities even there is no direct connection between them. We also plan to apply this method to other

datasets.

Acknowledgments

This work is funded by NSFC 61673338/61473260, and supported by Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *International Conference on Neural Information Processing Systems*, pages 2787–2795.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Meeting on Association for Computational Linguistics*, pages 423–429.

Javid Ebrahimi and Dejing Dou. 2010. Chain based rnn for relation classification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1244–1249.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2017. Reinforcement learning for relation classification from noisy data.

Zhou Guodong, Su Jian, Zhang Jie, and Zhang Min. 2002. Exploring various knowledge in relation extraction. In *ACL 2005, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, Usa*, pages 419–444.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text.

Zhengqiu He, Wenliang Chen, Zhenghua Li, Meishan Zhang, Wei Zhang, and Min Zhang. 2018. See: Syntax-aware entity embedding for neural relation extraction.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550.

Yankai Lin, Zhiyuan Liu, Xuan Zhu, Xuan Zhu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2181–2187.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Meeting of the Association for Computational Linguistics*, pages 2124–2133.

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. pages 430–439.

Mintz, Mike, Steven, Jurafsky, and Dan. 2009. Distant supervision for relation extraction without labeled data. In *Joint Conference of the Meeting of the ACL and the International Joint Conference on Natural Language Processing of the Afnlp: Volume*, pages 1003–1011.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen Tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms.

Sebastian Riedel, Limin Yao, and Andrew Mccallum. 2010. Modeling relations and their mentions without labeled text. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465.

Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Meeting of the Association for Computational Linguistics*, pages 1298–1307.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1112–1119.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. pages 1134–1137.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016. Ssp: Semantic space projection for knowledge graph embedding with text descriptions.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions.

Xu Yan, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency path. *Computer Science*, 42(1):56–61.

Bishan Yang, Wen Tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases.

D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. 2014. Relation classification via convolutional deep neural network.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257(000):1–8.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Meeting of the Association for Computational Linguistics*, pages 207–212.