Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach

Liyuan Liu†* Xiang Ren†* Qi Zhu† Huan Gui† Shi Zhi† Heng Ji‡ Jiawei Han†

† University of Illinois at Urbana-Champaign, Urbana, IL, USA
‡ Computer Science Department, Rensselaer Polytechnic Institute, USA
†{ll2, xren7, qiz3, huangui2, shizhi2, hanj}@illinois.edu ‡{jih} @rpi.edu

Abstract
Relation extraction is a fundamental task in information extraction. Most existing methods have heavy reliance on annotations labeled by human experts, which are costly and time-consuming. To overcome this drawback, we propose a novel framework, REHENSION, to conduct relation extractor learning using annotations from heterogeneous information source, e.g., knowledge base and domain heuristics. These annotations, referred as heterogeneous supervision, often conflict with each other, which brings a new challenge to the original relation extraction task: how to infer the true label from noisy labels for a given instance. Identifying context information as the backbone of both relation extraction and true label discovery, we adopt embedding techniques to learn the distributed representations of context, which bridges all components with mutual enhancement in an iterative fashion. Extensive experimental results demonstrate the superiority of REHENSION over the state-of-the-art.

1 Introduction
One of the most important tasks towards text understanding is to detect and categorize semantic relations between two entities in a given context. For example, in Fig. 1, with regard to the sentence of $c_1$, relation between Jesse James and Missouri should be categorized as died_in. With accurate identification, relation extraction systems can provide essential support for many applications. One example is question answering, regarding a specific question, relation among entities can provide valuable information, which helps to seek better answers (Bao et al., 2014). Similarly, for medical science literature, relations like protein-protein interactions (Fundel et al., 2007) and gene disease associations (Chun et al., 2006) can be extracted and used in knowledge base population. Additionally, relation extractors can be used in ontology construction (Schutz and Buitelaar, 2005).

Typically, existing methods follow the supervised learning paradigm, and require extensive annotations from domain experts, which are costly and time-consuming. To alleviate such drawback, attempts have been made to build relation extractors with a small set of seed instances or human-crafted patterns (Nakashole et al., 2012, 2011; Carlson et al., 2010), based on which more patterns and instances will be iteratively generated by bootstrap learning. However, these methods often suffer from semantic drift (Mintz et al., 2009). Besides, knowledge bases like Freebase have been leveraged to automatically generate training data and provide distant supervision (Mintz et al., 2009). Nevertheless, for many domain-specific applications, distant supervision is either non-existent or insufficient (usually less than 25% of relation mentions are covered (Ren et al., 2015; Lin et al., 2012)).

Only recently have preliminary studies been developed to unite different supervisions, including knowledge bases and domain specific patterns, which are referred as heterogeneous supervision. As shown in Fig. 1, these supervisions often conflict with each other (Ratner et al., 2016). To address these conflicts, data programming (Ratner et al., 2016) employs a generative model, which encodes supervisions as labeling functions, and adopts the source consistency assumption: a source is likely to provide true information with
the same probability for all instances. This assumption is widely used in true label discovery literature (Li et al., 2016) to model reliabilities of information sources like crowdsourcing and infer the true label from noisy labels. Accordingly, most true label discovery methods would trust a human annotator on all instances to the same level.

However, labeling functions, unlike human annotators, do not make casual mistakes but follow certain “error routine”. Thus, the reliability of a labeling function is not consistent among different pieces of instances. In particular, a labeling function could be more reliable for a certain subset (Varma et al., 2016) (also known as its proficient subset) comparing to the rest. We identify these proficient subsets based on context information, only trust labeling functions on these subsets and avoid assuming global source consistency.

Meanwhile, embedding methods have demonstrated great potential in capturing semantic meanings, which also reduce the dimension of overwhelming text features. Here, we present REHENSION, a novel framework capturing context’s semantic meaning through representation learning, and conduct both relation extraction and true label discovery in a context-aware manner. Specifically, as depicted in Fig. 1, we embed relation mentions in a low-dimension vector space, where similar relation mentions tend to have similar relation types and annotations. ‘True’ labels are further inferred based on reliabilities of labeling functions, which are calculated with their proficient subsets’ representations. Then, these inferred true labels would serve as supervision for all components, including context representation, true label discovery and relation extraction. Besides, the context representation bridges relation extraction with true label discovery, and allows them to enhance each other.

To the best of our knowledge, the framework proposed here is the first method that utilizes representation learning to provide heterogeneous supervision for relation extraction. The high-quality context representations serve as the backbone of true label discovery and relation extraction. Extensive experiments on benchmark datasets demonstrate significant improvements over the state-of-the-art.

The remaining of this paper is organized as follows. Section 2 gives the definition of relation extraction with heterogeneous supervision. We then present the REHENSION model and the learning algorithm in Section 3, and report our experimental evaluation in Section 4. Finally, we briefly survey related work in Section 5 and conclude this study in Section 6.

2 Preliminaries

In this section, we would formally define relation extraction and heterogeneous supervision.

Relation Extraction. For a sentence \(d\), an entity mention is a token span in \(d\) which represents an entity, and a relation mention is a triple \((e_1, e_2, d)\) which consists of an ordered entity pair \((e_1, e_2)\) and \(d\). The relation extraction task is to categorize relation mentions into a given set of relation types \(R\), or Not-Target-Type (None) which means the type of the relation mention does not belong to \(R\).

Heterogeneous Supervision. Similar to (Ratner et al., 2016), we employ labeling functions as basic units to encode supervision information and generate annotations. Since different supervision information may have different proficient subsets, we require each labeling function to encode only one elementary supervision information. Specifically, in the relation extraction scenario, we require each labeling function to only annotate one

![Figure 1: REHENSION Framework except Extraction and Representation of Text Features](image-url)
text feature embedding for \( f_i \in \mathcal{F} \)

\( v_i \) text feature embedding for \( f_i \in \mathcal{F} \)

\( z_i \) relation mention embedding for \( c \in \mathcal{C} \)

\( \lambda_i \) embedding for \( \lambda_i \)'s proficient subset, \( \lambda_i \in \Lambda \)

\( o_{c,i} \) annotation for \( c \), generated by labeling function \( \lambda_i \)

\( o_c^* \) underlying true label for \( c \)

\( \rho_{c,i} \) identify whether \( o_{c,i} \) is correct

\( S_c \) the proficient subset of labeling function \( \lambda_i \)

\( c_{j,i} \) identity whether \( c \) belongs to \( \lambda_i \)'s proficient subset

\( r_i \) relation type embedding for \( r_i \in \mathcal{R} \)

**Table 1: Notation Table.**

| Function | Description |
|----------|-------------|
| \( f_i \) | \( c \)'s text features set, where \( c \in \mathcal{C} \) |
| \( v_i \) | text feature embedding for \( f_i \in \mathcal{F} \) |
| \( z_i \) | relation mention embedding for \( c \in \mathcal{C} \) |
| \( \lambda_i \) | embedding for \( \lambda_i \)'s proficient subset, \( \lambda_i \in \Lambda \) |
| \( o_{c,i} \) | annotation for \( c \), generated by labeling function \( \lambda_i \) |
| \( o_c^* \) | underlying true label for \( c \) |
| \( \rho_{c,i} \) | identify whether \( o_{c,i} \) is correct |
| \( S_c \) | the proficient subset of labeling function \( \lambda_i \) |
| \( c_{j,i} \) | identity whether \( c \) belongs to \( \lambda_i \)'s proficient subset |
| \( r_i \) | relation type embedding for \( r_i \in \mathcal{R} \) |

For a POS-tagged corpus \( D \) with detected entities, we refer its relation mentions as \( \mathcal{C} = \{ c_1 = (e_{1,1}, e_{1,2}, d), \forall d \in D \} \). Our goal is to annotate entity mentions with relation types of interest \( (\mathcal{R} = \{ r_1, \ldots, r_K \}) \) or None. We require users to provide heterogeneous supervision in the form of labeling function \( \Lambda = \{ \lambda_1, \ldots, \lambda_M \} \), and mark the annotations generated by \( \Lambda \) as \( \mathcal{O} = \{ o_{c,i} | \lambda_i \text{ generate annotation } o_{c,i} \text{ for } c \in \mathcal{C} \} \). We record relation mentions annotated by \( \Lambda \) as \( C_1 \), and refer relation mentions without annotation as \( C_0 \). Then, our task is to train a relation extractor based on \( C_1 \) and categorize relation mentions in \( C_0 \).

### 3 The REHesSion Framework

Here, we present REHesSion, a novel framework to infer true labels from automatically generated noisy labels, and categorize unlabeled instances into a set of relation types. Intuitively, errors of annotations \( ( \mathcal{O} ) \) come from mismatch of contexts, e.g., in Fig. 1, \( \lambda_1 \) annotates \( c_1 \) and \( c_2 \) with ‘true’ labels but for mismatched contexts ‘killing’ and ‘killed’. Accordingly, we should only trust labeling functions on matched context, e.g., trust \( \lambda_1 \) on \( c_2 \) due to its context ‘was born in’, but not on \( c_1 \) and \( c_2 \). On the other hand, relation extraction can be viewed as matching appropriate relation type to a certain context. These two matching processes are closely related and can enhance each other, while context representation plays an important role in both of them.

**Framework Overview.** We propose a general framework to learn the relation extractor from automatically generated noisy labels. As plotted in Fig. 1, distributed representation of context bridges relation extraction with true label discovery, and allows them to enhance each other. Specifically, it follows the steps below:

1. After being extracted from context, text features are embedded in a low dimension space by representation learning (see Fig. 2);
2. Text feature embeddings are utilized to calculate relation mentions embeddings (see Fig. 2);
3. With relation mention embeddings, true labels are inferred by calculating labeling functions’ reliabilities in a context-aware manner (see Fig. 1);
4. Inferred true labels would ‘supervise’ all components to learn model parameters (see Fig. 1).

We now proceed by introducing these components of the model in further details.

#### 3.1 Modeling Relation Mention

As shown in Table 2, we extract abundant lexical features (Ren et al., 2016; Mintz et al., 2009; Chan and Roth, 2010) to characterize relation mentions. However, this abundance also results in the gigantic dimension of original text features (~10^7 in our case). In order to achieve better generalization ability, we represent relation mentions with low dimensional (~10^3) vectors. In Fig. 2, for example, relation mention \( c_3 \) is first represented as bag-of-features. After learning text feature embeddings, we use the average of feature embedding vectors to derive the embedding vector for \( c_3 \).

**Text Feature Representation.** Similar to other principles of embedding learning, we assume text features occurring in the same contexts tend to have similar meanings (also known as distributional hypothesis(Harris, 1954)). Furthermore, we let each text feature’s embedding vector to predict other text features occurred in the same relation mentions or context. Thus, text features with similar meaning should have similar embedding vectors. Formally, we mark text features as \( \mathcal{F} = \{ f_1, \ldots, f_{|\mathcal{F}|} \} \), record the feature set for \( \forall c \in \mathcal{C} \) as \( \mathcal{F}_c \), and represent the embedding vector for \( f_i \) as \( v_i \in \mathbb{R}^{n_{v_i}} \), and we aim to maximize the following
log likelihood: $\sum_{c \in C} \sum_{f_i, f_j \in E_c} \log p(f_i | f_j)$, where $p(f_i | f_j) = \exp(v_i^T v_j) / \sum_{f_k \in F} \exp(v_i^T v_k)$.

However, the optimization of this likelihood is impractical because the calculation of $\nabla p(f_i | f_j)$ requires summation over all text features, whose size exceeds $10^7$ in our case. In order to perform efficient optimization, we adopt the negative sampling technique (Mikolov et al., 2013) to avoid this summation. Accordingly, we replace the log likelihood with Eq. 1 as below:

$$J_p = \sum_{c \in C} \left( \log \sigma(v_i^T v_j) - \frac{1}{|E_c|} \sum_{k=1}^{|E_c|} \log \sigma(-v_i^T v_k) \right)$$

(1)

where $\hat{P}$ is noise distribution used in (Mikolov et al., 2013), $\sigma$ is the sigmoid function and $V$ is number of negative samples.

**Relation Mention Representation.** With text feature embeddings learned by Eq. 1, a naive method to represent relation mentions is to concatenate or average its text feature embeddings. However, text features embedding may be in a different semantic space with relation types. Thus, we directly learn a mapping $g$ from text feature representations to relation mention representations (Van Gysel et al., 2016a,b) instead of simple heuristic rules like concatenate or average (see Fig. 2):

$$z_c = g(f_c) = \tanh(W : \frac{1}{|E_c|} \sum_{f_i \in E_c} v_i)$$

(2)

where $z_c$ is the representation of $c \in C$, $W$ is a $n_z \times n_{vi}$ matrix, $n_z$ is the dimension of relation mention embeddings and $\tanh$ is the element-wise hyperbolic tangent function.

In other words, we represent bag of text features with their average embedding, then apply linear map and hyperbolic tangent to transform the embedding from text feature semantic space to relation mention semantic space. The non-linear $\tanh$ function allows non-linear class boundaries in other components, and also regularize relation mention representation to range $[-1, 1]$ which avoids numerical instability issues.

### 3.2 True Label Discovery

Because heterogeneous supervision generates labels in a discriminative way, we suppose its errors follow certain underlying principles, i.e., if a labeling function annotates an instance correctly / wrongly, it would annotate other similar instances correctly / wrongly. For example, $\lambda_1$ in Fig. 1 generates wrong annotations for two similar instances $c_1$, $c_2$ and would make the same errors on other similar instances. Since context representation captures the semantic meaning of relation mention and would be used to identify relation types, we also use it to identify the mismatch of context and labeling functions. Thus, we suppose for each labeling function $\lambda_i$, there exists a proficient subset $S_i$ on $R^{ntz}$, containing instances that $\lambda_i$ can precisely annotate. In Fig. 1, for instance, $c_3$ is in the proficient subset of $\lambda_1$, while $c_1$ and $c_2$ are not. Moreover, the generation of annotations are not really random, and we propose a probabilistic model to describe the level of mismatch from labeling functions to real relation types instead of annotations’ generation.

As shown in Fig. 3, we assume the indicator of whether $c$ belongs to $S_i$, $s_{c,i} = \delta(c \in S_i)$, would first be generated based on context representation

$$p(s_{c,i} = 1 | z_c, 1) = p(c \in S_i) = \sigma(z_i^T 1)$$

(3)

Then the correctness of annotation $o_{c,i}$, $\rho_{c,i} = \delta(o_{c,i} = o^*_i)$, would be generated. Furthermore, we assume $p(\rho_{c,i} = 1 | s_{c,i} = 1) = \phi_1$ and
\[
p(p_{c,i} = 1 | s_{c,i} = 0) = \phi_0 \text{ to be constant for all relation mentions and labeling functions.}
\]

Because \( s_{c,i} \) would not be used in other components of our framework, we integrate out \( s_{c,i} \) and write the log likelihood as

\[
\mathcal{J}_T = \sum_{s_{c,i} \in \mathcal{C}} \log(\sigma(z_i^T \mathbf{1}_L)) \phi_0^{\delta(o_{c,i} = o_{c}^*)} (1 - \phi_0)^{\delta(o_{c,i} \neq o_{c}^*)} + (1 - \sigma(z_i^T \mathbf{1}_L)) \phi_0^{\delta(o_{c,i} = o_{c}^*)} (1 - \phi_0)^{\delta(o_{c,i} \neq o_{c}^*)}
\]

(4)

Note that \( o_{c}^* \) is a hidden variable but not a model parameter, \( \mathcal{J}_T \) is the likelihood of \( \rho_{c,i} = \delta(o_{c,i} = o_{c}^*) \). Thus, we would first infer \( o_{c}^* = \text{argmax}_{o_{c}^*} \mathcal{J}_T \), then train the true label discovery model by maximizing \( \mathcal{J}_T \).

### 3.3 Modeling Relation Type

We now discuss the model for identifying relation types based on context representation. For each relation mention \( c \), its representation \( z_c \) implies its relation type, and the distribution of relation type can be described by the soft-max function:

\[
p(r_i | z_c) = \frac{\exp(z_i^T t_k)}{\sum_{r_j \in \mathcal{R}, (j \neq i)} \exp(z_j^T t_k)}
\]

(5)

where \( t_k \in \mathbb{R}^v \) is the representation for relation type \( r_k \). Moreover, with the inferred true label \( o_{c}^* \), the relation extraction model can be trained as a multi-class classifier. Specifically, we use Eq. 5 to approach the distribution

\[
p(r_i | o_{c}^*) = \begin{cases} 1 & r_i = o_{c}^* \\ 0 & r_i \neq o_{c}^* \end{cases}
\]

(6)

Moreover, we use KL-divergence to measure the dissimilarity between two distributions, and formulate model learning by maximizing \( \mathcal{J}_R \) in Eq. 7, where \( KL(p(z_i | p(o_{c}^*))) \) is the KL-divergence from \( p(r_i | o_{c}^*) \) to \( p(r_i | z_c) \) and \( p(r_i | o_{c}^*) \) has the form of Eq. 5 and Eq. 6.

\[
\mathcal{J}_R = - \sum_{c \in \mathcal{C}} KL(p(z_c | p(o_{c}^*)))
\]

(7)

### 3.4 Model Learning

Based on Eq. 1, Eq. 4 and Eq. 7, we form the joint optimization problem for model parameters as

\[
\min_{w, v, \nu^*, \lambda, \nu, \lambda, \phi_0} \mathcal{J} = - \mathcal{J}_R - \lambda_1 \mathcal{J}_E - \lambda_2 \mathcal{J}_T
\]

s.t. \( \forall c \in \mathcal{C}, o_{c}^* = \text{argmax}_{o_{c}^*} \mathcal{J}_T, z_c = g(\mathbf{E}) \)

(8)

Collectively optimizing Eq. 8 allows heterogeneous supervision guiding all three components, while these components would refine the context representation, and enhance each other.

### 3.5 Relation Type Inference

We now discuss the strategy of performing type inference for \( C_u \). As shown in Table 3, the proportion of \( \text{None} \) in \( C_u \) is usually much larger than in \( C_t \). Additionally, not like other relation types in \( \mathcal{R} \), \( \text{None} \) does not have a coherent semantic meaning. Similar to (Ren et al., 2016), we introduce a heuristic rule: identifying a relation mention as \( \text{None} \) when (1) our relation extractor predict it as \( \text{None} \), or (2) the entropy of \( p(z_c) \) over \( \mathcal{R} \) exceeds a pre-defined threshold \( \eta \). The entropy is calculated as \( H(p(z_c)) = -\sum_{r_i \in \mathcal{R}} p(r_i | z_c) \log(p(r_i | z_c)) \).

And the second situation means based on relation extractor this relation mention is not likely belonging to any relation types in \( \mathcal{R} \).

### 4 Experiments

In this section, we empirically validate our method by comparing to the state-of-the-art relation extraction methods on news and Wikipedia articles.

#### 4.1 Datasets and settings

In the experiments, we conduct investigations on two benchmark datasets from different domains: \(^1\) \text{NYT} (Riedel et al., 2010) is a news corpus sampled from \( \sim 294k \) 1989-2007 New York Times news articles. It consists of 1.18M sentences, while 395 of

| Datasets | NYT | Wiki-KBP |
|----------|-----|----------|
| % of None in Training | 0.6717 | 0.5552 |
| % of None in Test | 0.8972 | 0.8532 |

Table 3: Proportion of \( \text{None} \) in Training/Test Set

In order to solve the joint optimization problem in Eq. 8 efficiently, we adopt the stochastic gradient descent algorithm to update \( \{W, v, \nu^*, 1, t\} \) iteratively, and \( o_{c}^* \) is estimated by maximizing \( \mathcal{J}_T \) after calculating \( z_c \). Additionally, we apply the widely used dropout techniques (Srivastava et al., 2014) to prevent overfitting and improve generalization performance.

The learning process of REHESSION is summarized as below. In each iteration, we would sample a relation mention \( c \) from \( C_t \), then sample \( c \)'s text features and conduct the text features’ representation learning. After calculating the representation of \( c \), we would infer its true label \( o_{c}^* \) based on our true label discovery model, and finally update model parameters based on \( o_{c}^* \).

---

\(^1\) Codes and datasets used in this paper can be downloaded at: https://github.com/.
For both datasets, the training and test sets partitions are maintained in our experiments. Furthermore, we create validation sets by randomly sampling 10% mentions from each test set and used the remaining part as evaluation sets.

**Feature Generation.** As summarized in Table 2, we use a 6-word window to extract context features for each entity mention, apply the Stanford CoreNLP tool (Manning et al., 2014) to generate entity mentions and get POS tags for both datasets. Brown clusters (Brown et al., 1992) are derived for each corpus using public implementation. All these features are shared with all compared methods in our experiments.

**Labeling Functions.** In our experiments, labeling functions are employed to encode two kinds of supervision information. One is knowledge base, the other is handcrafted domain-specific patterns. For domain-specific patterns, we manually design a number of labeling functions; for knowledge base, annotations are generated following the procedure in (Ren et al., 2016; Riedel et al., 2010).

Regarding two kinds of supervision information, the statistics of the labeling functions are summarized in Table 4. We can observe that heuristic patterns can identify more relation types for KBP datasets, while for NYT datasets, knowledge base can provide supervision for more relation types. This observation aligns with our intuition that single kind of information might be insufficient while different kinds of information can complement each other.

We further summarize the statistics of annotations in Table 5. It can be observed that a large portion of instances is only annotated as None, while lots of conflicts exist among other instances. This phenomenon justifies the motivation to employ true label discovery model to resolve the conflicts among supervision. Also, we can observe most conflicts involve None type, accordingly, our proposed method should have more advantages over traditional true label discovery methods on the relation extraction task comparing to the relation classification task that excludes None type.

### 4.2 Compared Methods
We compare REHession with below methods:

**FIGER** (Ling and Weld, 2012) adopts multi-label learning with Perceptron algorithm.

**BFK** (Bunescu and Mooney, 2005) applies bag-of-feature kernel to train a support vector machine;

**DSL** (Mintz et al., 2009) trains a multi-class logistic classifier on the training data;

**Muitik** (Hoffmann et al., 2011) models training label noise by multi-instance multi-label learning;

**FCM** (Gormley et al., 2015) performs compositional embedding by neural language model.

**CoType-RM** (Ren et al., 2016) adopts partial-label loss to handle label noise and train the extractor.

Moreover, two different strategies are adopted to feed heterogeneous supervision to these methods. The first is to keep all noisy labels, marked as ‘NL’. Alternatively, a true label discovery method, Investment (Pasternack and Roth, 2010), is applied to resolve conflicts, which is based on the source consistency assumption and iteratively updates inferred true labels and label functions’ reliabilities. Then, the second strategy is to only feed the inferred true labels, referred as ‘TD’.

**Evaluation Metrics.** For relation classification task, which excludes None type from training / testing, we use the classification accuracy (Acc) for evaluation, and for relation extraction task, precision (Prec), recall (Rec) and F1 score (Bunescu and Mooney, 2005; Bach and Badaskar, 2007) are employed. Note that both relation extraction and relation classification are conducted and evaluated

---

**Table 4:** Number of labeling functions and the relation types they can annotated w.r.t. two kinds of information

| Pattern | Wiki-KBP | NYT |
|---------|----------|-----|
| KB      | #Types | #LF | #Types | #LF |
|         | 13 | 147 | 16 | 115 |
|         | 7 | 75 | 25 | 26 |

**Table 5:** Number of relation mentions (RM), relation mentions annotated as None, relation mentions with conflicting annotations and conflicts involving None

| Dataset | Wiki-KBP | NYT |
|---------|----------|-----|
| Total Number of RM | 225977 | 330767 |
| RM annotated as None | 100521 | 356497 |
| RM with conflicts | 32008 | 58198 |
| Conflicts involving None | 30599 | 38756 |

---

2https://github.com/percyliang/brown-cluster
3labeling functions can be accessed at: https://github.com/
4We use liblinear package from https://github.com/cjlin1/liblinear

---

(4.2)
as discussed in Sec. 4.3. On the other hand, the true labels inferred by Investment are actually yields better performance than TD strategy, since achieving the best performance. This observation accords with the significant on the NYT dataset than on the Wiki-KBP dataset. This observation justifies the motivation of avoiding the source consistency assumption and the effectiveness of proposed true label discovery model.

One could also observe the difference between REHESsion and the compared methods is more significant on the NYT dataset than on the Wiki-KBP dataset. This observation accords with the fact that the NYT dataset contains more conflicts than KBP dataset (see Table 5), and the intuition is that our method would have more advantages on more conflicting labels.

Among four tasks, the relation classification of Wiki-KBP dataset has highest label quality, i.e., conflicting label ratio, but with least number of training instances. And CoType-RM and DSL reach relatively better performance among all compared methods. CoType-RM performs much better than DSL on Wiki-KBP relation classification task, while DSL gets better or similar performance with CoType-RM on other tasks. This may be because the representation learning method is able to generalize better, thus performs better when the training set size is small. However, it is rather vulnerable to the noisy labels compared to DSL. Our method employs embedding techniques, and also integrates context-aware true label discovery to de-noise labels, making the embedding method rather robust, thus achieves the best performance on all tasks.

### 4.4 Case Study

**Context Awareness of True Label Discovery.**
In this paper, we propose REHession, an embedding framework to extract relation under heterogeneous supervision. It resolves conflicts based on context representation and avoid source consistency assumption. Heterogeneous supervision allows our model extract more relation types, while the inferred high-quality true labels allows our model to be more accurate. Experimental evaluation fully justifies the effectiveness of the proposed framework on two real-world datasets.

There exist many directions for future work. One is to apply transfer learning techniques to handle the difference between label distributions of training set and test set. Another is to further...
incorporate information like relation type hierarchy for fine-grained relation extraction.

References

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*.

Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. 2014. Knowledge-based question answering as machine translation. *Cell* 2(6).

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18(4):467–479.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *ACL*.

Razvan Bunescu and Raymond J Mooney. 2005. Subsequence kernels for relation extraction. In *NIPS*. pages 171–178.

Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pages 101–110.

Yee Seng Chan and Dan Roth. 2010. Exploiting background knowledge for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 152–160.

Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shibata, Naoki Nagata, Teruyoshi Hishiki, and Jun’ichi Tsujii. 2006. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Pacific Symposium on Biocomputing*. volume 11, pages 4–15.

Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie Strassel, and Jonathan Wright. 2012. Linguistic resources for 2013 knowledge base population evaluations. In *TAC*.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*. ACM, pages 100–110.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. Relexrelation extraction using dependency parse trees. *Bioinformatics* 23(3):365–371.

Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. *arXiv preprint arXiv:1505.02419*.

Zellig S Harris. 1954. Distributional structure. *Word* 10(2-3):146–162.

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 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 541–550.

David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*. pages 1953–1961.

David R Karger, Sewoong Oh, and Devavrat Shah. 2013. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review* 41(1):81–92.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *SIGKDD Explor. Newsl.* 17(2):1–16. https://doi.org/10.1145/2897350.2897352.

Thomas Lin, Oren Etzioni, et al. 2012. No noun phrase left behind: detecting and typing unlinkable entities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 893–903.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *AAAI*. Citeseer.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in 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 Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pages 1003–1011.
Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. 2011. Scalable knowledge harvesting with high precision and high recall. In Proceedings of the fourth ACM international conference on Web search and data mining. ACM, pages 227–236.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. Patty: a taxonomy of relational patterns with semantic types. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pages 1135–1145.

Thien Huu Nguyen and Ralph Grishman. 2015. Combining neural networks and log-linear models to improve relation extraction. arXiv preprint arXiv:1511.05926.

Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pages 877–885.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In Advances in Neural Information Processing Systems. pages 3567–3575.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pages 995–1004.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2016. Cotype: Joint extraction of typed entities and relations with knowledge bases. arXiv preprint arXiv:1610.08763.

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.

Alexander Schutz and Paul Buitelaar. 2005. Relext: A tool for relation extraction from text in ontology extension. In International semantic web conference. Springer, volume 2005, pages 593–606.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929–1958.

Sho Takase and Naoki Okazaki Kentaro Inui. 2016. Composing distributed representations of relational patterns. In Proceedings of ACL.

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2016a. Learning latent vector spaces for product search. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, pages 165–174.

Christophe Van Gysel, Maarten de Rijke, and Marcel Worrall. 2016b. Unsupervised, efficient and semantic expertise retrieval. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pages 1069–1079.

Paroma Varma, Bryan He, Dan Iter, Peng Xu, Rose Yu, Christopher De Sa, and Christopher Ré. 2016. Socratic learning: Correcting misspecified generative models using discriminative models. arXiv preprint arXiv:1610.08123.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In COLING. pages 2335–2344.

Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. Proceedings of the VLDB Endowment 5(6):550–561.

Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. 2015. Modeling truth existence in truth discovery. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pages 1543–1552.