Improving Distant Supervision of Relation Extraction with Unsupervised Methods

Conference Paper · November 2016

DOI: 10.1007/978-3-319-48740-3_42

7 authors, including:

Min Peng
Wuhan University
71 PUBLICATIONS **594** CITATIONS
SEE PROFILE

Hua Wang
Victoria University Melbourne
299 PUBLICATIONS **4,570** CITATIONS
SEE PROFILE

Improving Distant Supervision of Relation Extraction With Unsupervised Methods*

Min Peng[1], Jimin Huang[1], Zhaoyu Sun[1], Shizhong Wang[2], Hua Wang[3], Guangping Zhuo[4], and Gang Tian[1](✉)

[1] School of Computer, Wuhan University, Wuhan, China
[2] School of Economics and Management, Wuhan University, Wuhan, China
{pengm, huangjimin, sunzhaoyu, szwang, tiang2008}@whu.edu.cn
[3] Centre for Applied Informatics, Victoria University, Melbourne, Australia
hua.wang@vu.edu.au
[4] Department of Computer Science, Taiyuan Normal University, China
zhuoguangping@163.com

Abstract. Distant supervision has been widely adopted in relation extraction task since it does not require any labeled data. It can automatically align knowledge base with corpus to generate training data. However, the intuition base of alignment in this method may fail, resulting in wrong label problem. In this paper, we try to improve the intuition of distant supervision from the perspective of relation mentions, and propose a novel method called *Clustered DS* which employs our improved intuition in an unsupervised manner. By incorporating the information about the distribution of relation mentions, our method achieves a more precise alignment, thus it significantly reduces the number of wrong labels. Experimental results demonstrate the advantage of *Clustered DS* over existing distant supervision methods and show the effectiveness of our improved intuition.

Keywords: Information Extraction, Relation Extraction, Distant Supervision, Unsupervised Method.

1 Introduction

Distant supervision(DS)[1] is one of the most promising methods in information extraction(IE) to extend traditional supervised methods to web-scale dataset. It automatically generates training data by aligning an external knowledge base with free texts. In this paper, we focus on the task of relation extraction(RE), one of the subproblems of IE, which seeks to extract relations from corpora such as Wikipedia. For example, given an *entity pair* (`Barack Obama`, `United States`), the task of RE aims at extracting new relation instance (`Barack Obama`, `state_of_birth`, `United States`) from the first sentence of Fig.1. The alignment of DS in this field is based on an intuition, or a hypothesis, which leverages

* This research is supported by the Natural Science Fundation of China(No.61472291), and the Natural Science Fundation of Hubei Province, China(No.ZRY2014000901).

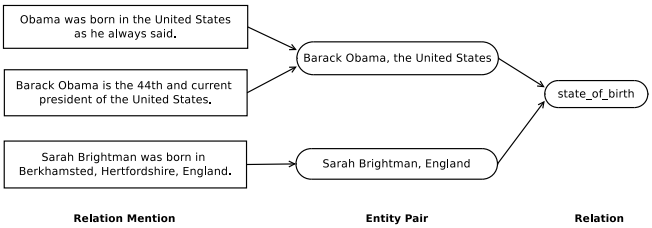

Fig. 1. Training examples generated via DS.

external knowledge base. Given a knowledge base D and a corpus L, the intuition assumes that for any entity pair (e_1, e_2) belonging to the relation r in D, all relation mentions of (e_1, e_2) express the same relation r. For instance, we can generate 3 training examples from sentences in Fig.1.

Though the intuition can significantly reduce human efforts in generating training data, it introduces two major challenges into the task. First, not all relation mentions express the same relation as their entity pairs, which is called the challenge of *false positive*. For instance, as shown in Fig.1, the entity pair (`Barack Obama`, `United States`) has two relations: `state_of_birth` and `employed_by`. However, relation mentions will always be labeled with `state_of_birth`, no matter which relation they actually express. Riedel et al. [2] reported that there are over 31% of false positives in the NYT corpus when they implied the intuition with Freebase [3]. Second, knowledge bases are usually highly incomplete. There are massive relation mentions explicitly expressing one of known relations, but their entity pairs are not in the knowledge base. For example, Min et al. [4] showed that 93.8% of `persons` from Freebase have no `state_of_birth`. These relation mentions will be generated as *false negatives* via the intuition of DS.

Different from previous researches, we improve the intuition of DS from the perspective of relation mentions to address these two challenges. Rather than focusing on entity pairs, we leverage the similarity between relation mentions to achieve a more precise alignment. Along this line, we propose a novel two-stage method, called *Clustered DS*. At the first stage, an unsupervised method is adopted to divide relation mentions into different bags, using the information of the relation mentions instead of their entity pairs. At the second stage, we assign relation mentions from the same bag to the same relation that holds most of entity pairs in the knowledge base. Experimental results show that *Clustered DS* yields an average increase of 3% F1 compared with previous DS methods, indicating the effectiveness of our method and the improved intuition.

2 Related Work

Since it was first proposed by Craven et al. [1], distant supervision has gained wide attraction in the relation extraction field as it can automatically generate training data without labeled data [2, 5–10, 4, 11–14]. To tolerate the wrong labels introduced by the intuition of distant supervision, Riedel et al. [2] adopt

Multiple Instance Learning(MIL) with the assumption that at-least-one of the relation mentions in each bag express the relation assigned via the intuition of distant supervision. MultiR [9] and Multi-Instance Multi-label(MIML) learning [10] further support multiple relations expressed by different sentences in a bag.

As these methods bypass the challenge of false negative, Min et al. [4] introduce a new latent variable in MIML, which allows them to learn from only positive and unlabeled data. Takamatsu et al. [11] model the probabilities of a pattern expressing relations and remove mentions that match low-probability patterns. Similar to our method, Min et al. [14] relabel examples to their most likely relations. However, they estimate the probabilities of patterns rather than relation mentions.

Several other researches also attempt to improve distant supervision with supervised learning. Pershina et al. [12] propose *Guided DS* to utilize labeled data to guide MIML during training. Angeli et al. [13] apply active learning to the distant supervision and design a novel criterion to sample examples which are both uncertain and representative. Compared with their methods, our *Clustered DS* gives the same improvement without requiring any labeled data.

3 Intuition Improvement

In this paper, our goal is to reduce the wrong labels introduced by the generation of training data. From the perspective of relation mentions, we observe that relation mentions of the same entity pair may express different relations. At the same time, those expressing the same relation may not share the same entity pair. We turn to explore what is the crux between relation mentions expressing the same relation if their entity pairs differ. Obviously, the relation mentions expressing the same relation are similar to each other in the semantic level. For instance, sentences of (`Barack Obama`, `United States`) and of (`Sarah Brightman`, `England`) are highly similar and indeed express the same relation, as shown in Fig.1. Based on our observation, we make an assumption about the distribution of relation mentions in the semantic space. We assume that relation mentions expressing the same relation are more similar than those expressing different relations. We then improve the intuition of DS with our assumption, which also can be deemed as a two-phase procedure:

1. Divide relation mentions into bags according to the similarity between relation mentions.
2. Align the whole bag to the relation that holds most of the entity pairs of the bag in the knowledge base.

The main difference of our improved intuition, motivated by our observation, is that we use the similarity of relation mentions to guide the generation of the division. According to our assumption, it guarantees that relation mentions from the same bag indeed express the same relation. In fact, the intuition of DS latently makes the same commitment, but the division generated based on entity pairs is not precise enough to support it. In Fig.2, we present a 2D visualization of the division generated by the traditional intuition and our improved intuition

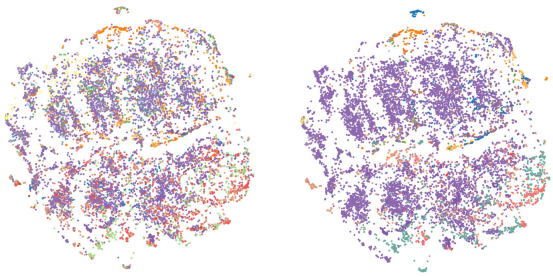

Fig. 2. The division generated by the traditional intuition(Left) and our improved intuition(Right). A point represents a relation mention in the semantic space whose color denote the bag it belongs. Best viewed in color.

using *t-Distributed Stochastic Neighbor Embedding*(t-SNE)[15]. It is now capable for relation mentions with different entity pairs to be divided to the same bag that is likely to express the same relation. Additionally, relation mentions whose entity pairs never appear in the knowledge base are now possible to be aligned to a known relation, rather than directly labeled with `no_relation`. Therefore, we can address both two issues with our improvement. Based on our improved intuition, we propose a novel two-stage method, called *Clustered DS*.

4 Division Generation

In the first stage of *Clustered DS*, an unsupervised method is adopted to generate the division of relation mentions. To deal with millions of relation mentions, Kmeans clustering, one of unsupervised methods, is a promising choice that is simple and efficient to perform in web-scale. The method iteratively refines centers in two steps with k initial cluster centers:

1. Assign datum to its nearest center.
2. Update cluster centers with the mean of its belonged datum.

The algorithm converges when there is no change in assignment. In our method, we first extract feature vectors from relation mentions by leveraging the feature set developed by Surdeanu et al. [16], which is the finest handcrafted feature set in relation extraction. Kmeans clustering then takes the input as feature vectors and generate k bags. It naturally leverages our improved intuition to reform the division according to the similarity between relation mentions.

The extracted vector is sparse and high-dimentional, which means we are unable to preform methods that automatically decides the cluster number k such as the *Chinese Resturant Process* [17]. Nevertheless, it is reasonble to set k twice the number of known relations or even higher, for the given set of relations obviously is just a small part of all relations. More importantly, even if the cluster number is much higher than the number of relations expressed in the corpus, small clusters can be assigned with the right relation in the following stage if the division is precise enough.

5 Relation Assignment

In the second stage of *Clustered DS*, bags generated in the previous stage are aligned to the known relations or `no_relation`. Since relation mentions from the same bag no longer share the same entity pair, we align the whole bag to the relation holding most of entity pairs of that bag. Given a bag M, the probability of the relation r is calculated in two steps:

1. For each relation mention m of M, calculate the probability $P(m \in r)$ according to its entity pair.

$$P(m \in r) = \begin{cases} 0 & \text{if entity pair of } m \notin r, \\ 1 & \text{if entity pair of } m \in r. \end{cases} \tag{1}$$

2. Calculate the probability $P(M \in r)$ by normalizing the sum of $P(m \in r)$.

$$P(M \in r) = \frac{\sum_{m \in M} P\left(m \in r\right)}{|M|} \tag{2}$$

We then assign M the most possible relation r.

6 Experiments

6.1 Dataset

We use the KBP dataset publicly released by Surdeanu et al. [10]. The dataset contains 1.5 million documents and nearly 110 000 entities. We analyze nearly 1 million relation mentions after extracting. The KBP shared task requires to extract all latent candidates from the corpus when the relation and the first entity are known. We use 200 queries from the 2010 and 2011 evaluation during the experiments, where 40 queries are used as development set and the rest are used as testing set. We adopt the adapted KBP scorer as Surdeanu et al. [10]

6.2 Implementation Detail

In the implementation of *Clustered DS*, we apply the *MiniBatchKmeans* [18] clustering from the open *scikit-learn* package [19] rather than Kmeans clustering for its much faster convergence speed. Our method generates a more precise training dataset, which allows us to incorporate previous DS methods as improved classifier. In this paper, we implement our method with serveral previous DS methods, including: 1) *Mintz++* [5], an improved implementation of DS as a strong baseline. 2) *MultiR* [9](denoted as *Hoffmann*), a multi-instance multi-label algorithm and 3) *MIML* [10], another multi-instance multi-label algorithm but trained with Bayesian framework. We then compare our implementation with the incorporated DS methods in the real world dataset.

System	DS				Clustered DS			
	P	R	F1	Time	P	R	F1	Time
Hoffmann	30.65	19.79	24.05	1h	25.15(±2.95)	30.15(±2.86)	27.03(±1.09)	1.1h
Mintz++	28.60	23.78	25.97	3h	28.04(±2.08)	27.32(±1.57)	27.59(±1.55)	3.1h
MIML	28.00	28.30	28.15	6h	**30.94(±2.83)**	33.51(±3.63)	31.66(±2.14)	6.1h
Angeli	29.40	**35.07**	**31.99**	6.3h				

Table 1. A summary of the results testing over KBP dataset. The bold items denote the best performance among all systems.

6.3 Results

6.3.1 Improve Distant Supervision Performance. In this section, we summarize all of the results in Table 1. It shows that when using the training data generated by *Clustered DS*, both Mintz++, Hoffmann and MIML yield better performance. When employing MIML as the classifier, *Clustered DS* achieves the best performance, which is comparable to the result generated by the MIML model with an external set of labeled data (Angeli) [13]. Though we can not incorporate Angeli in our method for its published model is already trained with DS, it is still reasonable to believe that we can use an external set of labeled data to improve our method as Angeli. Table 1 also shows that the time consumption increased by *Clustered DS* is negligible when compared with the training of classifier.

In Fig.3, we show the P/R curves of our implementations compared with their incorporated DS methods. When we employ Mintz++ as the classifier of

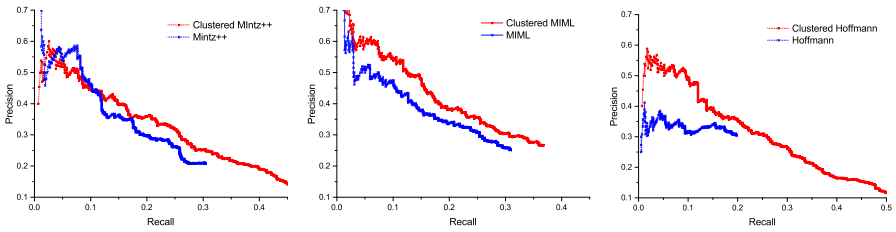

Fig. 3. The result of *Clustered DS* with DS methods incorporated as classifier which is shown in red curve. We also compare our implementations with the incorporated method which is shown in blue curve.

our method, the overall performance shows nearly 2% F1 increase comparing with the original method. Mintz++ is the original DS method in relation extraction, which can be deemed as a basic classifier. It appears that our method can significantly reduce wrong labels in training data introduced by the intuition of DS. Furthermore, for other improved DS methods such as Hoffmann and MIML, which derive a more complex classify model to tolerate noisy inputs, the system of *Clustered DS* can yield an improvement of 3.5% F1. The reason why

improved DS methods benefit more is that our method introduces new wrong labels due to an unsuccessful division. Thus those methods are more robust to leverage our training data when they can tolerate wrong labels.

6.3.2 Cluster Number Setting. The Kmeans clustering adopted in our method is highly sensitive to the cluster number k. However, we can set k much higher than the number of known relations. Our method still performs well. Here we implement a series of *Clustered DS* with employing Mintz++ as classifier and set k to {50, 100, 150, 200, 250}, for the number of known relation in the KBP dataset is 42. The context of implementations is guaranteed to be the same except k. The performance is displayed in Fig.4 with Mintz++ as the baseline. As

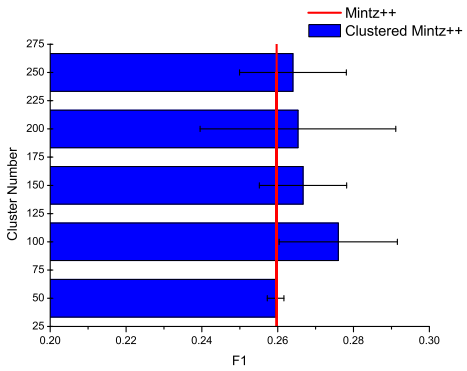

Fig. 4. The performance of a series of Clustered DS with different cluster number.

shown in Fig.4, we note that when k is set above the number of known relations, our method reaches the best performance around twice of the number of known relations and provides a consistent performance when k keeps growing. There is a slow decline when the cluster number keeps growing, due to the generation of more small clusters with a higher k. Nevertheless, our method still outperforms Mintz++ for we can aggregate small clusters in the stage of relation assignment.

7 Conclusion

Traditional distant supervision methods in relation extraction face challenges of false positives and false negatives due to the invalid intuition base. In this paper, we proposed a novel two-stage method called *Clustered DS* to tackle with these two chanlleges. By using our improved intuition of distant supervision, we adopted an unsupervised method to generate a more precise division, resulting in a considerable decrease of wrong labels in training data. We showed that *Clustered DS* can improve the distant supervision methods, with an increment of 3% in average F-score on the KBP dataset. Empirically, we noted that *Clustered DS* performs well as long as the cluster number is set above the number of known

relations. In the future, we plan to combine our method with other appoaches to further improve the performance.

References

1. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In: ISMB. (1999)
2. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: PKDD. (2010)
3. Bollacker, K.D., Cook, R.P., Tufts, P.: Freebase: A shared database of structured general human knowledge. In: AAAI. (2007)
4. Min, B., Grishman, R., Wan, L., Wang, C., Gondek, D.: Distant supervision for relation extraction with an incomplete knowledge base. In: NAACL. (2013)
5. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: ACL. (2009)
6. Bunescu, R.C., Mooney, R.J.: Learning to extract relations from the web using minimal supervision. In: ACL. (2007)
7. Wu, F., Weld, D.S.: Autonomously semantifying wikipedia. In: CIKM. (2007)
8. Nguyen, T.V.T., Moschitti, A.: End-to-end relation extraction using distant supervision from external semantic repositories. In: ACL. (2011)
9. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L.S., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: ACL. (2011)
10. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: EMNLP. (2012)
11. Takamatsu, S., Sato, I., Nakagawa, H.: Reducing wrong labels in distant supervision for relation extraction. In: ACL. (2012)
12. Pershina, M., Min, B., Xu, W., Grishman, R.: Infusion of labeled data into distant supervision for relation extraction. In: ACL. (2014)
13. Angeli, G., Tibshirani, J., Wu, J., Manning, C.D.: Combining distant and partial supervision for relation extraction. In: EMNLP. (2014)
14. Min, B., Li, X., Grishman, R., Sun, A.: New york university 2012 system for kbp slot filling. In: TAC. (2011)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(2579-2605) (2008) 85
16. Surdeanu, M., Gupta, S., Bauer, J., McClosky, D., Chang, A.X., Spitkovsky, V.I., Manning, C.D.: Stanford's distantly-supervised slot-filling system. In: TAC. (2011)
17. Aldous, D.J.: Exchangeability and related topics. Springer (1985)
18. Sculley, D.: Web-scale k-means clustering. In: WWW. (2010)
19. Michel, V., Pedregosa, F., Passos, A., VanderPlas, J., Weiss, R., Dubourg, V., Duchesnay, E., Grisel, O., Cournapeau, D., Blondel, M., Varoquaux, G., Prettenhofer, P., Thirion, B., Perrot, M., Gramfort, A., Brucher, M.: Scikit-learn: Machine learning in python. CoRR **abs/1201.0490** (2011)