Global Relation Embedding for Relation Extraction

Yu Su[*], **Honglei Liu**[*], **Semih Yavuz, Izzeddin Gür**
University of California, Santa Barbara
{ysu,honglei,syavuz,izzeddingur}@cs.ucsb.edu

Huan Sun
The Ohio State University
sun.397@osu.edu

Xifeng Yan
University of California, Santa Barbara
xyan@cs.ucsb.edu

Abstract

We study the problem of textual relation embedding with distant supervision. To combat the wrong labeling problem of distant supervision, we propose to embed textual relations with *global statistics* of relations, i.e., the co-occurrence statistics of textual and knowledge base relations collected from the entire corpus. This approach turns out to be more robust to the training noise introduced by distant supervision. On a popular relation extraction dataset, we show that the learned textual relation embedding can be used to augment existing relation extraction models and significantly improve their performance. Most remarkably, for the top 1,000 relational facts discovered by the best existing model, the precision can be improved from 83.9% to 89.3%.

1 Introduction

Relation extraction requires deep understanding of the relation between entities. Early studies mainly use hand-crafted features (Kambhatla, 2004; Zhou et al., 2005), and later kernel methods are introduced to automatically generate features (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2006). Recently neural network models have been introduced to embed words, relations, and sentences into continuous feature space, and have shown a remarkable success in relation extraction (Socher et al., 2012; Zeng et al., 2014; Xu et al., 2015b; Zeng et al., 2015; Lin et al., 2016). In this work, we study the problem of embedding *textual relations*, defined as the shortest dependency path[1] between two entities in the dependency graph of a sentence, to improve relation extraction.

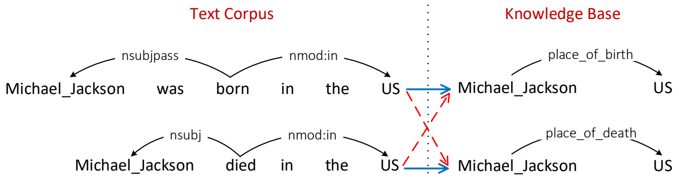
Textual relations are one of the most discriminative textual signals that lay the foundation of many relation extraction models (Bunescu and Mooney, 2005). A number of recent studies have explored textual relation embedding under the supervised setting (Xu et al., 2015a,b, 2016; Liu et al., 2016), but the reliance on supervised training data limits their scalability. In contrast, we embed textual relations with *distant supervision* (Mintz et al., 2009), which provides much larger-scale training data without the need of manual annotation. However, the assertion of distant supervision, "*any* sentence containing a pair of entities that participate in a knowledge base (KB) relation is likely to express the relation," can be violated more often than not, resulting in many wrongly labeled training examples. A representative example is shown in Figure 1. Embedding quality is thus compromised by the noise in training data.

Our main contribution is a novel way to combat the wrong labeling problem of distant supervision. Traditional embedding methods (Xu et al., 2015a,b, 2016; Liu et al., 2016) are based on *local statistics*, i.e., individual textual-KB relation pairs like in Figure 1 (Left). Our key hypothesis is that *global statistics is more robust to noise than local statistics*. For individual examples, the relation label from distant supervision may be wrong from time to time. But when we zoom out to consider the entire corpus, and collect the global co-occurrence statistics of textual and KB relations, we will have a more comprehensive view of relation semantics: The semantics of a textual relation can then be represented by its co-occurrence distribution of KB relations. For example, the distribution in Figure 1 (Right) indicates that the textual relation $\textsc{Subject} \xleftarrow{\text{nsubjpass}} \textit{born} \xrightarrow{\text{nmod:in}} \textsc{Object}$ mostly means `place_of_birth`, and is also a good indicator of `nationality`, but not `place_of_death`. Although it is still wrongly labeled with `place_of_death` a number of times, the negative impact becomes negligible. Similarly,

[*] Equally contributed.

[1] We use fully lexicalized shortest dependency path with directional and typed dependency relations.

Proceedings of NAACL-HLT 2018, pages 820–830
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics



	$\xleftarrow{\text{nsubjpass}}$ *born* $\xrightarrow{\text{nmod:in}}$	$\xleftarrow{\text{nsubj}}$ *died* $\xrightarrow{\text{nmod:in}}$
place_of_birth	1868	14
nationality	389	20
place_of_death	37	352
...	...	...

Figure 1: The wrong labeling problem of distant supervision, and how to combat it with global statistics. *Left*: conventional distant supervision. Each of the textual relations will be labeled with both KB relations, while only one is correct (blue and solid), and the other is wrong (red and dashed). *Right*: distant supervision with global statistics. The two textual relations can be clearly distinguished by their co-occurrence distribution of KB relations. Statistics are based on the annotated ClueWeb data released in (Toutanova et al., 2015).

we can confidently believe that SUBJECT $\xleftarrow{\text{nsubj}}$ *died* $\xrightarrow{\text{nmod:in}}$ OBJECT means `place_of_death` in spite of the noise. Textual relation embedding learned on such global statistics is thus more robust to the noise introduced by the wrong labeling problem.

We augment existing relation extractions using the learned textual relation embedding. On a popular dataset introduced by Riedel et al. (2010), we show that a number of recent relation extraction models, which are based on local statistics, can be greatly improved using our textual relation embedding. Most remarkably, a new best performance is achieved when augmenting the previous best model with our relation embedding: The precision of the top 1,000 relational facts discovered by the model is improved from 83.9% to 89.3%, a 33.5% decrease in error rate. The results suggest that relation embedding with global statistics can capture complementary information to existing local statistics based models.

The rest of the paper is organized as follows. In Section 2 we discuss related work. For the modeling part, we first describe how to collect global co-occurrence statistics of relations in Section 3, then introduce a neural network based embedding model in Section 4, and finally discuss how to combine the learned textual relation embedding with existing relation extraction models in Section 5. We empirically evaluate the proposed method in Section 6, and conclude in Section 7.

2 Related Work

Relation extraction is an important task in information extraction. Early relation extraction methods are mainly feature-based (Kambhatla, 2004; Zhou et al., 2005), where features in various levels, including POS tags, syntactic and dependency parses, are integrated in a max entropy model. With the popularity of kernel methods, a large number of kernel-based relation extraction methods have been proposed (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2006). The most related work to ours is by Bunescu and Mooney (Bunescu and Mooney, 2005), where the importance of shortest dependency path for relation extraction is first validated.

More recently, relation extraction research has been revolving around neural network models, which can alleviate the problem of exact feature matching of previous methods and have shown a remarkable success (e.g., (Socher et al., 2012; Zeng et al., 2014)). Among those, the most related are the ones embedding shortest dependency paths with neural networks (Xu et al., 2015a,b, 2016; Liu et al., 2016). For example, Xu et al. (2015b) use a RNN with LSTM units to embed shortest dependency paths without typed dependency relations, while a convolutional neural network is used in (Xu et al., 2015a). However, they are all based on the supervised setting with a limited scale. In contrast, we embed textual relations with distant supervision (Mintz et al., 2009), which provides much larger-scale training data at a low cost.

Various efforts have been made to combat the long-criticized wrong labeling problem of distant supervision. Riedel et al. (2010), Hoffmann et al. (2011), and Surdeanu et al. (2012) have attempted a multi-instance learning (Dietterich et al., 1997) framework to soften the assumption of distant supervision, but their models are still feature-based. Zeng et al. (2015) combine multi-instance learning with neural networks, with the assumption that at least one of the contextual sentences of an entity pair is expressing the target relation, but this will lose useful information in the neglected sentences. Instead, Lin et al. (2016) use all the contextual sentences, and introduce an attention mechanism to weight the contextual sentences. Li et al. (2017) also use an attention mechanism to weight contextual sentences, and incorporate additional entity description information from knowledge bases. Luo et al. (2017) manage to alleviate the negative impact of noise by modeling and learning noise transition patterns from data. Liu et al. (2017) propose to infer the true label of a context sentence using a truth discovery approach (Li et al., 2016). Wu et al. (2017) incorporate adversarial training, i.e., injecting random perturbations in training, to improve the robustness of relation extraction. Using PCNN+ATT (Lin et al., 2016) as base model, they show that adversarial training can improve its performance by a good margin. However, the base model implementation used by them performed inferior to the one in the original paper and in ours, and therefore the results are not directly comparable. No prior study has exploited global statistics to combat the wrong labeling problem of distant supervision. Another unique aspect of this work is that we focus on compact textual relations, while previous studies along this line have focused on whole sentences.

In universal schema (Riedel et al., 2013) for KB completion and relation extraction as well as its extensions (Toutanova et al., 2015; Verga et al., 2016), a binary matrix is constructed from the entire corpus, with entity pairs as rows and textual/KB relations as columns. A matrix entry is 1 if the relational fact is observed in training, and 0 otherwise. Embeddings of entity pairs and relations, either directly or via neural networks, are then learned on the matrix entries, which are still individual relational facts, and the wrong labeling problem remains. Global co-occurrence frequencies (see Figure 1 (Right)) are not taken into account, which is the focus of this study. Another distinction is that our method directly models the association between textual and KB relations, while universal schema learns embedding for shared entity pairs and use that as a bridge between the two types of relations. It is an interesting venue for future research to comprehensively compare these two modeling approaches.

3 Global Statistics of Relations

When using a corpus to train statistical models, there are two levels of statistics to exploit: *local* and *global*. Take word embedding as an example. The skip-gram model (Mikolov et al., 2013) is based on local statistics: During training, we sweep through the corpus and slightly tune the embedding model in each local window (e.g., 10 consecutive words). In contrast, in global statistics based methods, exemplified by latent semantic analysis (Deerwester et al., 1990) and GloVe (Pennington et al., 2014), we process the entire corpus to collect global statistics like word-word co-occurrence counts, normalize the raw statistics, and train an embedding model directly on the normalized global statistics.

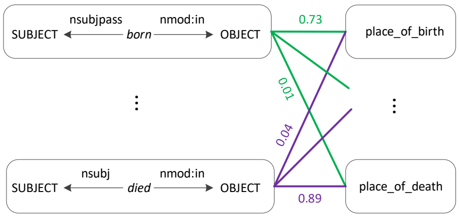


Figure 2: Relation graph. The left node set is textual relations, and the right node set is KB relations. The raw co-occurrence counts are normalized such that the KB relations corresponding to the same textual relation form a valid probability distribution. Edges are colored by textual relation and weighted by normalized co-occurrence statistics.

Most existing studies on relation extraction are based on local statistics of relations, i.e., models are trained on individual relation examples. In this section, we describe how we collect global co-occurrence statistics of textual and KB relations, and how to normalize the raw statistics. By the end of this section a bipartite *relation graph* like Figure 2 will be constructed, with one node set being textual relations $\mathcal{T}$, and the other being KB relations $\mathcal{R}$. The edges are weighted by the normalized co-occurrence statistics of relations.

3.1 Relation Graph Construction

Given a corpus and a KB, we first do entity linking on each sentence, and do dependency parsing if at least two entities are identified[2]. For each entity pair (e, e') in the sentence, we extract the fully lexicalized shortest dependency path as a textual relation t, forming a *relational fact* (e, t, e'). There are two outcomes from this step: a set of textual relations $\mathcal{T} = \{t_i\}$, and the *support* $S(t_i)$ for each t_i. The support of a textual relation is a *multiset* containing the entity pairs of the textual relation. The *multiplicity* of an entity pair, $m_{S(t_i)}(e, e')$, is the number of occurrences of the corresponding relational fact (e, t_i, e') in

[2] In the experiments entity linking is assumed given, and dependency parsing is done using Stanford Parser (Chen and Manning, 2014) with universal dependencies.

the corpus. For example, if the support of t_i is $S(t_i) = \{(e_1, e_1'), (e_1, e_1'), (e_2, e_2'), \dots\}$, entity pair (e_1, e_1') has a multiplicity of 2 because the relational fact (e_1, t_i, e_1') occur in two sentences. We also get a set of KB relations $\mathcal{R} = \{r_j\}$, and the support $S(r_j)$ of a KB relation r_j is the set of entity pairs having this relation in the KB, i.e., there is a relational fact (e, r_j, e') in the KB. The number of *co-occurrences* of a textural relation t_i and a KB relation r_j is

$$n_{ij} = \sum_{(e,e') \in S(r_j)} m_{S(t_i)}(e, e'), \qquad (1)$$

i.e., every occurrence of relational fact (e, t_i, e') is counted as a co-occurrence of t_i and r_j if $(e, e') \in S(r_j)$. A bipartite relation graph can then be constructed, with $\mathcal{T}$ and $\mathcal{R}$ as the node sets, and the edge between t_i and r_j has weight n_{ij} (no edge if $n_{ij} = 0$), which will be normalized later.

3.2 Normalization

The raw co-occurrence counts have a heavily skewed distribution that spans several orders of magnitude: A small portion of relation pairs co-occur highly frequently, while most relation pairs co-occur only a few times. For example, a textual relation, SUBJECT $\xleftarrow{\text{nsubjpass}}$ *born* $\xrightarrow{\text{nmod:in}}$ OBJECT, may co-occur with the KB relation `place_of_birth` thousands of times (e.g., "*Michelle Obama was born in Chicago*"), while a synonymous but slightly more compositional textual relation, SUBJECT $\xleftarrow{\text{nsubjpass}}$ *born* $\xrightarrow{\text{nmod:in}}$ *city* $\xrightarrow{\text{nmod:of}}$ OBJECT, may only co-occur with the same KB relation a few times in the entire corpus (e.g., "*Michelle Obama was born in the city of Chicago*"). Learning directly on the raw co-occurrence counts, an embedding model may put a disproportionate amount of weight on the most frequent relations, and may not learn well on the majority of rarer relations. Proper normalization is therefore necessary, which will encourage the embedding model to learn good embedding not only for the most frequent relations, but also for the rarer relations.

A number of normalization strategies have been proposed in the context of word embedding, including correlation- and entropy-based normalization (Rohde et al., 2005), positive pointwise mutual information (PPMI) (Bullinaria and Levy, 2007), and some square root type transformation (Lebret and Collobert, 2014). A shared goal is to reduce the impact of the most frequent words, e.g., "the" and "is," which tend to be less informative for the purpose of embedding.

We have experimented with a number of normalization strategies and found that the following strategy works best for textual relation embedding: For each textual relation, we normalize its co-occurrence counts to form a probability distribution over KB relations. The new edge weights of the relation graph thus become $w_{ij} = \tilde{p}(r_j|t_i) = n_{ij} / \sum_{j'} n_{ij'}$. Every textual relation is now associated with a set of edges whose weights sum up to 1. We also experimented with PPMI and smoothed PPMI with $\alpha = 0.75$ (Levy et al., 2015) that are commonly used in word embedding. However, the learned textual relation embedding turned out to be not very helpful for relation extraction. One possible reason is that PPMI (even the smoothed version) gives inappropriately large weights to rare relations (Levy et al., 2015). There are many textual relations that correspond to none of the target KB relations but are falsely labeled with some KB relations a few times by distant supervision. PPMI gives large weights to such falsely labeled cases because it thinks these events have a chance significantly higher than random.

4 Textual Relation Embedding

Next we discuss how to learn embedding of textual relations based on the constructed relation graph. We call our approach **Glo**bal **R**elation **E**mbedding (GloRE) in light of global statistics of relations.

4.1 Embedding via RNN

Given the relation graph, a straightforward way of relation embedding is matrix factorization, similar to latent semantic analysis (Deerwester et al., 1990) for word embedding. However, textual relations are different from words in that they are sequences composed of words and typed dependency relations. Therefore, we use recurrent neural networks (RNNs) for embedding, which respect the compositionality of textual relations and can learn the shared sub-structures of different textual relations (Toutanova et al., 2015). For the examples in Figure 1, an RNN can learn, from both textual relations, that the shared dependency relation "nmod:in" is indicative of location modifiers. It is worth noting that other models like convolutional neural networks can also be used, but it is not the focus of this paper to compare all the alternative embedding models; rather, we aim to show

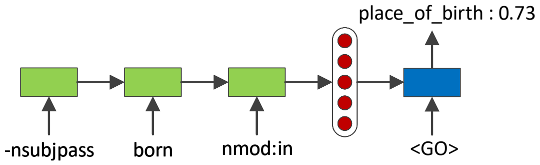


Figure 3: Embedding model. *Left*: A RNN with GRU for embedding. *Middle*: embedding of textual relation. *Right*: a separate GRU cell to map a textual relation embedding to a probability distribution over KB relations.

the effectiveness of global statistics with a reasonable embedding model.

For a textual relation, we first decompose it into a sequence of tokens $\{x_1, ..., x_m\}$, which includes lexical words and directional dependency relations. For example, the textual relation SUBJECT $\xleftarrow{\text{nsubjpass}}$ *born* $\xrightarrow{\text{nmod:in}}$ OBJECT is decomposed to a sequence of three tokens $\{$−nsubjpass, born, nmod:in$\}$, where "−" represents a left arrow. Note that we include directional dependency relations, because both the relation type and the direction are critical in determining the meaning of a textual relation. For example, the dependency relation "nmod:in" often indicates a location modifier and is thus strongly associated with location-related KB relations like `place_of_birth`. The direction also plays an important role. Without knowing the direction of the dependency relations, it is impossible to distinguish `child_of` and `parent_of`.

An RNN with gated recurrent units (GRUs) (Cho et al., 2014) is then applied to consecutively process the sequence as shown in Figure 3. We have also explored more advanced constructs like attention, but the results are similar, so we opt for a vanilla RNN in consideration of model simplicity.

Let ϕ denote the function that maps a token x_l to a fixed-dimensional vector, the hidden state vectors of the RNN are calculated recursively:

$$\boldsymbol{h}_l = \text{GRU}\big(\phi(x_l), \boldsymbol{h}_{l-1}\big). \tag{2}$$

GRU follows the definition in Cho et al. (2014).

4.2 Training Objective

We use global statistics in the relation graph to train the embedding model. Specifically, we model the semantics of a textual relation as its co-occurrence distribution of KB relations, and learn textual relation embedding to reconstruct the corresponding co-occurrence distributions.

We use a separate GRU cell followed by softmax to map a textual relation embedding to a distribution over KB relations; the full model thus resembles the sequence-to-sequence architecture (Sutskever et al., 2014). Given a textual relation t_i and its embedding $\boldsymbol{h}_m$, the predicted conditional probability of a KB relation r_j is thus:

$$p(r_j|t_i) = \text{softmax}(\text{GRU}(\phi(\text{<GO>}), \boldsymbol{h}_m))_j, \tag{3}$$

where $()_j$ denotes the j-th element of a vector, and <GO>is a special token indicating the start of decoding. The training objective is to minimize

$$\Theta = \frac{1}{|\mathcal{E}|} \sum_{i,j:\tilde{p}(r_j|t_i)>0} (\log p(r_j|t_i) - \log \tilde{p}(r_j|t_i))^2, \tag{4}$$

where $\mathcal{E}$ is the edge set of the relation graph. It is modeled as a regression problem, similar to GloVe (Pennington et al., 2014).

Baseline. We also define a baseline approach where the unnormalized co-occurrence counts are directly used. The objective is to maximize:

$$\Theta' = \frac{1}{\sum_{i,j} n_{ij}} \sum_{i,j:n_{ij}>0} n_{ij} \log p(r_j|t_i). \tag{5}$$

It also corresponds to local statistics based embedding, i.e., when the embedding model is trained on individual occurrences of relational facts with distant supervision. Therefore, we call it **Lo**cal **R**elation **E**mbedding (LoRE).

5 Augmenting Relation Extraction

Learned from global co-occurrence statistics of relations, our approach provides semantic matching information of textual and KB relations, which is often complementary to the information captured by existing relation extraction models. In this section we discuss how to combine them together to achieve better relation extraction performance.

We follow the setting of distantly supervised relation extraction. Given a text corpus and a KB with relation set $\mathcal{R}$, the goal is to find new relational facts from the text corpus that are not already contained in the KB. More formally, for each entity pair (e, e') and a set of *contextual sentences* C containing this entity pair, a relation extraction model assigns a score $E(z|C)$ to each candidate relational fact $z = (e, r, e'), r \in \mathcal{R}$. On the other hand, our textual relation embedding model works on the sentence level. It assign a score $G(z|s)$ to each contextual sentence s in C as for how well the textual relation t between the entity pair in the sentence matches the KB relation r, i.e., $G(z|s) = p(r|t)$. It poses a challenge to aggregate the sentence-level scores to get a set-level score $G(z|C)$, which can be used to combine with the original score $E(z|C)$ to get a better evaluation of the candidate relational fact.

One straightforward aggregation is max pooling, i.e., only using the largest score $\max_{s \in C} G(z|s)$, similar to the at-least-one strategy used by Zeng et al. (2015). But it will lose the useful signals from those neglected sentences (Lin et al., 2016). Because of the wrong labeling problem, mean pooling is problematic as well. The wrongly labeled contextual sentences tend to make the aggregate scores more evenly distributed and therefore become less informative. The number of contextual sentences positively supporting a relational fact is also an important signal, but is lost in mean pooling.

Instead, we use summation with a trainable *cap*:

$$G(z|C) = \min\left(cap, \sum_{s \in C} G(z|s)\right), \qquad (6)$$

In other words, we additively aggregate the signals from all the contextual sentences, but only to a bounded degree.

We simply use a weighted sum to combine $E(z|C)$ and $G(z|C)$, where the trainable weights will also handle the possibly different scale of scores generated by different models:

$$\tilde{E}(z|C) = w_1 E(z|C) + w_2 G(z|C). \qquad (7)$$

The original score $E(z|C)$ is then replaced by the new score $\tilde{E}(z|C)$. To find the optimal values for w_1, w_2 and *cap*, we define a hinge loss:

$$\Theta_{Merge} = \frac{1}{K} \sum_{k=1}^{K} \max\left\{0, 1 + \tilde{E}(z_k^-) - \tilde{E}(z_k^+)\right\}, \qquad (8)$$

where $\{z_k^+\}_{k=1}^K$ are the true relational facts from the KB, and $\{z_k^-\}_{k=1}^K$ are false relational facts generated by replacing the KB relation in true relational facts with incorrect KB relations.

Data	# of sentences	# of entity pairs	# of relational facts from KB
Train	570,088	291,699	19,429
Test	172,448	96,678	1,950

Table 1: Statistics of the NYT dataset.

6 Experiments

In this experimental study, we show that GloRE can greatly improve the performance of several recent relation extraction models, including the previous best model on a standard dataset.

6.1 Experimental Setup

Dataset. Following the literature (Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016), we use the relation extraction dataset introduced in (Riedel et al., 2010), which was generated by aligning New York Times (NYT) articles with Freebase (Bollacker et al., 2008). Articles from year 2005-2006 are used as training, and articles from 2007 are used as testing. Some statistics are listed in Table 1. There are 53 target KB relations, including a special relation NA indicating that there is no target relation between entities.

We follow the approach described in Section 3 to construct the relation graph from the NYT training data. The constructed relation graph contains 321,447 edges with non-zero weight. We further obtain a training set and a validation set from the edges of the relation graph. We have observed that using a validation set totally disjoint from the training set leads to unstable validation loss, so we randomly sample 300K edges as the training set, and another 60K as the validation set. The two sets can have some overlap. For the merging model (Eq. 8), 10% of the edges are reserved as the validation set.

Relation extraction models. We evaluate with four recent relation extraction models whose source code is publicly available[3]. We use the optimized parameters provided by the authors.

- **CNN+ONE** and **PCNN+ONE** (Zeng et al., 2015): A convolutional neural network (CNN) is used to embed contextual sentences for relation classification. Multi-instance learning with at-least-one (ONE) assumption is used to combat the wrong labeling problem. In PCNN, piecewise max pooling is

[3] https://github.com/thunlp/NRE

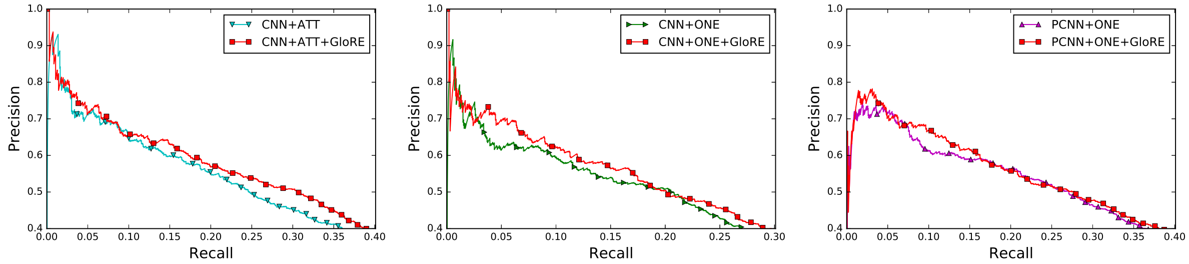


Figure 4: Held-out evaluation: other base relation extraction models and the improved versions when augmented with GloRE.

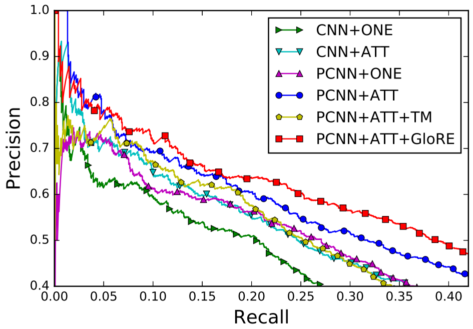


Figure 5: Held-out evaluation: the previous best-performing model can be further improved when augmented with GloRE. PCNN+ATT+TM is a recent model (Luo et al., 2017) whose performance is slightly inferior to PCNN+ATT. Because the source code is not available, we did not experiment to augment this model with GloRE. Another recent method (Wu et al., 2017) incorporates adversarial training to improve PCNN+ATT, but the results are not directly comparable (see Section 2 for discussion). Finally, Ji et al. (2017) propose a model similar to PCNN+ATT, but the performance is inferior to PCNN+ATT and is not shown here for clarity.

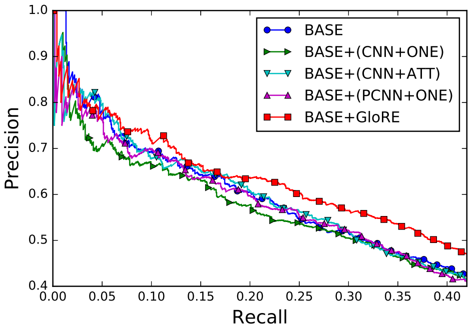


Figure 6: Held-out evaluation: GloRE brings the largest improvement to BASE (PCNN+ATT), which further shows that GloRE captures useful information for relation extraction that is complementary to existing models.

used to handle the three pieces of a contextual sentence (split by the two entities) separately.

- **CNN+ATT** and **PCNN+ATT** (Lin et al., 2016): Different from the at-least-one assumption which loses information in the neglected sentences, these models learn soft attention weights (ATT) over contextual sentences and thus can use the information of all the contextual sentences. *PCNN+ATT is the best-performing model on the NYT dataset.*

Evaluation settings and metrics. Similar to previous work (Riedel et al., 2010; Zeng et al., 2015), we use two settings for evaluation: (1) Held-out evaluation, where a subset of relational facts in KB is held out from training (Table 1), and is later used to compare against newly discovered relational facts. This setting avoids human labor but can introduce some false negatives because of the incompleteness of the KB. (2) Manual evaluation, where the discovered relational facts are manually judged by human experts. For held-out evaluation, we report the precision-recall curve. For manual evaluation, we report $Precision@N$, i.e., the precision of the top N discovered relational facts.

Implementation. Hyper-parameters of our model are selected based on the validation set. For the embedding model, the mini-batch size is set to 128, and the state size of the GRU cells is 300. For the merging model, the mini-batch size is set to 1024. We use Adam with parameters recommended by the authors for optimization. Word embeddings are initialized with the 300-dimensional word2vec vectors pre-trained on the Google News corpus[4]. Early stopping based on the validation set is employed. Our model is implemented using Tensorflow (Abadi et al., 2016), and the source code is available at `https://github.com/ppuliu/GloRE`.

[4] `https://code.google.com/archive/p/word2vec/`

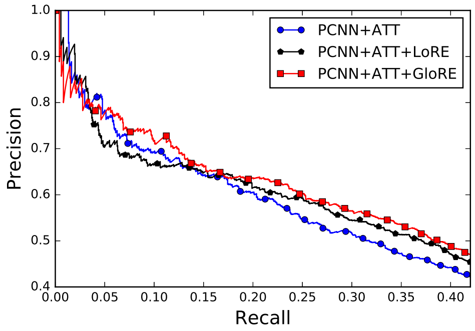


Figure 7: Held-out evaluation: LoRE vs. GloRE.

6.2 Held-out Evaluation

Existing Models + GloRE. We first show that our approach, GloRE, can improve the performance of the previous best-performing model, PCNN+ATT, leading to a new state of the art on the NYT dataset. As shown in Figure 5, when PCNN+ATT is augmented with GloRE, a consistent improvement along the precision-recall curve is observed. It is worth noting that although PCNN+ATT+GloRE seems to be inferior to PCNN+ATT when recall < 0.05, as we will show via manual evaluation, it is actually due to false negatives.

We also show in Figure 4 that the improvement brought by GloRE is general and not specific to PCNN+ATT; the other models also get a consistent improvement when augmented with GloRE.

To investigate whether the improvement brought by GloRE is simply from ensemble, we also augment PCNN+ATT with the other three base models in the same way as described in Section 5. The results in Figure 6 show that pairwise ensemble of existing relation extraction models does not yield much improvement, and GloRE brings much larger improvement than the other models.

In summary, the held-out evaluation results suggest that GloRE captures useful information for relation extraction that is not captured by these local statistics based models.

LoRE v.s. GloRE. We compare GloRE with the baseline approach LoRE (Section 4) to show the advantage of normalization on global statistics. We use PCNN+ATT as the base relation extraction model. As shown in Figure 7, GloRE consistently outperforms LoRE. It is worth noting that LoRE can still improve the base relation extraction model when recall > 0.15, further confirming the usefulness of directly embedding textual relations in addition to sentences.

Precision@N	100	300	500	700	900	1000
PCNN+ATT	**97.0**	93.7	92.8	89.1	85.2	83.9
PCNN+ATT+LoRE	**97.0**	95.0	94.2	91.6	89.6	87.0
PCNN+ATT+GloRE	**97.0**	**97.3**	**94.6**	**93.3**	**90.1**	**89.3**

Table 2: Manual evaluation: false negatives from held-out evaluation are manually corrected by human experts.

6.3 Manual Evaluation

Due to the incompleteness of the knowledge base, held-out evaluation introduces some false negatives. The precision from held-out evaluation is therefore a lower bound of the true precision. To get a more accurate evaluation of model performance, we have human experts to manually check the false relational facts judged by held-out evaluation in the top 1,000 predictions of three models, PCNN+ATT, PCNN+ATT+LoRE and PCNN+ATT+GloRE, and report the corrected results in Table 2. Each prediction is examined by two human experts who reach agreement with discussion. To ensure fair comparison, the experts are not aware of the provenance of the predictions. Under manual evaluation, PCNN+ATT+GloRE achieves the best performance in the full range of N. In particular, for the top 1,000 predictions, GloRE improves the precision of the previous best model PCNN+ATT from 83.9% to 89.3%. The manual evaluation results reinforce the previous observations from held-out evaluation.

6.4 Case Study

Table 3 shows two examples. For better illustration, we choose entity pairs that have only one contextual sentence.

For the first example, PCNN+ATT predicts that most likely there is no KB relation between the entity pair, while both LoRE and GloRE identify the correct relation with high confidence. The textual relation clearly indicates that the head entity is (appos) a criminologist at (nmod:at) the tail entity.

For the second example, there is no KB relation between the entity pair, and PCNN+ATT is indeed able to rank NA at the top. However, it is still quite confused by `nationality`, probably because it has learned that sentences about a person and a country with many words about profession ("poet," "playwright," and "novelist")

Contextual Sentence	Textual Relation	PCNN+ATT Predictions	LoRE Predictions	GloRE Predictions
[**Alfred Blumstein**]$_{head}$, a criminologist at [**Carnegie Mellon University**]$_{tail}$, called ...	$\xleftarrow{\text{appos}}$ criminologist $\xrightarrow{\text{nmod:at}}$	NA (0.63) **employee_of** (0.36) founder_of (0.00)	**employee_of** (1.00) NA (0.00) founder_of (0.00)	**employee_of** (0.96) NA (0.02) founder_of (0.02)
[**Langston Hughes**]$_{head}$, the American poet, playwright and novelist, came to [**Spain**]$_{tail}$ to ...	$\xleftarrow{\text{-nsubj}}$ came $\xrightarrow{\text{to}}$	**NA** (0.58) nationality (0.38) place_lived (0.01)	place_of_death (0.35) **NA** (0.33) nationality (0.21)	**NA** (0.73) contain_location (0.07) employee_of (0.06)

Table 3: Case studies. We select entity pairs that have only one contextual sentence, and the head and tail entities are marked. The top 3 predictions from each model with the associated probabilities are listed, with the correct relation bold-faced.

likely express the person's nationality. As a result, its prediction on NA is not very confident. On the other hand, GloRE learns that if a person "came to" a place, likely it is not his/her birthplace. In the training data, due to the wrong labeling problem of distant supervision, the textual relation is wrongly labeled with `place_of_death` and `nationality` a couple of times, and both PCNN+ATT and LoRE suffer from the training noise. Taking advantage of global statistics, GloRE is more robust to such noise introduced by the wrong labeling problem.

7 Conclusion

Our results show that textual relation embedding trained on global co-occurrence statistics captures useful relational information that is often complementary to existing methods. As a result, it can greatly improve existing relation extraction models. Large-scale training data of embedding can be easily solicited from distant supervision, and the global statistics of relations provide a natural way to combat the wrong labeling problem of distant supervision.

The idea of relation embedding based on global statistics can be further expanded along several directions. In this work we have focused on embedding textual relations, but it is in principle beneficial to jointly embed knowledge base relations and optionally entities. Recently a joint embedding approach has been attempted in the context of knowledge base completion (Toutanova et al., 2015), but it is still based on local statistics, i.e., individual relational facts. Joint embedding with global statistics remains an open problem. Compared with the size of the training corpora for word embedding (up to hundred of billions of tokens), the NYT dataset is quite small in scale. Another interesting venue for future research is to construct much larger-scale distant supervision datasets to train general-purpose textual relation embedding that can help a wide range of downstream relational tasks such as question answering and textual entailment.

Acknowledgements

The authors would like to thank the anonymous reviewers for their thoughtful comments. This research was sponsored in part by the Army Research Laboratory under cooperative agreements W911NF09-2-0053 and NSF IIS 1528175. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467* .

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International conference on Management of data*. ACM, pages 1247–1250.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods* 39(3):510–526.

Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 724–731.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 740–750.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078* .

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89(1):31–71.

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

Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL on Interactive poster and demonstration sessions*. Association for Computational Linguistics.

Rémi Lebret and Ronan Collobert. 2014. Word embeddings through hellinger PCA. *European Chapter of the Association for Computational Linguistics* .

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *Acm Sigkdd Explorations Newsletter* 17(2):1–16.

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

Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, and Jiawei Han. 2017. Heterogeneous supervision for relation extraction: A representation learning approach. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Yang Liu, Sujian Li, Furu Wei, and Heng Ji. 2016. Relation classification via modeling augmented dependency paths. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24(9):1585–1594.

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. pages 430–439.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. pages 3111–3119.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1532–1543.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 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 the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Douglas LT Rohde, Laura M Gonnerman, and David C Plaut. 2005. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM* 8:627–633.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1201–1211.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. pages 3104–3112.

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 *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1499–1509.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015a. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv:1506.07650* .

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv:1601.03651* .

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015b. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1785–1794.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research* 3(Feb):1083–1106.

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

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the International Conference on Computational Linguistics*. pages 2335–2344.

Min Zhang, Jie Zhang, and Jian Su. 2006. Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, pages 288–295.

Zhou Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 427–434.