Relation Discovery with Out-of-Relation Knowledge Base as Supervision

Yan Liang[1], Xin Liu[1], Jianwen Zhang[2], and Yangqiu Song[1]
[1]Department of CSE, Hong Kong University of Science and Technology, HK
[2]Microsoft, USA
[1]{yliangav, xliucr, yqsong}@cse.ust.hk
[2]{jiazhan}@microsoft.com

Abstract

Unsupervised relation discovery aims to discover new relations from a given text corpus without annotated data. However, it does not consider existing human annotated knowledge bases even when they are relevant to the relations to be discovered. In this paper, we study the problem of how to use out-of-relation knowledge bases to supervise the discovery of unseen relations, where out-of-relation means that relations to discover from the text corpus and those in knowledge bases are not overlapped. We construct a set of constraints between entity pairs based on the knowledge base embedding and then incorporate constraints into the relation discovery by a variational auto-encoder based algorithm. Experiments show that our new approach can improve the state-of-the-art relation discovery performance by a large margin.

1 Introduction

Relation extraction has been widely used for many applications, such as knowledge graph construction (Dong et al., 2014), information retrieval (Liu et al., 2014), and question answering (Ravichandran and Hovy, 2002). Traditional supervised approaches require direct annotation on sentences with a relatively small number of relations (Roth and Yih, 2002; Kambhatla, 2004).[1] With the development of large-scale knowledge bases (KBs) such as Freebase (Bollacker et al., 2008), relation extraction has been extended to larger scales comparable to KBs using the distant supervision (Mintz et al., 2009). However, when the training corpus does not support the annotated relations showing in the KB, such approach could fail to find sufficient training examples. Distant supervision assumption can be violated by up to 31% for some relations when aligning to NYT corpus (Riedel et al., 2010). More importantly, either traditional supervised learning or distantly supervised learning cannot discover new relations unseen in the training phase.

Unsupervised relation discovery tries to overcome the shortcomings of supervised or distantly supervised learning approaches. Existing approaches either extract surface or syntactic patterns from sentences and use relation expressions as predicates (which result in many noisy relations) (Etzioni et al., 2004; Banko et al., 2007), or cluster the relation expressions based on the extracted triplets to form relation clusters (Yao et al., 2011, 2012; Marcheggiani and Titov, 2016). However, these approaches do not use existing high-quality and large-scale KBs when they are relevant to the relations to be discovered.

In this paper, we consider a new relation discovery problem where both the training corpus for relation clustering and a KB are available, but the relations in the training corpus and those in the KB are not overlapped. As shown in Figure 1, in the KB, we have entities `Pink Floyd`, `Animals`, etc., with some existing relations *notable_work* and *has_member* in the KB. However, when doing relation discovery, we can only get supporting sentences that suggest new relations *based_on* and *influenced_by*. This is a common and practical problem since predicates in KBs are limited to the annotator defined relations while the real relations in the world are always open and creative.

It is challenging when there is no overlapped relation between target relation clusters and the KB because in this case the KB is not a direct supervision. But if target relation clusters and the KB share some entities, we can use the shared entities as a bridge to introduce indirect supervision for the relation discovery problem. Specifically, we build constraints between pairs of tuples based on the

[1]We distinguish a relation (e.g., a predicate in a knowledge base) from the relation expression (e.g., the text surface between entities in a sentence) throughout the paper.

Proceedings of NAACL-HLT 2019, pages 3280–3290
Minneapolis, Minnesota, June 2 - June 7, 2019.

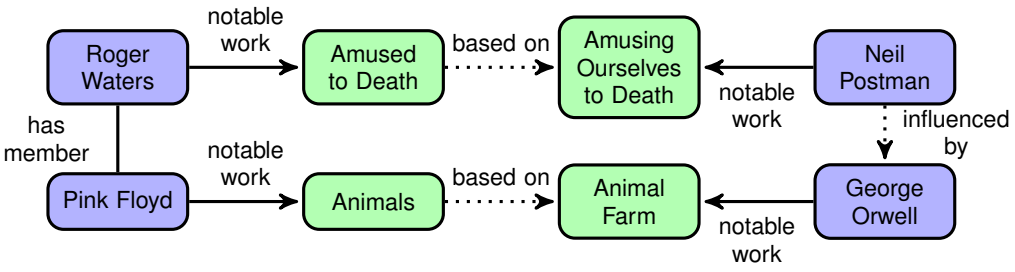


Figure 1: An illustration of our new relation discovery setting. The konwledge base contains relations *notable_work* and *has_member*. However, the training corpus to perform relation discovery only contains new relations *based_on* and *influenced_by*.

KB. For example, in Figure 1, when we cluster the *based_on* relation, we can evaluate the similarity between the tuple (`Animals`, `Animal Farm`) and the tuple (`Amused to Death`, `Amusing Ourselves to Death`) based on the KB. If the KB tells us these two pairs of tuples are close to each other, then we put a constraint to force our relation clustering algorithm to group them together.

We use the discrete-state variational autoencoder (DVAE) framework (Marcheggiani and Titov, 2016) as our base relation discovery model since this framework is flexible to incorporate different features and currently the state-of-the-art. We use KB embedding (Bordes et al., 2013) to obtain entity embeddings in the KB and use entity embeddings to evaluate the similarity between a pair of tuples. Then constraints are constructed and incorporated into the DVAE framework in a way inspired by the must-link and cannot-link based constrained clustering (Basu et al., 2004). We show that with no overlapped relations between the KB and the training corpus, we can improve the relation discovery by a large margin.

Our contributions are summarized as follows.

- We study a new prevalent but challenging task of relation discovery where the training corpus and the KB have no overlapped relation.
- We propose a new kind of indirect supervision to relation discovery which is built based on pairwise constraints between two tuples.
- We show promising results using existing relation discovery datasets to demonstrate the effectiveness of our proposed learning algorithm for the new relation discovery task.

The code we used to train and evaluate our models is available at `https://github.com/HKUST-KnowComp/RE-RegDVAE`.

2 Problem Definition

We use $\mathcal{X}$ to denote the set of all training sentences. $\mathcal{V}$ is the set of named entities that are recognized by an NER system in $\mathcal{X}$, and (e_1, e_2) is the pair of first and second entities in a given sentence $x \in \mathcal{X}$. $\mathcal{R}_\mathcal{X}$ is the set of relation labels for $\mathcal{X}$. In addition, there exists an external knowledge base $\mathcal{G}(\mathcal{E}_\mathcal{G}, \mathcal{T}_\mathcal{G})$, consisting of a set of entities $\mathcal{E}_G$ and relations $\mathcal{R}_G$ and triplets $\mathcal{T}_G$ where a triplet consists of two entities with their relation.

Our model is a relation extractor to predict the underlying semantic relation $r \in \mathcal{R}_\mathcal{X}$ given sentences $\mathcal{X}$, with the help of $\mathcal{G}(\mathcal{E}_\mathcal{G}, \mathcal{T}_\mathcal{G})$. In particular, we focus on the challenging scenario where $\mathcal{R}_\mathcal{X} \cap \mathcal{R}_G = \emptyset$.

3 Model

In this section, we first review the discrete-state variational autoencoder (DVAE) in §3.1. Then we introduce our new framework in §3.2.

3.1 DVAE for Relation Discovery

Assuming that we perform generative modeling, where each latent relation r follows a uniform prior distribution $p_u(r)$, we follow (Marcheggiani and Titov, 2016) to optimize a pseudo-likelihood:

$$\mathcal{L}(\theta) = \log \sum_{r \in \mathcal{R}_\mathcal{X}} p(e_i, e_{-i}|r, \theta) p_u(r) \quad (1)$$

$$\approx \sum_{i=1}^{2} \log \sum_{r \in \mathcal{R}_\mathcal{X}} p(e_i|e_{-i}, r, \theta) p_u(r), \quad (2)$$

where e_i and e_{-i} are entities, $i \in \{1, 2\}$ and e_{-i} denotes the complement $\{e_1, e_2\} \setminus \{e_i\}$. $p(e_i|e_{-i}, r, \theta)$ is the probability of one entity given another entity as well as the relation, where θ denotes the set of parameters. Note that this probability p is defined on the triplet (e_1, r, e_2) which is universal across different sentences containing the two entities.

The pseudo-likelihood $\mathcal{L}(\theta)$ can be lower-bounded based on Jensen's inequality through a variational posterior $q(r|x, \psi)$:

$$\mathcal{L}(\theta, \psi) = \sum_{i=1}^{2} \sum_{r \in \mathcal{R}_T} q(r|x, \psi) \log p(e_i|e_{-i}, r, \theta) + \alpha H\left[q(r|x, \psi)\right], \quad (3)$$

where $q(r|x, \psi)$ predicts the relation based on the whole sentence x as an input and ψ as the set of parameters. H is the entropy to regularize the probability distribution q, and α is the hyper-parameter to balance the regularization strength.

This model consists of two components, an encoder $q(r|x, \psi)$ which encodes sentence features into a relation distribution, and a decoder $p(e_i|r, e_{-i}, \theta)$ which predicts an entity given the relation cluster and another entity. Both are modeled by softmax functions:

$$q(r|x, \psi) = \frac{\exp\left(\mathbf{w}_r^\intercal \mathbf{g}(x)\right)}{\sum_{r' \in \mathcal{R}_\mathcal{X}} \exp\left(\mathbf{w}_{r'}^\intercal \mathbf{g}(x)\right)}, \quad (4)$$

$$p(e_i|e_{-i}, r, \theta) = \frac{\exp\left(\phi(e_i, e_{-i}, r, \theta)\right)}{\sum_{e_i' \in \mathcal{V}} \exp\left(\phi(e_i', e_{-i}, r, \theta)\right)}, \quad (5)$$

where $\psi = \{\mathbf{w}_r | r \in \mathcal{R}_\mathcal{X}\}$ and $\mathbf{g}(x)$ is a vector representation of sentence x, which can be high-dimensional one-hot feature encodings or low-dimensional sentence embeddings encoded by deep neural networks. $\phi(e_1, e_2, r, \theta)$ can be a general scoring function defined over triplets. We use the instantiation with the best performance shown by (Marcheggiani and Titov, 2016), which is a combination of bilinear model and selectional preference model:

$$\phi(e_1, e_2, r, \theta) = \mathbf{e}_1^\intercal \mathbf{C}_r \mathbf{e}_2 + [\mathbf{e}_1, \mathbf{e}_2]^\intercal \mathbf{r} \quad (6)$$

where $\theta = \{\mathbf{C}_r, \mathbf{r}, \mathbf{e}_i | r \in \mathcal{R}_T, e_i \in \mathcal{V}\}$, $\mathbf{C}_r$ is a matrix, $\mathbf{r}$ is a vector for the relation r, $\mathbf{e}_1$ and $\mathbf{e}_2$ are the vectors for head and tail entities respectively, and $[\mathbf{e}_1, \mathbf{e}_2]$ is the concatenation of the vector representations of the two entities.

The DVAE model directly optimizes the variational lower bound by doing gradient ascent for θ and ψ jointly. Both encoder $q(r|x, \psi)$ and decoder $p(e_i|r, e_{-i}, \theta)$ are implemented as neural networks. Standard training techniques and tricks can be applied.

3.2 Knowledge Base Constraint

Our KB constraint framework can be summarized as a two-step procedure: KB constraints construction and regularization for the learning model. In the constraints construction step, a set of sentences is formed as a query to KB and retrieves a set of constraints back. Then in the regularization step, we apply the constraint to regularize posterior distributions of the relation extractor.

Conceptually, given a set of sentences $\mathcal{X}$, we want to bias the learning result: After the entities are linked to the KB, if KB inference indicates that some pairs should be in a relation based on a set of rules Υ, then the extractor should be constrained to output it. This constraint can be encoded into a feature function $Q(\mathcal{X})$ = "entity pairs in the same relation based on Υ" and put into the posterior regularization framework (Gillenwater et al., 2011). However, the computational complexity of the feature function is exponential since we need to traverse the KB to find Υ. We instead consider the must-link and cannot-link constraints (Basu et al., 2004), indicating respectively that a pair of sentences should be or should not be labeled as the same relation. For each pairwise constraint, the model assigns an associated cost of violating that constraint for the model regularization.

3.2.1 KB Constraints Construction

From the perspective of KB, a must-link constraint on sentences (x_1, x_2) exists if two pairs of entities $(p_1, p_2) = [(e_{1,1}, e_{1,2}), (e_{2,1}, e_{2,2})]$ are similar given the KB, where $(e_{i,1}, e_{i,2})$ is the entity pair

Euclidean L_2 distance	$d_{Euc}\left(q_1(r), q_2(r)\right) = \sqrt{\sum_r \left\|q_1(r) - q_2(r)\right\|^2}$
Kullback-Leibler (KL) divergence	$d_{KL}(q_1(r), q_2(r)) = \sum_r q_1(r) \log\left(\frac{q_1(r)}{q_2(r)}\right)$
Jensen-Shannon (JS) divergence	$d_{JS}(q_1(r), q_2(r)) = \frac{1}{2}\sum_r q_1(r) \log\left(\frac{2q_1(r)}{q_1(r)+q_2(r)}\right) + \frac{1}{2}\sum_r q_2(r) \log\left(\frac{2q_2(r)}{q_1(r)+q_2(r)}\right)$

Table 1: Cluster regularization with different distance or divergences.

belongs to sentence x_i. This motivates us to define a similarity score for a pair of entity pairs. Instead of modeling the common relation paths or logic rules, which is computationally infeasible, we compare them in the latent embedding space. In particular, we model the KB using the TransE (Bordes et al., 2013) model, where a relation is interpreted as a translation from the head entity to the tail entity, with a score function, $\mathbf{e}_1 + \mathbf{r} = \mathbf{e}_2$ for each gold triplet (e_1, r, e_2) in the KB. This operation is fast and the latent embeddings are expressive in many cases. Then we can reason the latent relation representation of a particular pair in vector space by $\mathbf{r}_i = \mathbf{e}_{i,2} - \mathbf{e}_{i,1}$, without the need for extra parameters. Here $\mathbf{r}_i$ is not necessarily a real relation between two entities in the KB but just reflects the geometric property. The penalty for violating a must-link constraint between a pair of sentences with a high KB score should be higher than those with low KB scores. This further inspires us to define a soft constraint penalty based on the similarity of latent KB relations.

Here, we use the adjusted cosine similarity (Sarwar et al., 2001) between two latent relations as a must-link confidence score

$$s^+(x_1, x_2) = [\cos(\mathbf{e}_{1,2} - \mathbf{e}_{1,1}, \mathbf{e}_{2,2} - \mathbf{e}_{2,1})]^+_{\gamma^+} \tag{7}$$

where $[x]^+_{\gamma^+} = x$ if $x > \gamma^+$ otherwise 0, $\gamma^+ \in [0, 1]$ is a threshold we defined to control the must-link scope, $e_{i,j}$ is named entity in x_i and $\mathbf{e}_{i,j}$ is its embedding. The similarity between $\mathbf{e}_{1,2} - \mathbf{e}_{1,1}$ and $\mathbf{e}_{2,2} - \mathbf{e}_{2,1}$ evaluates whether two sentences indicate similar relations according to the KB embedding.

We also define the cannot-link in a similar way, where two sentences cannot be in the same cluster with a confidence

$$s^-(x_1, x_2) = [\cos(\mathbf{e}_{1,2} - \mathbf{e}_{1,1}, \mathbf{e}_{2,2} - \mathbf{e}_{2,1})]^-_{\gamma^-} \tag{8}$$

where $[x]^-_{\gamma^-} = x$ if $x < -\gamma^-$ otherwise 0, and $\gamma^- \in [0, 1]$ is a threshold we defined to control the cannot-link scope. We simply set $\gamma^+ = \gamma^- = \gamma$.

3.2.2 Clustering Regularization

For each pair of sentences (x_1, x_2), the relation extractor will predict a clustering posterior $q_i(r|x_i, \psi)$, $i = 1, 2$, which can be computed based on Eq. (4). We regularize the clustering result on the probability distance between sentence pairs, using either Euclidean L_2 distance, Kullback-Leibler (KL) divergence, or Jensen-Shannon (JS) divergence. The computation of the distance or divergences can be found in Table 1.

Then the soft constraints introduced in §3.2.1 are applied on the corresponding distance to calculate the regularization terms:

$$D^+(x_1, x_2) = -d_*\left(q_1(r), q_2(r)\right) s^+(x_1, x_2), \tag{9}$$

$$D^-(x_1, x_2) = d_*\left(q_1(r), q_2(r)\right) |s^-(x_1, x_2)|, \tag{10}$$

for must and cannot links respectively, where d_* can be d_{Euc}, d_{KL}, or d_{JS}. Taking must-link constraint as an example, if the posterior distributions $q_1(r|x_1, \psi)$ and $q_2(r|x_2, \psi)$ are different from each other but KB suggests that these two sentences should be in the same cluster where $s^+(x_1, x_2)$ is large, then $d_*\left(q_1(r), q_2(r)\right)$ being large means there is a large cost when q_1 and q_2 being different. Then in the training phase, we want to reduce this cost given the constraint.

The constraints above are defined in a $|\mathcal{X}| \times |\mathcal{X}|$ space. It is almost impossible to enumerate all of the constraints. To make it trainable, we instead gather the constraints within a mini-batch. Since in different training epochs we randomly permute the training samples, it is possible to touch many pairs of sentences in practice.

3.3 Learning

The model parameters only exist in original autoencoder components (i.e., ψ and θ), which can be jointly optimized by maximizing the following

objective function with L_2 regularization:

$$\mathcal{L}(\theta,\psi) = \sum_{x\in X}\sum_{i=1}^{2}\sum_{r\in R_T} q(r|x,\psi)\log p(e_i|e_{-i},r,\theta) + \sum_{x\in X}\alpha H\left[q(r|x,\psi)\right] + \sum_{X_i\sim X}\sum_{(x_1,x_2)\in X_i}\beta D(x_1,x_2) + \lambda\|(\psi,\theta)\|_2, \tag{11}$$

where α, β, γ, and λ are hyper-parameters to control the regularization strength. D can be D^+ or D^- depending on the cosine similarity between pairs. In practice, we apply annealing method over α in an exponential way:

$$\alpha_t = \alpha_0\exp(-\eta t) \text{ and } \eta = \frac{\log(\alpha_0/\alpha_T)}{T},$$

where α_0 is the initial value, and α_T is the final value, t and T are the current and total training steps respectively. This method enables the extractor to explore more possibilities first and finally converge to a stable distribution.

It is difficult to directly compute the partition function in Eq. (5), as it requires to sum over $|\mathcal{V}|$. We use the same negative sampling method as (Marcheggiani and Titov, 2016) to substitute $\log p(e_i|e_{-i},r,\theta)$ in Eq. (11) with:

$$\log p(e_i|e_{-i},r,\theta) \approx \log\sigma(\phi(e_i,e_{-i},r,\theta)) + \sum_{e^{\text{neg}}\in\mathcal{N}}\log\sigma\left(-\phi(e^{\text{neg}},e_i,r,\theta)\right),$$

where $\mathcal{N}$ is the set of randomly sampled entities in $\mathcal{V}$ and σ is the sigmoid function.

4 Experiments

In this section, we show the experimental results.

4.1 Dataset and Preprocessing

We evaluate our model in the context of unsupervised relation discovery and compare to the baseline model, DVAE (Marcheggiani and Titov, 2016) which is the current state-of-the-art of relation discovery. Distant supervision assumes that the relations should be aligned between the KB and the training text corpus, which is not available in our setting.

We tested our model on three different subsets of New York Times corpus (NYT) (Sandhaus and Evan, 2008).

	Data	NYT122	NYT71	NYT27
Text	# sentences	67,123	14,210	87,144
	# facts	9,207	2,274	8,559
	# entity pairs	20,939	3,539	36,714
	# entities	5,865	2,489	4,803
	# relations	122	71	27
KB	# triplets	401,490	456,146	439,507
	# entity pairs	331,008	373,875	354,960
	# entities	14,907	14,933	14,911
	# relations	705	1,009	1,031

Table 2: Statistics of datasets. # facts in the text corpus is the number of sentences with relation labels.

- The first one is widely used in unsupervised settings, which was developed by Yao et al. (2011) and has also been used by Marcheggiani and Titov (2016). This dataset contains articles 2000 to 2007, with named entities annotated and features processed (POS tagging, NER, and syntactic parsing). We use this dataset to compare with previous work directly (Marcheggiani and Titov, 2016).

- The second and third ones are usually applied by supervised models. So when they generated the data, they tended to focus on relations with more supporting sentences. The second one was developed by Zeng et al. (2017). The data is built by aligning Wikidata (Vrandečić, 2012) relations with NYT corpus, as a result of 99 possible relations. It is built to contain more updated facts and richer structures of relations, e.g., a larger number of relation/relation paths. We use this dataset to amplify the effects coming from relation paths in KB, as the data was used to train a path-based relation extraction model.

- The third one was developed by Riedel et al. (2010) and has also been used by Lin et al. (2016). This dataset was generated by aligning Freebase (Bollacker et al., 2008) relations with NYT in 2005-2007, and with 52 possible relations. We use this data to test the clustering result with a narrow relation domain.

We align these datasets against FB15K, which is a randomly sampled subset of Freebase developed by Bordes et al. (2013). For each of the datasets above, we hold out the triplets in FB15K that contains relations in corresponding text data, so that we ensure that KB cannot give any direct supervision on any relation labels. We then discard named

Model	Metrics							
	Prediction based on encoder				Prediction based on decoder			
	F1		NMI		F1		NMI	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DVAE	0.417	0.011	0.339	0.009	0.419	0.011	0.337	0.014
RegDVAE (Euclidean at encoder)	**0.469**	0.014	**0.430**	0.020	**0.448**	0.020	**0.384**	0.020
RegDVAE (KL at encoder)	0.375	0.009	0.359	0.014	0.380	0.011	0.355	0.014
RegDVAE (JS at encoder)	0.435	0.038	0.370	0.042	0.409	0.012	0.336	0.005
RegDVAE (Euclidean at decoder)	0.416	0.019	0.329	0.017	0.350	0.012	0.201	0.054

Table 3: Comparison results on NYT122 with different prediction and regularization strategies (using encoder or decoder).

entities in text corpus if they are not shown in KB, so that we can directly test the influence of our KB constraint model. Finally, we only keep a single label for each sentence, and e_1, e_2 follow the occurrence order in the sentence. The resulting datasets contain 122, 71, and 27 relation labels respectively, so we name them as NYT122, NYT71, and NYT27. The statistics of the three datasets are shown in Table 2. For NYT71 and NYT27, we perform the same feature extraction as NYT122 shown in (Marcheggiani and Titov, 2016).

4.2 Implementation Details

All the model parameters are initialized randomly. The number of negative samples is set to 5, mini-batch size is set to 100 with 80 epochs. We optimize all the models using AdaGrad (Duchi et al., 2011) with initial learning rate at 0.5. For NYT122, we induce 40 relations clusters, with $\alpha_0 = 4$, $\alpha_T = 10^{-5}$, $\beta = 0.6$, and $\gamma = 0.9$. For NYT71, we induce 30 relations clusters, with $\alpha_0 = 2$, $\alpha_T = 10^{-4}$, $\beta = 0.8$, and $\gamma = 0.95$. For NYT27, we induce 20 relations clusters, with $\alpha_0 = 2$, $\alpha_T = 10^{-4}$, $\beta = 0.8$, and $\gamma = 0.3$. We train TransE as our KB embedding model with 50 dimensions and 1,000 epochs.

We report the average and standard deviation based on five different runs. We randomly split the data into validation:test=4:6. All the model selections were based on validation sets, and final evaluation results will be only based on test sets.

4.3 Evaluation and Discussion

As the scoring function, we use the B^3F_1 (Bagga and Baldwin, 1998) which has also been used by our baseline (Marcheggiani and Titov, 2016), and Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002) metrics. Both are standard measures for evaluating clustering tasks.

Regularization and Prediction Strategies. We first report our results on NYT122 using different regularization and prediction settings, as this dataset was used by our baseline model DVAE.

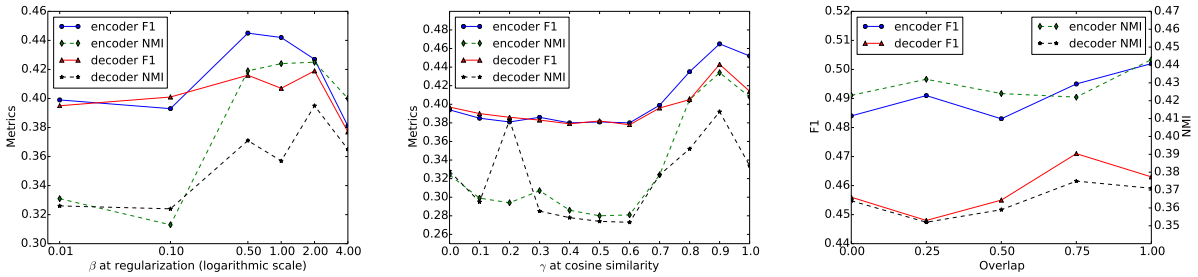
Note that both encoder and decoder components can make relation predictions. In fact, the way of using encoder $q(r|x, \psi)$ for each sentence is straightforward. Then based on the encoder, we predict relation on the basis of single occurrence of entity pair. When using the decoder, we need to re-normalize $p(e_i|r, e_{-i}, \theta)$ as $p(r|e_1, e_2, \theta)$ to make predictions. Based on the decoder, we make predictions for each unique entity pair. As a consequence, our constraints can be imposed on both encoder and decoder. The way of computing decoder probability distribution is the same as making predictions. So in this experiment, we report both results.

The results are shown in Table 3. From the table, we can see that regularization with Euclidean distance performs the best compared to KL and JS. Moreover, the regularization over encoder is better than the regularization over decoder. This may be because the way that we put constraints only over sampled sentences in a batch may hurt the regularization of decoder, since sampled unique pairs may be less than sample sentences. If we look at results comparing original DVAE prediction based on the encoder and the decoder, both result in similar F1 and NMI numbers. Thus, we can only conclude that currently in the way we do sampling, constraining over encoder is a better choice.

Comparison on Different Datasets. We also compare our algorithm on the three datasets with different baseline settings. In order to evaluate our model rigorously, besides the original DVAE model, we compare two additional augmented baseline models with the same hyper-parameter setting: DVAE with TransE embeddings appended to encoder input features (DVAE+E) and DVAE

Model	NYT122				NYT71				NYT27			
	F1		NMI		F1		NMI		F1		NMI	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Majority	0.355	-	0	-	0.121	-	0	-	0.549	-	0	-
DVAE	0.417	0.011	0.339	0.009	0.325	0.011	0.375	0.023	0.433	0.018	0.384	0.021
DVAE+E	0.385	0.021	0.341	0.043	0.339	0.021	0.418	0.022	0.396	0.034	0.381	0.039
DVAE+D	0.452	0.033	0.438	0.022	0.352	0.038	0.339	0.009	0.499	0.040	0.469	0.027
RegDVAE	0.469	0.014	0.430	0.020	0.377	0.020	0.466	0.036	0.587	0.005	0.451	0.005
RegDVAE+D	**0.499**	0.022	**0.497**	0.013	**0.432**	0.028	**0.589**	0.071	**0.665**	0.022	**0.562**	0.038

Table 4: Comparison of prediction results based on encoder using NYT122, NYT71, and NYT27 datasets with different KB regularization strategies.



(a) Parameter sensitivity analysis at β. (b) Parameter sensitivity analysis at γ. (c) Relation overlap ratios.

Figure 2: Comparison results on NYT122 with different parameters and relation overlaps. The predictions are based on either the encoder or the decoder.

with decoder entity vectors replaced by pre-trained KB embeddings (DVAE+D). For our method, we report RegDVAE with the best setting where we use Euclidean distance based constraints to regularize the encoder. Moreover, we report a setting with fixed embeddings in the decoder as the ones obtained from TransE (RegDVAE+D). This also makes sense since even though the TransE embeddings are not trained with the observation of the same relations as the text corpus, the embeddings already contain much semantic information about entities. Then by fixing the embeddings of entities in the decoder, we can significantly reduce the number of parameters that need to be trained. The results are shown in Table 4. As we can see that, RegDVAE+D can outperform the original DVAE by 8∼23 points on F1. DVAE+D is also good but may fail when there are a lot of out-of-sample entities in the training corpus.

Hyper-parameter Sensitivity. We have three hyper-parameters in our algorithm: α_0 for the regularization of encoder entropy, β for the regularization with our constraints, and γ for the threshold of KB based cosine similarities. Here, we test β and γ, since the sensitivity result of α_0 is the same as the original DVAE work (Marcheggiani and Titov, 2016). The sensitivity of β is shown in Figure 2(a). The results are good in a wide range from $\beta = 0.5$ to $\beta = 2$. The sensitivity of γ is shown in Figure 2(b). It reveals some interesting patterns. At the beginning when γ is small, it hurts the performance. After γ getting greater than 0.7, it improves the performance, which means that only very similar relations indicated by KB embeddings are useful relations as constraints. In addition, $\gamma = 1$ (meaning only finding identical relations) is worse than $\gamma = 0.9$, which means we indeed find some relations in our KB so that different triplets will be constrained.

KB Relation Overlap. Although we assume that there is no overlapped relation between the KB and the training text corpus, in practice, we may find a lot of applications that the relations are partially observed in KB. Thus, we also test a setting when the KB has different proportions of overlapped relations with training text corpus. In this case, we train different KB embeddings for different percentages of overlapped relations, and then apply the embeddings into the constraints. The results are shown in Figure 2(c). As we can see, in general, more overlapped relations will result in better performance. The best number can be better than the number without overlapped relation by about two points. This again verifies that the KB embedding is very robust and represent the semantic meanings of entities even with part of the

Contextual Sentence	Cluster	Similarity
...**Spain** will become the third country in **Europe**...	12	0.926
Portugal, with all that talent, goes home to **Europe**...	12	
Brazil, **Latin America** 's largest economy ...	12	0.916
...**Argentina** was perhaps the most expensive country in **Latin America** for tourists....	12	

Table 5: Examples for relation: /location/contained_by.

relations observed (Bordes et al., 2013).

Case Study. We also show some examples of entity pair similarities in Table 5. From the Table we can see that our target relation cluster is */location/contained_by*. In the first example, the similarity between entity pairs (`Spain, Europe`) and (`Portugal, Europe`) are high, which indicates the same cluster of pairs of sentences. The same constraint is applied in the second example, although there's no direct connection between (`Brazil, Latin America`), (`Argentina, Latin America`).

5 Related Work

Supervised and Distantly Supervised Relation Extraction. Traditional supervised relation extraction focuses on a limited number of relations (Roth and Yih, 2002; Kambhatla, 2004; Chan and Roth, 2010). Distant supervision uses KBs to obtain a lot of automatically annotated data (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Xu et al., 2013a; Zhang et al.; Zeng et al., 2015; Lin et al., 2016; Zeng et al., 2017). There are two important assumptions behind these models, namely multi-instance learning (Riedel et al., 2010) and multi-instance multi-label learning (Hoffmann et al., 2011; Surdeanu et al., 2012). Our setting is similar to multi-instance learning but we assume there is no overlapped relation between KB and training text corpus. Universal schema (Riedel et al., 2013; Verga et al., 2016; Toutanova et al., 2015; McCallum et al., 2017) can also exploit KB to help extract relations. It needs a lot of entity pairs in text to co-occur with KB triplets, which is under the same setting with distant supervision. Those surface patterns are pre-extracted and shown in the training phase, which makes it also a weakly supervised learning method.

Unsupervised Relation Extraction. Open Domain Information Extraction (Open-IE) assumes that every relation expression can represent a unique relation (Etzioni et al., 2004; Banko et al., 2007; Fader et al., 2011; Mausam et al., 2012; Xu et al., 2013b; Angeli et al., 2015). On the other hand, relation clustering approaches group all the related relation expressions to represent a relation (Lin and Pantel, 2001; Mohamed et al., 2011; Takamatsu et al., 2011; Yao et al., 2011, 2012; Nakashole et al., 2012a,b; Marcheggiani and Titov, 2016). Our setting is based on (Marcheggiani and Titov, 2016) but we also introduce KB as a different kind of weak and indirect supervision.

Knowledge Base Representation. Embedding based knowledge base representation learning methods (Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015; Trouillon et al., 2016) represent entities and relations as vectors, denoted as $\mathbf{e}$ and $\mathbf{C}_r$ respectively such that for a distances function f, the value $f(\mathbf{e}_1, \mathbf{C}_r, \mathbf{e}_2)$ is maximized for all (e_1, r, e_2) facts. Among all these methods, TransE model has a favorable property that the translation operation can be easily recovered by entity vectors ($\mathbf{r}_{1,2} = \mathbf{e}_1 - \mathbf{e}_2$). With its simplicity and high performance, TransE is enough for demonstration. Though our method is not restricted to the representation form of KB, we leave it for future evaluation.

Constraints can be made more explainable by paths finding. For instance, the Path Ranking Algorithm (PRA) (Lao and Cohen, 2010; Lao et al., 2011) uses random walk to perform multi-hop reasoning based on logic rules. Later on, reinforcement Learning (Toutanova et al., 2015; Xiong et al., 2017; Das et al., 2017; Chen et al., 2018) is used to search for paths more effectively. Though heuristics are used to further reduce the number of mined relations, it is still very costly to find the paths for KB with hundreds of relations, if not impossible.

Constraint Modeling. Originated from semi-supervised learning (Chapelle et al., 2006), must-link and cannot-link modeling has been well studied in machine learning community (Wagstaff et al., 2001; Basu et al., 2004, 2008). Such constraints were usually generated based on the ground truth labels of data. For document clustering, word constraints constructed based on WordNet similarities have been applied (Song et al., 2013) and entity constraints based on entity types

in an external KB have been used (Wang et al., a, 2016), both being considered as a kind of indirect supervision based on side information. For triplet relation clustering, relation surface similarity and entity type constraints have been explored (Wang et al., b). However the above constraints are applied to a particular form of models, co-clustering models. Compared to existing approaches, our constraints are constructed based on more recently developed KB embeddings, which is more flexible and easy to incorporate into different models.

In natural language processing community, constraints based on background knowledge are also well studied. For example, constrained conditional models (CCM) (Chang et al., 2012) provides a very flexible framework to decouple learning and inference, where in the inference step, background knowledge can be incorporated as an ILP (integer linear programming) problem. Posterior regularization (PR) (Ganchev et al., 2010) generalizes this idea so that it uses a joint learning and inference framework to incorporate the background knowledge. Both CCM and PR have many applications including the application to relation extraction (Chan and Roth, 2010; Chen et al., 2011). Compared to these existing approaches, our constraints are derived from the general-purpose KB, which is quite different from their way of manually crafting some background knowledge as declarative rules.

It is very interesting that we are similar to the PR framework. Since we use a DVAE framework as the base algorithm, there is no traditional E-step and M-step in the variational inference. Instead, only q and p probabilities parameterized by neural networks are updated. In our framework, we can add constraints to either q or p probabilities (applying to p needs modification of normalization). It is the same that we draw a biased learning process when estimating the posteriors as PR does.

6 Conclusion

In this paper, we propose a new relation discovery setting where there is no overlapped relations between the training text corpus and the KB. We propose a new learning framework of KB regularization which uses must-link and cannot-link constraints derived based on similarities in the KB embedding space. Our method improves the results over all baseline models without harming the scalability. We believe this framework is as flexible as other constraint models to be applied to many applications when we think the semantics of entities and relations provided by the KB is useful.

Acknowledgments

This paper was supported by the Early Career Scheme (ECS, No. 26206717) from Research Grants Council in Hong Kong. We thank Intel Corporation for supporting our deep learning related research. We also thank the anonymous reviewers for their valuable comments and suggestions that help improve the quality of this manuscript.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *ACL*, pages 344–354.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *LREC*, pages 563–566.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, pages 2670–2676.

Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. 2004. A probabilistic framework for semi-supervised clustering. In *KDD*, pages 59–68.

Sugato Basu, Ian Davidson, and Kiri Wagstaff. 2008. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Yee Seng Chan and Dan Roth. 2010. Exploiting background knowledge for relation extraction. In *COLING*, pages 152–160.

Ming-Wei Chang, Lev-Arie Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine Learning*, 88(3):399–431.

O. Chapelle, B. Schölkopf, and A. Zien, editors. 2006. *Semi-Supervised Learning*. MIT Press.

Harr Chen, Edward Benson, Tahira Naseem, and Regina Barzilay. 2011. In-domain relation discovery with meta-constraints via posterior regularization. In *ACL-HLT*, pages 530–540.

Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Yang Wang. 2018. Variational knowledge graph reasoning. In *NAACL-HLT*, pages 1823–1832.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alexander J. Smola, and Andrew McCallum. 2017. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *CoRR*, abs/1711.05851.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *KDD*, pages 601–610.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Oren Etzioni, Michael Cafarella, and Doug Downey. 2004. Webscale information extraction in knowitall (preliminary results). In *WWW*, pages 100–110.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Jennifer Gillenwater, Kuzman Ganchev, João Graça, Fernando Pereira, and Ben Taskar. 2011. Posterior sparsity in unsupervised dependency parsing. *Journal of Machine Learning Research*, 12:455–490.

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 *ACL*, pages 541–550.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *ACL - Poster and Demonstration*.

Ni Lao and William W. Cohen. 2010. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81(1):53–67.

Ni Lao, Tom M. Mitchell, and William W. Cohen. 2011. Random walk inference and learning in A large scale knowledge base. In *EMNLP*, pages 529–539.

Dekang Lin and Patrick Pantel. 2001. DIRT – discovery of inference rules from text. In *KDD*, pages 323–328.

Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*, pages 705–714.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *ACL*, pages 2124–2133.

Xitong Liu, Fei Chen, Hui Fang, and Min Wang. 2014. Exploiting entity relationship for query expansion in enterprise search. *Inf. Retr.*, 17(3):265–294.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistic*, 4:231–244.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *EMNLP-CoNLL*, pages 523–534.

Andrew McCallum, Arvind Neelakantan, and Patrick Verga. 2017. Generalizing to unseen entities and entity pairs with row-less universal schema. In *EACL*, pages 613–622.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/AFNLP*, pages 1003–1011.

Thahir Mohamed, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2011. Discovering relations between noun categories. In *EMNLP*, pages 1447–1455.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012a. Discovering and exploring relations on the web. *PVLDB*, 5(12):1982–1985.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012b. PATTY: A taxonomy of relational patterns with semantic types. In *EMNLP*, pages 1135–1145.

Deepak Ravichandran and Eduard H. Hovy. 2002. Learning surface text patterns for a question answering system. In *ACL*, pages 41–47.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML PKDD*, pages 148–163.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *NAACL-HLT*, pages 74–84.

Dan Roth and Wen-tau Yih. 2002. Probabilistic reasoning for entity & relation recognition. In *COLING*.

Sandhaus and Evan. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295.

Yangqiu Song, Shimei Pan, Shixia Liu, Furu Wei, Michelle X. Zhou, and Weihong Qian. 2013. Constrained text coclustering with supervised and unsupervised constraints. *IEEE Trans. Knowl. Data Eng.*, 25(6):1227–1239.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP-CoNLL*, pages 455–465.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2011. Probabilistic matrix factorization leveraging contexts for unsupervised relation extraction. In *PAKDD*, pages 87–99.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, pages 1499–1509.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *ICML*, pages 2071–2080.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *NAACL-HLT*, pages 886–896.

Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *WWW*, pages 1063–1064.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. 2001. Constrained k-means clustering with background knowledge. In *ICML*, pages 577–584.

Chenguang Wang, Yangqiu Song, Ahmed El-Kishky, Dan Roth, Ming Zhang, and Jiawei Han. a. Incorporating world knowledge to document clustering via heterogeneous information networks. In *KDD*, pages 1215–1224.

Chenguang Wang, Yangqiu Song, Dan Roth, Chi Wang, Jiawei Han, Heng Ji, and Ming Zhang. b. Constrained information-theoretic tripartite graph clustering to identify semantically similar relations. In *IJCAI*, pages 3882–3889.

Chenguang Wang, Yangqiu Song, Dan Roth, Ming Zhang, and Jiawei Han. 2016. World knowledge as indirect supervision for document clustering. *ACM Transactions on Knowledge Discovery from Data*, 11(2):13:1–13:36.

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

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *EMNLP*, pages 564–573.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013a. Filling knowledge base gaps for distant supervision of relation extraction. In *ACL*, pages 665–670.

Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. 2013b. Open information extraction with tree kernels. In *NAACL-HLT*, pages 868–877.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *EMNLP*, pages 1456–1466.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2012. Unsupervised relation discovery with sense disambiguation. In *ACL*, pages 712–720.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *EMNLP*, pages 1768–1777.

Xingxing Zhang, Jianwen Zhang, Junyu Zeng, Jun Yan, Zheng Chen, and Zhifang Sui. Towards accurate distant supervision for relational facts extraction. In *ACL (2)*, pages 810–815.