Distant Supervision for Relation Extraction via Sparse Representation

Daojian Zeng[1], Siwei Lai[1], Xuepeng Wang[1], Kang Liu[1], Jun Zhao[1], and Xueqiang Lv[2]

[1] National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
{djzeng,swlai,xpwang,kliu,jzhao}@nlpr.ia.ac.cn
[2] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research
Beijing Information Science & Technology University
lxq@bistu.edu.cn

Abstract. In relation extraction, *distant supervision* is proposed to automatically generate a large amount of labeled data. *Distant supervision* heuristically aligns the given knowledge base to free text and consider the alignment as labeled data. This procedure is effective to get training data. However, this heuristically label procedure is confronted with wrong labels. Thus, the extracted features are noisy and cause poor extraction performance. In this paper, we exploit the sparse representation to address the noise feature problem. Given a new test feature vector, we first compute its sparse linear combination of all the training features. To reduce the influence of noise features, a noise term is adopted in the procedure of finding the sparse solution. Then, the residuals to each class are computed. Finally, we classify the test sample by assigning it to the object class that has minimal residual. Experimental results demonstrate that the noise term is effective to noise features and our approach significantly outperforms the state-of-the-art methods.

1 Introduction

Relation extraction refers to the task of predicting semantic relations between entities expressed in text, e.g. to detect an */business/company/founders* relation between the company *WhatsApp* and the person *Jan Koum* from the following text: *WhatsApp Inc. was founded in 2009 by Americans Brian Acton and Jan Koum (also the CEO), both former employees of Yahoo.* Most approaches to relation extraction use supervised paradigm, which achieve high precision and recall [1, 20, 22]. Unfortunately, fully supervised methods are limited by the availability of training data and cannot satisfy the demands of extracting thousands of relations with explosion of Web text.

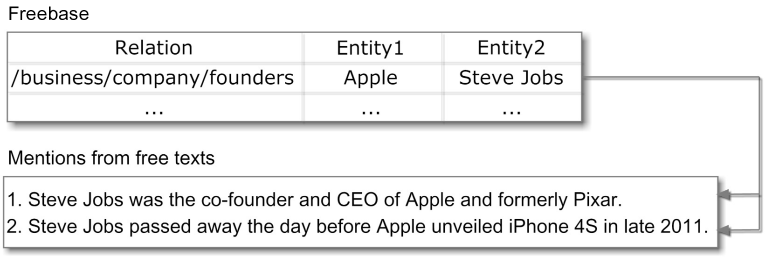
Although it lacks of off-the-shelf explicitly labeled data, an abundance of knowledge bases exist, such as DBpedia[1], YAGO[2] and Freebase[3]. To address the

[1] http://dbpedia.org/

[2] http://www.mpi-inf.mpg.de/yago-naga/yago/

[3] http://www.freebase.com

issue of lacking a large amount of labeled data, a particularly attractive approach to relation extraction is based on *distant supervision* [12]. The *distant supervision* assumes that if two entities participate in a relation of a known knowledge base, all of the sentences that mention these two entities express that relation in some way. Thus, *distant supervision* heuristically align the given knowledge base to free text and consider the alignment as labeled data. An example accounting for the training instances generated through *distant supervision* is illustrated in Figure 1. The entity pairs $\langle Apple,\ SteveJobs \rangle$ participate in a known Freebase relation. The relation mentions 1 and 2 are selected as training instances. In succession, we usually extract diverse lexical and syntactic features from all of the mentions and combine them into a richer feature vector [12], which subsequently fed to a classifier. This procedure is effective to get training examlpes. However, it is confronted with noise features. For instance, the mention 2 does not express the corresponding relation, but selected as an training instances as well in Figure 1. Therefore, features extracted from this mention are noisy. This analogous case is widespread in relation extraction based on *distant supervision*.



Freebase

Relation	Entity1	Entity2
/business/company/founders	Apple	Steve Jobs
...	...	...

Mentions from free texts

1. Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
2. Steve Jobs passed away the day before Apple unveiled iPhone 4S in late 2011.

Fig. 1. Training instances generated through distant supervision.

In this paper, we exploit the sparse representation to address the noise feature problem mentioned above. The rationale of this method is that if sufficient training samples are available from each class, it will be possible to represent the test samples as a linear combination of just those training samples from the same class [18]. In relation extraction, given a new test feature vector, we first compute its sparse representation of all the features extracted from training samples. It may not be possible to express the test instance exactly as a sparse superposition of the training samples due to the noise features. To reduce the influence of noise feature, a noise term is adopted in the procedure of finding the sparse solution. Then, the residuals to each class are computed. Finally, we classify the test sample by assigning it to the object class that has minimal residual. To the best of our knowledge, this work is the first trial to apply this technique on relation extraction based on *distant supervision*. It is not at all clear that the sparse representation should have relevance to the noise features. Nevertheless, our experiments demonstrate that the sparse representation can be quite effective to alleviate the noise feature problem. Moreover, our noise-tolerant approaches significantly outperform the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 reviews previous works in the area of relation extraction. Section 3 gives a detailed description of relation extraction via sparse representation. The experimental results are presented in Section 4. Finally, there is a conclusion in Section 5.

2 Related Work

Relation extraction is an important topics in Natural Language Processing (NLP). Many approaches have been explored for relation extraction, such as bootstrapping, unsupervised relation discovery and supervised method. Researchers have proposed various features to identify the relations between two entities by using different methods.

In bootstrapping and unsupervised paradigms, contextual features are used. The Distributional Hypothesis [6] theory indicates that words occur in the same context tend to have similar meanings. Accordingly, it is assumed that the entity pairs occurring in similar context tend to have similar relations. Hasegawa *et al.* [7] adopted a hierarchical clustering method to cluster the contexts of entity pairs and simply selected the most frequent words in the contexts to represent the relation that held between the entities. Chen *et al.* [3] proposed a novel unsupervised method based on model order selection and discriminative label identification to deal with this problem.

In the supervised paradigm, relation extraction is usually considered as a multi classification problem and researchers concentrate on extracting more complex features. Generally, these methods can be categorized into two types: feature-based and kernel-based. In feature-based methods, a diverse set of strategies have been exploited to convert the classification clues in structures such as sequences, parse trees into feature vectors [11, 15]. Feature-based methods suffer from the problem of selecting a suitable feature-set when converting the structured representation into feature vectors. Kernel methods provide a natural alternative to exploit rich representation of the input classification clues like syntactic parse trees, etc. Kernel methods allow the usage of a large set of features without explicitly extracting them. So far, various kernels have been proposed to solve relation extraction, such as convolution tree kernel [13], subsequence kernel [1] and dependency tree kernel [2].

The methods mentioned above, however, suffered from lacking of a large amount of labeled data for training. To address this problem, Mintz *et al.* [12] adopted Freebase, a large-scale knowledge base which contains billions of relation instances, to distantly supervise Wikipedia corpus. The *distant supervision* paradigm selected the sentences that matched the facts in knowledge base as positive examples. As mentioned in section 1, this training data generating algorithm sometimes exposed to wrong label problem and brought noise labeled data. To address the shortcoming, Riedel *et al.* [14] and Hoffmann *et al.* [8] cast the relaxed *distant supervision* assumption as multi-instance learning. In addition, Surdeanu *et al.*[16] proposed a novel approach to multi-instance multi-label learning for relation extraction, which can model all of the sentences in texts and all of the labels in knowledge bases. Takamatsu *et al.* [17] pointed out the relaxed assumption would fail and proposed a novel generative model to model the heuristic labeling process in order to reduce the wrong labels. Zhang *et al.* [21] analyzed some critical factors in *distant supervision* which have great impact on the accuracy to improve the performance. These previous studies mainly pay attention to errors generated in the procedure of *distant supervision.* In contrast, our work alternatively resolves the noise features and exploit the sparse representation to solve the problem.

3 Methodology

In this paper, we apply sparse representation with convex optimization for *distant supervision* relation extraction. Sparse representations are representations that account for most or all of the information of a signal with a linear combination of a small number of elementary signals called atoms [9]. Often, the atoms are chosen from an over-complete dictionary. This technology has been successfully applied on many active research areas, such as computer vision [19] and speech signal processing [10]. Our models for relation extraction are based on the theoretic framework proposed by Wright *et al.* [19], which reveals that occlusion and corruption in face recognition can be handled uniformly and robustly within the sparse representation classification framework. In distant supervision relation extraction, the noise features are analogous to the occlusion and corruption in face recognition. The feature vectors of all the train instances are selected as the dictionary and the classifier enhances the robustness to noise feature by adopting noise term in the procedure of finding the sparse solution.

3.1 Problem Statement

The problem in *distant supervision* relation extraction is to use labeled training samples from k distinct classes to correctly determine the class to which a new test sample belongs. The distant supervision assumption is that if two entities participate in a relation, all of the sentences that mention these two entities might express that relation. To leverage the information from different mentions, the features for identical tuples from different sentences are usually combined, creating a richer feature vector [12]. Let $\mathbf{v} \in \mathbb{R}^m$ denote the richer feature vector. The given n_i training samples from the ith class can be represented as the columns of a matrix $\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \cdots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$. The entire training set is denoted as the concatenation of the n training samples of all k classes: $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_k] \in \mathbb{R}^{m \times n}$ $(n = \sum\limits_{i=1}^{k} n_i)$. The *distant supervision* relation extraction is formulated as given matrix $\mathbf{A}$ to determine the class to which the richer feature vector $\mathbf{y} \in \mathbb{R}^m$ of a test sample belongs.

3.2 Sparse Linear Combination

So far, a variety of statistical, generative and discriminative methods have been proposed for exploiting the structure of the matrix $\mathbf{A}$ for classification. One particularly simple and effective method is to consider the samples from a single class as a linear combination of the subspace. In relation extraction, we select the training examples as the subspace. Given sufficient training samples of the i-th class, $\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \cdots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$, a test sample $\mathbf{y} \in \mathbb{R}^m$ can be represented as follows:

$$\mathbf{y} = \mathbf{v}_{i,1} w_{i,1} + \mathbf{v}_{i,2} w_{i,2} + \cdots + \mathbf{v}_{i,n_i} w_{i,n_i}, \tag{1}$$

where $w_{i,j} \in \mathbb{R}$ is the weight for the j-th training sample $\mathbf{v}_{i,j}$ associated with class i.

The class of the test sample $\mathbf{y}$ is unknown before the linear combination. Therefore, we select the matrix $\mathbf{A}$ of all the training samples as a dictionary and the test sample is further represented as a sparse linear combination of matrix $\mathbf{A}$.

$$\mathbf{y} = \mathbf{A}\mathbf{w} \in \mathbb{R}^m, \tag{2}$$

where $\mathbf{w} = [0, \cdots, 0, w_{i,1}, w_{i,2}, \cdots, w_{i,n_i}, 0, \cdots, 0] \in \mathbb{R}^n$ is the weight for all of the training samples and the entries are zero except for those associated with the i-th class.

As the entries of the vector $\mathbf{w}$ encode the proportion of the test sample, the relation extraction is converted to get optimal $\mathbf{w}$ in equation (2) when given $\mathbf{y}$ and $\mathbf{A}$. Obviously, this equation is overdetermined when $m >$ n and we can get its unique solution. In relation extraction, various syntactic and semantic features are usually exploited and the feature dimension is very high. In addition, $\mathbf{y} = \mathbf{A}\mathbf{w}$ may not be perfectly satisfied in the presence of noise features. Thus, equation is underdetermined and the solution $\mathbf{w}$ is not unique. At first glance, it seems impossible to get the solution of w in this case. However, we can get the following observations: a valid test sample $\mathbf{y}$ can be represented using only the training samples from the same class. This representation is naturally sparse if the number of the relation classes is reasonably large. Thus, to get the optimal $\mathbf{w}$, we transform the problem to find the sparsest solution to $\mathbf{y} = \mathbf{A}\mathbf{w}$, solving the following optimization problem:

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\|_0 \qquad \textit{subject to} \quad \mathbf{A}\mathbf{w} = \mathbf{y}, \tag{3}$$

where $\| \cdot \|$ denotes the ℓ^0-norm, which means to find the solution vector that has zero entries as much as possible. However, this ℓ^0-minimization problem is NP-hard and difficult even to get an approximate solution. Recent work in the sparse representation prove that if the solution w is sparse enough, ℓ^0-norm can be replaced by ℓ^1-norm [5]. We further transform the problem to ℓ^1-minimization problem:

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\|_1 \qquad \textit{subject to} \quad \mathbf{A}\mathbf{w} = \mathbf{y}, \tag{4}$$

This is a convex optimization problem and can be solved in polynomial time by standard linear programming methods [4].

3.3 Dealing with Noise Features

In the above section, we assumed that the test feature vectors are exactly represented as a sparse linear combination of all the training feature vectors. Since the features are noisy in *distant supervision*, the exact representation is not reasonable and we cannot assume that $\mathbf{Aw}$ is known with arbitrary precision. To tolerate the noise entries in the feature vector, a noise term is adopted in the procedure of finding the sparse solution and the equation (2) is modified to explicitly model the noise as follows:

$$\mathbf{y} = \mathbf{Aw} + \mathbf{e}, \tag{5}$$

where $\mathbf{e} \in \mathbb{R}^m$ is the added noise term. There are mainly two approach to solve the feature noise in equation (5). On the one hand, $\mathbf{e}$ is considered as a term with bounded energy $\|\mathbf{e}\|_2 < \varepsilon$. This model is called **N**oise **T**erm as **B**ounded **E**nergy(**NTBE**). Therefore, the sparse solution $\mathbf{w}$ is then approximately recovered by solving the following convex optimization problem:

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\|_1 \qquad \textit{subject to} \quad \|\mathbf{Aw} - \mathbf{y}\|_2 \leqslant \varepsilon. \tag{6}$$

Similarly, this convex optimization problem can be efficiently solved. The optimal solution subjects to the constraint that the energy of the reconstruction error of $\mathbf{y}$ is no bigger than ε.

On the other hand, $\mathbf{e}$ is considered as an error vector, some entries of which are nonzero. This model is called **N**oise **T**erm as **E**rror **V**ector (**NTEV**). The noise features may affect any entry of the feature vector and may be arbitrarily large in magnitude. In this model, the noise features are represented as a linear combination of error basis and handled uniformly through sparse representation. Then, equation (5) is replaced as:

$$\mathbf{y} = \mathbf{Aw} + \mathbf{e} = [\mathbf{A}, \mathbf{I}] \begin{bmatrix} \mathbf{w} \\ \mathbf{e} \end{bmatrix} = \mathbf{A}'\mathbf{w}' \tag{7}$$

where the matrix $\mathbf{A}$ and sparse weight w are respectively extended to $\mathbf{A}' = [\mathbf{A}, \mathbf{I}] \in \mathbb{R}^{m\times(n+m)}$ and $\mathbf{w}' = \begin{bmatrix} \mathbf{w} \\ \mathbf{e} \end{bmatrix} \in \mathbb{R}^{n+m}$. Compared to the first approach, this method can explicitly model the location of feature corruption. The nonzero entries of $\mathbf{e}$ indicate corrupted dimensions of the feature vector. Apparently, equation (8) is underdetermined and does not have a unique solution for $\mathbf{w}'$. Similarly to equation (4), we attempt to get the the sparsest solution $\mathbf{w}'$ from solving the following extended constrained ℓ^1 minimization problem:

$$\hat{\mathbf{w}'} = \arg\min \|\mathbf{w}'\|_1 \qquad \textit{subject to} \quad \mathbf{A}'\mathbf{w}' = \mathbf{y}. \tag{8}$$

3.4 Relation Classification

This section introduces how to get class label of test samples based on the results of sparse representation. Given a new test sample vector $\mathbf{y}$, we first compute its sparse representation $\hat{\mathbf{w}}$ via equation (6) or (8). Ideally, the entries in the estimate $\hat{\mathbf{w}}$ will be zero except for the entries that associated with the target class. However, the feature noise and modeling error may lead to small nonzero entries for multiple classes. To resovle this problem, it usually classifies $\mathbf{y}$ based on how well the coefficients associated with all of the training samples of each class reproduce $\hat{\mathbf{y}}_i$. The test sample is classified according to the reconstruction error between $\mathbf{y}$ and its approximations.

For class i, $\tau_i(\mathbf{w})$ represents a new vector whose only nonzero entries are the entries in $\mathbf{w}$ that are associated with class i. We can approximately reproduce $\mathbf{y}$ in class i as $\hat{\mathbf{y}}_i = \mathbf{A}\tau_i(\hat{\mathbf{w}})$. Then the classifier assign $\mathbf{y}$ to the class that minimizes the residual between $\mathbf{y}$ and $\hat{\mathbf{y}}_i$. If we use **NTBE** to model the noise, we can predict the relation class as follows:

$$\arg\min_i r_i(\mathbf{y}) = \|\mathbf{y} - \hat{\mathbf{y}}_i\|_2 = \|\mathbf{y} - A\tau_i(\hat{\mathbf{w}})\|_2 \tag{9}$$

If use **NTEV** to model the noise, we approximately get $\hat{\mathbf{w}'} = \begin{bmatrix} \hat{\mathbf{w}} \\ \hat{\mathbf{e}} \end{bmatrix}$. $\hat{\mathbf{e}}$ represents the estimated noise of the test sample feature vector $\mathbf{y}$ and $\mathbf{y} - \hat{\mathbf{e}}$ recovers the clean feature vector. To get the class label, we slightly modify the residual $r_i(\mathbf{y})$ and compute as follows:

$$\arg\min_i r_i(\mathbf{y}) = \|\mathbf{y} - \hat{\mathbf{y}}_i\|_2 = \|\mathbf{y} - \hat{\mathbf{e}} - A\tau_i(\hat{\mathbf{w}})\|_2 \tag{10}$$

Furthermore, the residual is interpreted as a conditional probability by applying a softmax operation:

$$p(i|\mathbf{y}, \mathbf{A}) = \frac{e^{1-r_i(\mathbf{y})}}{\sum\limits_{m=1}^{k} e^{1-r_m(\mathbf{y})}} \tag{11}$$

4 Experiments

To evaluate the performance of our proposed approach, we conduct three sets of experiments. The first is to test **NTBE** with different noise term ε, to gain some understanding of how the choice of noise term impacts upon the performance. In the second set of experiments, we make comparison of the performance among our method and other three kinds of landmark methods [12, 8, 16] using held-out evaluation. As the held-out evaluation is confronted with false negative problem, the goal of the third one is to evaluate the extracted results manually.

4.1 Dataset and Experiment Settings

Dataset: We select a real world dataset[4], NYT'10, to evaluate our method. The dataset was developed by Riedel *et al.* [14] and also used by Hoffmann *et al.* [8].

[4] http://iesl.cs.umass.edu/riedel/ecml/

In the dataset, Freebase was used as the *distant supervision* source and the New York Times (NYT) was selected as the text corpus. Four categories of Freebase relations are used, including "people", "business", "person" and "location". The NYT data contains over 1.8 million articles written and published between January 1, 1987 and June 19, 2007. The Freebase relations were divided into two parts, one for training and one for testing. The former is aligned to the years 2005-2006 of the NYT corpus, the latter to the year 2007. As we need negative examples for training, this dataset generally pick 10% of the entity pairs that appear in the same sentence but are not related according to Freebase. Moreover, three kinds of features, namely, lexical, syntactic and named entity tag features, were extracted from relation mentions. The statistics about the dataset is presented in Table 1. There are 51 relationships and an *NA* class.

# of relation labels	# of training examples	# of NA training examples	# of testing examples	# of NA testing examples	# of features
51	4,700	63,596	1,950	94,917	1,071,684

Table 1. Statistics about the NYT'10 dataset.

Experiment Settings: As the number of *NA* training examples is too large, we randomly choose 1% *NA* training examples as the negative examples in the following experiments. Table 1 presents that the number of features exceeds one million. Using all of this features will lead to the very high dimension of feature vector and data sparsity. To control the feature sparsity degree, Surdeanu *et al.* [16] released the source code[5] to reproduce their experiments and use a threshold θ to filter the features that appears less than θ times. They set $\theta = 5$ in the original code by default. To guarantee the fair comparison for all of the methods, we follow their settings and adopt the same way to filter the features in our experiments. In the sparse representation-based relation extraction, the main problem is to solve the ℓ^1-minimization problem. In this paper, we use SPGL1[6] to solve this problem and set the parameters as default.

4.2 The Effect of Noise Term

In the **NTBE** model, an extra input parameter ε is needed. This parameter determines the tolerance to the noise features. The optimal value of ε varies with different datasets. In this part, we experimentally study the effect of the noise term ε in our proposed method. Since there is no development dataset in NYT'10, we tuned the noise term ε by trying different value via five-fold cross-validation. Figure 2 illustrates the curves of F1 score for each fold. $\varepsilon = 0$ means that the test feature vector are exactly represent as a sparse linear combination of all the training feature vectors. From Figure 2, we can observe a phenomenon

[5] http://nlp.stanford.edu/software/mimlre.shtml

[6] http://www.cs.ubc.ca/ mpf/spgl1/

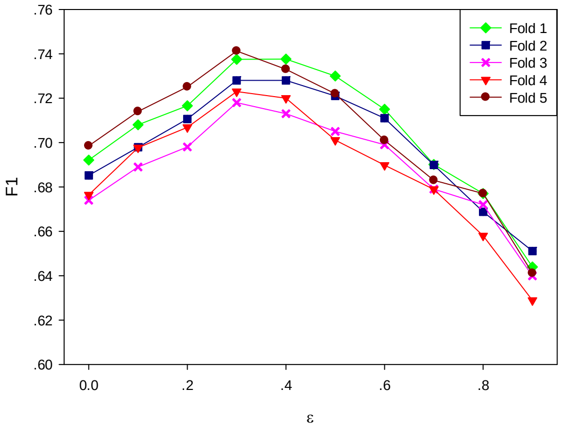


Fig. 2. Five-fold cross validation for the error energy ε.

that the performance gradually increases as the error energy ε increases before reaching the optimum. However, it sharply decreases if we continue increasing the optimal error energy. We get the optimal results when $\varepsilon = 0.3$. An intuitive explanation can be accounted for this phenomenon: The feature vector contains much noise when the error energy is very small and the model tends to be overfitting. Whereas the feature vector is likely to lose principal information and the model tends to be underfitting when the error energy is excessively large.

4.3 Held-out Evaluation

In the held-out evaluation, half of the Freebase relations were divided for testing. The relation instances discovered from testing articles were automatically compared with those in Freebase. Held-out evaluation gives a rough measure of precision without requiring expensive human evaluation. To get the final performance of our proposed method, we select three approaches as competitors to be compared with our method in Figure 3. *Mintz* represents the baseline distant supervision model for relation extraction proposed by Mintz *et al.* [12]. *MultiR* is a state-of-the-art multi instance learning system proposed by Hoffmann *et al.* [8]. *MIML* indicates the multi-instance multi-label learning system [16], which jointly models all of the instances of a pair of entities in text and all of their labels. We compare these three approached with our proposed sparse representation based methods. In *NTBE*, the error energy is set to the optimal value $\varepsilon = 0.3$. For *MultiR* and *MIML*, we used the authors' implementation. We re-implemented Mintz's algorithm.

From Figure 3, we have the following observations. *NTBE* outperforms the state-of-the-art over the whole precision-recall curve. *NTEV* achieves the best

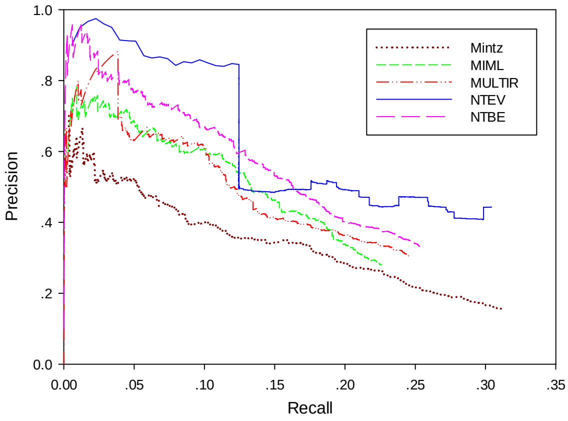


Fig. 3. Precision and recall for the held out evaluation.

results except that the recall is between 0.12 and 0.17. The precision-recall curve of *NTEV* has a sharp decline when the recall is about 0.12. The reason for this phenomenon is that the number of *NA* testing examples is far more than the number of testing examples. By analyzing the results, we find that a large number of the *NA* testing examples have conditional probabilities greater than the testing examples. The precision will declines gradually while the recall rate remains unchange. Therefore, there has been a sharp decline in the precision-recall curve. The similar phenomenon can be observed in *MultiR* when the recall rate is about 0.03. In the held out evaluation, we compare newly discovered relation instances to the held out Freebase. As the incompleteness of Freebase, held out evaluation suffers from false negatives and the precision is underestimated. We address this problem through manual evaluation in the next section.

4.4 Manual Evaluation

Top-N	Mintz	MultiR	MIML	NTBE	NTEV
Top-100	0.77	0.83	0.88	0.89	**0.91**
Top-200	0.71	0.74	0.81	0.82	**0.85**
Top-500	0.55	0.59	0.63	0.68	**0.72**
Average	0.676	0.720	0.773	0.796	**0.826**

Table 2. Precision of the Top-100, Top-200, Top-500 for manual evaluation.

For the manual evaluation, we choose those entity pairs which at least one participating entity is not in Freebase. Since the number of the relation instances expressed in the test data is unknown, we cannot calculate recall in this case. Alternatively, we calculate the precision of the Top-N extracted relation instances. Table 2 presents the manually evaluated precisions for the Top-100, Top-200, Top-500, from which we can see that both of our methods outperform all of the compared methods. *NTEV* achieves the best performance. We also perform one-tailed t-test ($p \leqslant 0.05$) which demonstrates that our method significantly outperforms all of the baselines.

5 Conclusion

The *distant supervision* assumes that every sentence that mentions two related entities in a knowledge base express the corresponding relation. The *distant supervision* assumption can fail, which results in noise feature problem and causes poor extraction performance. In this paper, we exploit sparse representation for *distant supervision* relation extraction. To tolerance the noise features, a noise term is adopted in our model. Two models, **NTBE** and **NTEV**, are adopted to model the noise. Experiment results demonstrate that the sparse representation is quite effective to alleviate the noise feature problem.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61202329, 61272332, 61333018), CCF-Tencent Open Research Fund and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research(ICDD201301). We thank the anonymous reviewers for their insightful comments.

References

1. Bunescu, R., Mooney, R.: Subsequence kernels for relation extraction. Advances in neural information processing systems 18, 171 (2006)
2. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 724–731 (2005)
3. Chen, J., Ji, D., Tan, C.L., Niu, Z.: Unsupervised feature selection for relation extraction. In: Proceedings of IJCNLP (2005)
4. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. 43(1), 129–159 (Jan 2001)
5. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ1-norm solution is also the sparsest solution. Comm. Pure and Applied Math. Math 59, 797–829 (2006)
6. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
7. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (2004)

8. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 541–550 (2011)
9. Huang, K., Aviyente, S.: Sparse representation for signal classification. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19, pp. 609–616 (2006)
10. Jafari, M.G., Plumbley, M.D.: Fast dictionary learning for sparse representations of speech signals. J. Sel. Topics Signal Processing 5(5), 1025–1031 (2011)
11. Kambhatla, N.: Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (2004)
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: Volume 2 - Volume 2. pp. 1003–1011 (2009)
13. Qian, L., Zhou, G., Kong, F., Zhu, Q., Qian, P.: Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In: Proceedings of the 22nd International Conference on Computational Linguistics. pp. 697–704 (August 2008)
14. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III. pp. 148–163 (2010)
15. Suchanek, F.M., Ifrim, G., Weikum, G.: Combining linguistic and statistical analysis to extract relations from web documents. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 712–717 (2006)
16. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multilabel learning for relation extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 455–465 (2012)
17. Takamatsu, S., Sato, I., Nakagawa, H.: Reducing wrong labels in distant supervision for relation extraction. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. pp. 721–729 (2012)
18. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(2), 210–227 (Feb 2009)
19. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S.: Sparse representation for computer vision and pattern recognition. Proceedings of the IEEE 98(6), 1031–1044 (2010)
20. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. The Journal of Machine Learning Research 3, 1083–1106 (2003)
21. Zhang, X., Zhang, J., Zeng, J., Yan, J., Chen, Z., Sui, Z.: Towards accurate distant supervision for relational facts extraction. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. pp. 810–815 (2013)
22. Zhou, G., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 427–434 (2005)