Noise-Clustered Distant Supervision for Relation Extraction: A Nonparametric Bayesian Perspective

Qing Zhang and **Houfeng Wang**
Key Laboratory of Computational Linguistics (Peking University) Ministry of Education, China
{zqicl,wanghf}@pku.edu.cn

Abstract

For the task of relation extraction, distant supervision is an efficient approach to generate labeled data by aligning knowledge base with free texts. The essence of it is a challenging incomplete multi-label classification problem with sparse and noisy features. To address the challenge, this work presents a novel nonparametric Bayesian formulation for the task. Experiment results show substantially higher top-precision improvements over the traditional state-of-the-art approaches.

1 Introduction

To efficiently generate structured relation information from free texts, the research on distantly supervised Relation Extraction (RE) (Mintz et al., 2009; Riedel et al., 2013; Hoffmann et al., 2011) has been attracting much attention, because it can greatly reduce the manual annotation for training. It essentially based on the assumption that the relation between two entities in a Knowledge Base (KB), is also likely hold within a sentence that mentions the two entities in free texts. This assumption plays a crucial role in distant supervision, which is quite effective in real applications.

However, the assumption of distant alignment can also lead to the noisy training corpus problem (Fan et al., 2014), which is challenging for the task as follows: **i) Noisy features.** Not all relations existed in a KB keep the same meaning of that relation for the corresponding entities in a free text. For example, the second relation mention in Figure 1 does not explicitly describe any relation instance, so features extracted from this sentence can be noisy. Such analogous cases commonly exist in feature extraction. **ii) Incomplete labels.** Similar to noisy features, the generated label can be incomplete due to the incomplete knowledge base (Ritter et al., 2013). For example, the fourth relation mention in Figure 1 should be labeled by the relation Senate-of. However, the corresponding relation instance (Senate-of(Barack Obama, U.S.)) is missing in the knowledge base. Such analogous cases are also common in real applications. **iii) Sparse features.** Sophisticated features extracted from the mentions can result in a large number of sparse features (Riedel et al., 2013). The generalization ability of feature based prediction models will be badly hurt, when the features do not match between testing and training.

Entity pair	<Barack Obama,U.S.>
Relation instances from knowledge bases	1.President of (Barack Obama,U.S.) 2.Born in (Barack Obama,U.S.)
Relation mentions from free texts	1.**Barack Obama** is the 44th and current President of the **U.S.**(**President of**) 2.**Barack Obama** ended **U.S.**military involvement in the Iraq War.(-) 3.**Barack Obama** was born in honolulu, Hawaii, **U.S.**(**Born in**) 4.**Barack Obama** ran for the **U.S.**Senate in 2004.(**Senate of**)

Figure 1: Aligned Example (Fan et al., 2014): the relation instances related to the entity pair $\langle BarackObama, U.S.\rangle$ in the KB, and its mentions in the free text.

To tackle the problem, we develop a novel distant supervision approach from a nonparametric Bayesian perspective (Blei et al., 2016), along with the previously most effective research line (Petroni et al., 2015) of using matrix completion (Fan et al., 2014) for relation extraction. Our goal is to design a noise-tolerant relation extraction model for distantly supervised corpus with noise and sparsity problems. Different from (Fan et al., 2014) as one state-of-the-art method in this line, we model noisy data corpus using adaptive variance modeling approach (Chen et al., 2015), based on *Dirichlet Process* (Blei and Jordan, 2004) instead of a fixed way of controlling complex noise weighting. To the best of our knowledge, we are

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1808–1813
Copenhagen, Denmark, September 7–11, 2017.

the first to apply this technique on relation extraction with distant supervision.

2 Approach

The essence of the task is a multi-label classification problem (Cabral et al., 2011) with noisy patterns (Han and Sun, 2014). One simple way, to solve the problem, is to learn separate classifiers for each of relation labels, using n samples with d features, by optimizing $b \in R^{1\times 1}$ and $\mathbf{w} \in R^{d\times 1}$,

$$\text{argmin}_{b,\mathbf{w}}\ l(\mathbf{y_{train}}, [\mathbf{1}\ X_{train}] \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}), \quad (1)$$

where $\mathbf{1}$ is the all-one column vector; $X_{train} \in R^{n\times d}$ and $\mathbf{y_{train}} \in R^{n\times 1}$ are the corresponding feature matrix and label vector respectively. However, label correlations are not considered in the above formulation. To jointly consider feature correlations and label correlations, (Cabral et al., 2011) formulated the multi-label classification as a matrix completion problem. As a powerful framework, it has been successfully applied to relation extraction task with distant supervision.

2.1 Previous Formulation

The work in (Fan et al., 2014) first adopted the mentioned framework, as a general joint learning and inference framework (Cabral et al., 2011), to learn *noise-tolerant* distant supervision for relation extraction. It achieves the state-of-the-art performance. Suppose we have a training corpus, including n instances (entity pairs) including both training and test data, with d-dimensional features and t relation labels, which is built according to the basic alignment assumption. The task can be modeled with a sparse matrix $Z \in R^{n\times(d+t)}$, defined as

$$Z = \begin{bmatrix} X_{train} & Y_{train} \\ X_{test} & Y_{test} \end{bmatrix}, \quad (2)$$

where each row in Z represents entity pair, and each column represents noisy textual feature in X or incomplete relation label in Y. In such a way, relation extraction is transformed into a problem of completing the unknown labels in Y_{test} for the test data X_{test} in Z. The rational of this modeling is that noisy features and incomplete labels are semantically correlated, which can be explained in an underlying low-rank structure (Riedel et al., 2013). Taking noise into consideration, Z is further defined as

$$Z = Z^* + E, \quad (3)$$

where Z^* is the underlying low-rank matrix

$$Z^* = \begin{bmatrix} X^*_{train} & Y^*_{train} \\ X^*_{test} & Y^*_{test} \end{bmatrix}, \quad (4)$$

and E is the error (noise) matrix

$$E = \begin{bmatrix} E_{X_{train}} & E_{Y_{train}} \\ E_{X_{test}} & 0 \end{bmatrix}. \quad (5)$$

This error (noise) modeling approach has been successfully applied to distantly supervised relation extraction. However, it still has clear limitations. The noise model is limited to a single source without considering the intrinsic clustering structures of data. In addition, the true rank is usually hard to determine, for adaptively modeling the correlations among features and labels.

2.2 Nonparametric Bayesian Modeling

The use of nonparametric Bayesian modeling has been widely adopted in Natural Language Processing (NLP) (Chen et al., 2014). Instead of imposing assumptions that might be wrong, it "lets the data speak for itself", without requiring optimizing parameters blindly by hands (Blei and Jordan, 2004). To take advantage of these merits, we here adopt it for the task, with the following motivations:

Motivation 1: Adaptive Noise-Clustered Attention. The goal is to find an adaptive *cluster specific* noise *parameterization* for the complex noisy corpus, *without* making overly strong assumptions about the noise distribution in real applications.

Motivation 2: Adaptive Latent Feature Space Selection. The goal is to automatically find better *dense representations* of latent entity-pair, feature and label *without* pre-specifying the rank values by laboriously retraining models.

2.2.1 Nonparametric Bayesian Formulation

We develop a novel formulation for distantly supervised relation extraction, using a nonparametric Bayesian approach, based on the Dirichlet Process, which can been seen as an infinite Dirichlet distribution, with clustering effect for modeling categorical variables adaptively.

Noise component modeling. Instead of using a single fixed noise model, we redefine $E = [\varepsilon_{i,j}] \in R^{n\times(d+t)}$ in Eq.(5). $\varepsilon_{i,j}$ is modeled by a summation of infinite noise models (Chen et al., 2015),

$$p(\varepsilon_{i,j}) = \sum_{k=1}^{\infty} \theta_k N(\varepsilon_{i,j}|0, \sigma_k), \quad (6)$$

where θ_k is the mixing proportion for the k-th gaussian component $N(\varepsilon_{i,j}|0, \sigma_k)$ with mean zero and variance σ_k. The θ is obtained from the stick-breaking process (Blei and Jordan, 2004), with $\sum_{k=1}^{\infty} \theta_k = 1$,

$$\theta_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_l), \quad (7)$$

where β_k are independent draws from the beta distribution $\beta(1, \alpha)$. As a result, the noise entries will cluster themselves into K groups without requiring a complicated model selection procedure. Since a mixture of Gaussians can approximate any continuous probability distribution (Zhao et al., 2014), this structural noise formulation can adapt much wider range of real noises than previous formulation (Fan et al., 2014) for relation extraction.

Low-rank component modeling. Different from (Fan et al., 2014), instead of directly minimizing the rank of Z^* in Eq.(3), we decompose Z^* into two low-rank matrices U and V, from probabilistic perspective (Salakhutdinov and Mnih, 2007). This modeling approach can lead to a more flexible way of estimating the optimal rank values for latent feature spaces. To determine the appropriate rank automatically, we adopt the *Automatic Relevance Determination* (ARD) method (Babacan et al., 2012) by imposing a prior on each dimmension (column) of U and V. Specifically, we impose the Gaussian priors with variance λ_r on the r-th columns of U and V, i.e., $u_{\cdot r}$ and $v_{\cdot r}$:

$$\begin{aligned} p(\mathbf{U}|\lambda) &= \prod_{r=1}^{R} N(u_{\cdot r}|0, \lambda_r \mathbf{I}_{\mathbf{U}}), \\ p(\mathbf{V}|\lambda) &= \prod_{r=1}^{R} N(v_{\cdot r}|0, \lambda_r \mathbf{I}_{\mathbf{V}}), \\ \lambda_r &\sim IG(a_1, b_1), \end{aligned} \quad (8)$$

where IG is an Inverse Gamma distribution for modeling the variance λ_r. Considering a column as latent factor in U or V with a zero mean in the prior, a very small variance indicates that this column will shrink to zero. Thus, the irrelevant columns hurting the performance will be eliminated adaptively, without pre-specifying the rank values by retraining models laboriously as in the previous modeling (Fan et al., 2014) for the task.

Prediction component modeling. We can leverage the above presented low-rank component for U, V and noise component for $\varepsilon_{i,j}$, to build Eq.(9) for prediction. Different from the state-of-the-art multi-label classification framework as adopted in (Fan et al., 2014), for simplicity, we design noise model for features and labels jointly,

$$p(y_{i,j}) = N(y_{i,j}| \underbrace{u_{i\cdot}v_{j\cdot}^T}_{low-rank\ component}, \underbrace{\varepsilon_{i,j}}_{noise\ component}), \quad (9)$$

where $u_{i\cdot}$ and $v_{j\cdot}$ are defined in Eq.(8) as rows of U and V respectively.

For each interaction between entity-pair and feature (or relation), $\varepsilon_{i,j}$ as defined in Eq.(6) can be injected into Eq.(9) (Chen et al., 2015) by

$$\begin{aligned} \varepsilon_{i,j} &= \sigma_{z_{ij}}, \\ \sigma_{z_{ij}} &\sim IG(a_0, b_0), \\ z_{ij} &= k \sim Mult(\theta_k), \end{aligned} \quad (10)$$

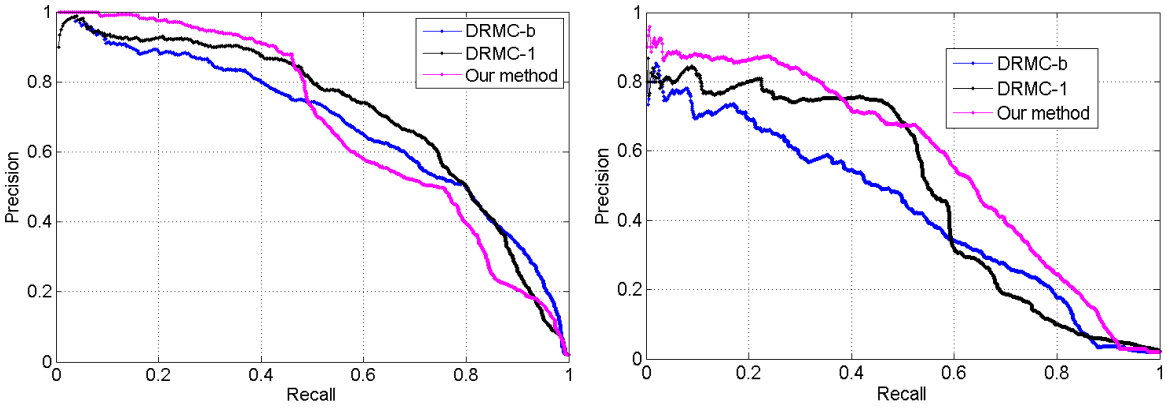
where θ_k is modeled in Eq.(7); Mult is a Multinomial distribution.

The mechanism of the introduced clustered noise component for relation extraction can be easily understood through considering its role in the Gaussian distribution. As shown in Eq.(9), $\varepsilon_{i,j}$ is used to control the variance. Large variance value means low confidence, and the small value means high confidence, for fitting $y_{i,j}$ with $u_{i\cdot}v_{j\cdot}^T$. The variance parameter $\varepsilon_{i,j}$, generated by noise component Eq.(7,10), serves as a confidence parameter for training instance. In the algebra view of likelihood, variance parameter is just the weight of training instance (i.e., the interaction between "entity pair and feature" or "entity pair and label"), measuring the importance for its contribution to the total likelihood. We can treat this mechanism as an importance weighting mechanism, for selecting noisy interactions y_{ij} with different clustering structures adaptively.

In this mechanism, for each $y_{i,j}$ in noisy corpus, it allows $1 \rightarrow 0$ (noisy feature) for features, and allow $1 \rightarrow 0$ (label with no supportive features) or $0 \rightarrow 1$ (incomplete label) for labels. In addition, for the task, we expect that our method can automatically adjust the importance weight for reducing the effect of common features, to differentiate two instances with different labels. To achieve the goal, in matrix Z, we fit both "1" (observed) and "0" for training labels as discriminative supervision, while we only fit "1" (observed) for features.

Dataset	#training	#testing	%more than one label	#features	#relation labels
NYT'10	4,700	1,950	7.5%	244,903	51
NYT'13	8,077	3,716	0%	1,957	51

Table 1: Statistics about the two widely used datasets.



(a) NYT'10 dataset (b) NYT'13 dataset

Figure 2: Precision-Recall curve on NYT'10 and NYT'13 datasets. DRMC-b(1) (Fan et al., 2014).

Models	P	R	F1
Mintz	63.59%	61.20%	62.37%
Hoffmann	67.18%	36.41%	47.23%
Surdeanu	76.23%	53.18%	62.65%
DRMC-b	61.03%	66.82%	63.79%
DRMC-1	64.17%	**71.74%**	**67.75%**
Our	**87.94%**	46.00%	63.44%

Table 2: Results at the highest F1 point in the Precision-Recall (P-R) curve on NYT'10 dataset. Mintz (Mintz et al., 2009); Hoffmann (Hoffmann et al., 2011); Surdeanu (Surdeanu et al., 2012); DRMC-b(1) (Fan et al., 2014);

Learning. To combine Eqs. (7)-(10), we can construct the full Bayesian model. The goal turns to infer the posterior of all involved variables:

$$p(\mathbf{U}, \mathbf{V}, \lambda, \sigma, \mathbf{z}, \beta | \mathbf{X}_{observed}, \mathbf{Y}_{observed}), \quad (11)$$

where $\mathbf{X}_{observed}, \mathbf{Y}_{observed}$ are the observed binary features (fitting 1) and labels (fitting both 1 and 0). Variational inference is adopted as shown in (Chen et al., 2015).

Prediction. After learning[1], we use the expectation $E(P(y_{i,j}))$ in Eq.(9) to complete the entries in Y_{test}. Finally, we can acquire Top-N predicted relations via ranking the values $E(P(y_{i,j}))$, given entity pair i, for different relations j.

[1]We implement the system for relation extraction, based on the code at http://peixianc.me/amf_codes.zip.

3 Experiments

We evaluate our method on two widely used datasets as shown in Table 1 with the same setting in (Fan et al., 2014).

Dataset. *NYT'10*, was developed by (Riedel et al., 2010). *NYT'13*, was also released by (Riedel et al., 2013), in which they only regarded the lexicalized dependency path between two entities as features. Both are automatically generated by aligning Freebase to New York Times corpus.

Parameter setting. For all the conducted experiments, the model hyperparameters are fixed without further tuning: $a_0 = b_0 = 10^{-4}$, $a_1 = b_1 = 0.1$ and $\alpha = 1$.

Model comparison. Since (Fan et al., 2014) achieves the state-of-the-art performance on the two datasets, we mainly compare our method with that in the same setting, to verify the effectiveness. *NYT'10 dataset*: Table 2 indicates that our model achieves the highest precision performance among all of the competitors. Although the recall performance is not competitive, the F1 score is also comparable to DRMC-b. Figure 2(a) further shows the strong precision performance when the recall is not large. *NYT'13 dataset*: Figure 2(b) illustrates that our approach outperforms the state-of-the-art methods, which shows that our approach can maintain a fairly high precision even when recall is larger. In addition, in practical applications, we also concern about the precision on

Top-N	NFE-13	DRMC-b	DRMC-1	Our
Top-100	62.9%	82.0%	80.0%	**92.0%**
Top-200	57.1%	77.0%	80.0%	**88.2%**
Top-500	37.2%	70.2%	77.0%	**86.3%**
Average	52.4%	76.4%	79.0%	**88.8%**

Table 3: Precision of Top-N predicted instances on NYT'13 dataset. NFE-13 (Riedel et al., 2013); DRMC-b(1) (Fan et al., 2014).

Models	P	R	F1
DRMC-b	47.70%	49.58%	48.62%
DRMC-1	**67.99%**	50.42%	57.90%
Our	66.46%	**53.30%**	**59.16%**

Table 4: Results at the highest F1 point in the Precision-Recall (P-R) curve on NYT'13 dataset. DRMC-b(1) (Fan et al., 2014).

Top-N predicted instances. Table 3 shows that our model achieves much significant improvements on that. Moreover, Table 4 shows that our method can achieve the best F1, compared with the baselines.

NYT'10 and NYT'13 have different performance records, which could be explained as follows. From the dataset perspective, NYT'10 is a dataset with multi-label instances, which is more complex than NYT'13 only having single label instances. This is one reason of why the trends are quite different between them. More essentially, we further discuss the differences from the model mechanism perspective, to explain the reasons. In (Fan et al., 2014)'s work, it has no explicit noise modeling mechanism. The noise is modeled implicitly as the error of cost functions. From the probabilistic view, that error is sampled from single Gaussian with zero mean and fixed variance. In contrast, our method uses infinite Gaussian with automatically learnt variance. It may cause overfitting for complex dataset with sparse features. In addition, we guess the reason is that in (Fan et al., 2014)'s work, they use two separate cost functions for features and labels, while in our work we use one unified noise component for both of them, which shows the promising precision performance in NYT'10 when recall is less than 0.4.

In addition, in our experiments, we found that early stopping is crucial for achieving good results while model learning. This also verifies that the potential overfitting problem should be further considered while using the more flexible nonparametric method for NLP task.

4 Related Work

Our work is closest to (Fan et al., 2014), since we focus on the same noisy corpus problem. Although from different perspectives, we study it along with the same line of using matrix factorization (Petroni et al., 2015) for relation extraction. In this line, (Riedel et al., 2013) initially considered the task as a matrix factorization problem. Their method consists of several models, such as PCA (Collins et al., 2001) and collaborative filtering (Koren, 2008). However, the data noise brought by the assumption of distant supervision (Mintz et al., 2009), is not considered in the work. Another line addressing the problem uses deep neural networks (Zeng et al., 2015; Wang et al., 2015). The difference is that it is a supervised learning approach, while our focused one is a joint learning approach with transductive style, in which both training and test data are exploited simultaneously. In addition, (Han and Sun, 2016) explored Markov logic technique to enrich supervision knowledge, which can incorporate indirect supervision globally. Our method could be further augmented by that idea, using additional logical constraint to reduce the uncertainty for the clustered noise modeling.

5 Conclusion

In this paper, building on recent advances from the nonparametric Bayesian literature, we reformulate the task of relation extraction with distant supervision, based on the adaptive variance learning with intrinsic clustering structures. For the task, it can solve the sparsity problem via the learnt low-rank dense representations and can allow fitting noisy corpus through adaptive variance adjustment. Meanwhile, it can avoid turning a large number of parameters. Experiments suggest substantially higher top-precision than the competitors. In the future work, we plan to develop more sophisticated noise models for features and labels separately, and try to explore logical information, particularly in this context of nonparametric noise modeling, for further benefiting this task.

Acknowledgments

Our work is supported by National Natural Science Foundation of China (No.61370117 & No.61433015). The corresponding author of this paper is Houfeng Wang.

References

S. Derin Babacan, Martin Luessi, Rafael Molina, and Aggelos K. Katsaggelos. 2012. Sparse bayesian methods for low-rank matrix estimation. *IEEE Transactions on Signal Processing*, 60(8):3964–3977.

David M. Blei and Michael I. Jordan. 2004. Variational methods for the dirichlet process. In *Proceedings of the International Conference on Machine Learning, Banff, Alberta, Canada*.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2016. Variational inference: A review for statisticians. *CoRR*, abs/1601.00670.

Ricardo Silveira Cabral, Fernando De la Torre, João Paulo Costeira, and Alexandre Bernardino. 2011. Matrix completion for multi-label image classification. In *Proceedings of Advances in Neural Information Processing Systems, Granada, Spain.*, pages 190–198.

Miaohong Chen, Baobao Chang, and Wenzhe Pei. 2014. A joint model for unsupervised chinese word segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar*, pages 854–863.

Peixian Chen, Naiyan Wang, Nevin L. Zhang, and Dit-Yan Yeung. 2015. Bayesian adaptive matrix factorization with automatic model selection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, MA, USA*, pages 1284–1292.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. 2001. A generalization of principal components analysis to the exponential family. In *Proceedings of Advances in Neural Information Processing Systems, British Columbia, Canada*, pages 617–624.

Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y. Chang. 2014. Distant supervision for relation extraction with matrix completion. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA*, pages 839–849.

Xianpei Han and Le Sun. 2014. Semantic consistency: A local subspace based method for distant supervised relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA*, pages 718–724.

Xianpei Han and Le Sun. 2016. Global distant supervision for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA.*, pages 2950–2956.

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

Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA*, pages 426–434.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, Singapore*, pages 1003–1011.

Fabio Petroni, Luciano Del Corro, and Rainer Gemulla. 2015. CORE: context-aware open relation extraction with factorization machines. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal*, pages 1763–1773.

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

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of Conference of the North American Chapter of the Association of Computational Linguistics, Atlanta, Georgia, USA*, pages 74–84.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *TACL*, 1:367–378.

Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic matrix factorization. In *Proceedings of Advances in Neural Information Processing Systems, British Columbia, Canada*, pages 1257–1264.

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

Zhen Wang, Baobao Chang, and Zhifang Sui. 2015. Distantly supervised neural network model for relation extraction. In *Proceedings of Chinese Computational Linguistics and Natural Language Processing, Guangzhou, China*, pages 253–266.

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

Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Lei Zhang. 2014. Robust principal component analysis with complex noise. In *Proceedings of the International Conference on Machine Learning, Beijing, China*, pages 55–63.