Improving Distantly Supervised Relation Extraction by Natural Language Inference

Kang Zhou, Qiao Qiao, Yuepei Li, Qi Li

Department of Computer Science, Iowa State University, Ames, Iowa, USA
{kangzhou, qqiao1, liyp0095, qli}@iastate.edu

Abstract

To reduce human annotations for relation extraction (RE) tasks, distantly supervised approaches have been proposed, while struggling with low performance. In this work, we propose a novel DSRE-NLI framework, which considers both distant supervision from existing knowledge bases and indirect supervision from pretrained language models for other tasks. DSRE-NLI energizes an off-the-shelf natural language inference (NLI) engine with a semi-automatic relation verbalization (SARV) mechanism to provide indirect supervision and further consolidates the distant annotations to benefit multi-classification RE models. The NLI-based indirect supervision acquires only one relation verbalization template from humans as a semantically general template for each relationship, and then the template set is enriched by high-quality textual patterns automatically mined from the distantly annotated corpus. With two simple and effective data consolidation strategies, the quality of training data is substantially improved. Extensive experiments demonstrate that the proposed framework significantly improves the SOTA performance (up to 7.73% of F1) on distantly supervised RE benchmark datasets. Our code is available at https://github.com/kangISU/DSRE-NLI.

1 Introduction

Relation extraction (RE) has been studied intensively in the past years (Zhou and Chen 2021; Zhang et al. 2017; Zelenko, Aone, and Richardella 2003). It aims to extract the relations among entities from text and plays an important role in various natural language processing (NLP) tasks such as knowledge graph construction (Distiawan et al. 2019), question answering (Yu et al. 2017), and text summarization (Hachey 2009). In this paper, we define the RE task as identifying the pre-defined relation for a pair of entity mentions in a given sentence.

Due to the cost of human annotations for RE tasks, researchers have been trying to develop alternative approaches without requiring human annotations. Specially, two practically appealing learning strategies have shown promising results: distantly supervised learning using large noisy training data and zero-shot learning using indirect supervision.

Distant supervision acquires massive annotations using existing in-domain knowledge bases (Ma et al. 2021; Zheng et al. 2019; Jia et al. 2019). A commonly adopted distant annotation process for RE tasks is that if two entities participate in a relationship in the knowledge base, then all sentences in the training corpus with these two entities are labeled as positive examples of that relation (Mintz et al. 2009; Riedel, Yao, and McCallum 2010).

	Sentences in the training corpus	DS	IS
1	Akio Morita, who founded Sony in 1946, lived in NY.	Y (TP)	Y (TP)
2	Mr. Morita is the brother of Akio Morita, the co-founder of Sony.	Y (TP)	N (FN)
3	He profiles Akio Morita of Sony and Sumner Redstone of Viacom.	Y (FP)	N (TN)
4	Charles Murray is a fellow at the American Enterprise Institute.	N (TN)	N (TN)
5	James Lenox, founder of the Lenox Library, bought it in 1845.	N (FN)	Y (TP)

subject object true positive or negative false positive or negative

Figure 1: Annotation examples of distant supervision (DS) and indirect supervision (IS) for relation `founders`.

Distantly supervised methods usually face high label noise in training data due to the annotation process, since not all sentences with the entity pair express the relationship. Figure 1 shows some examples of distant annotations for the `founders` relationship. Since `Akio Morita` and `Sony` have the `founders` relationship in the knowledge base, all sentences with the two entities are labeled as positive, introducing the false positive problem (Sentence 3). On the other hand, due to the limited coverage of the knowledge base, Sentence 5 is labeled as negative, introducing the false negative problem. Therefore, existing distantly supervised methods are proposed to tackle the noise in training data (Ma et al. 2021; Jia et al. 2019; Zheng et al. 2019; Lin et al. 2016; Zeng et al. 2015).

Recently, indirectly supervised methods have taken advantage of pretrained models for other NLP tasks to solve RE tasks in the zero-shot setting. For example, RE tasks have been reformulated as question answering problems (Levy et al. 2017), as natural language inference (NLI) tasks (Obamuyide and Vlachos 2018; Sainz et al. 2021), and as text summarization tasks (Lu et al. 2022). Performance in the zero-shot setting, however, still has a significant gap, and highly relies on the quality and diversity of relation paraphrase templates. For example, we apply the NLI-based method (Sainz et al. 2021) with a relation verbalization template `{subj} was founded by {obj}`

for sentences in Figure 1. For Sentence 2, the model cannot align `co-founder` meaning with it and thus mislabels.

In this work, we introduce indirect supervision into distantly supervised RE (DSRE) tasks for the first time to improve their performance. Specifically, the proposed DSRE-NLI energizes an off-the-shelf NLI engine with a novel semi-automatic relation verbalization (SARV) mechanism to diagnose label noise in distant annotations. To involve as little as possible human effort in the relation verbalization process, we acquire only one semantically general template from humans for each relationship. To improve the semantic diversity of relation templates, we conduct an NLI-involved textual pattern mining and grouping process to enrich the template set of each relationship by choosing high-quality textual patterns from the distantly annotated corpus. These relation verbalization templates are used as-is for an NLI-based zero-shot RE on the training corpus to provide indirect supervision, which further consolidates the distant annotations with two simple and effective strategies. Finally, the consolidated training data are used to train traditional multi-class RE models for prediction.

In empirical studies, we use two real DSRE benchmark datasets and a simulated dataset to evaluate the proposed method. The results show that the proposed DSRE-NLI consistently outperforms the state-of-the-art models by large margins. The ablation study demonstrates that the mined relation patterns by SARV can significantly benefit indirect supervision. The simulation study shows that DSRE-NLI can create high-quality training data.

2 Related Work

2.1 Distantly Supervised Relation Extraction

Mintz et al. (2009) propose the DSRE task for the first time. It assumes that if two entities participate in a relation, then all sentences mentioning the two entities express that relation. Riedel, Yao, and McCallum (2010) argue that the assumption is too strong in real practice, so they modify the assumption as if two entities participate in a relation, at least one sentence that mentions the two entities expresses that relation. This assumption is further modified to allow multiple labels for an entity pair (Hoffmann et al. 2011; Surdeanu et al. 2012).

The research focus of distantly supervised methods, belonging to weakly supervised learning in a broad sense, is to tackle the noise in training data, especially the false positive problem. One strategy is to apply the multi-instance learning framework (Hoffmann et al. 2011; Surdeanu et al. 2012; Zeng et al. 2015; Lin et al. 2016; Jiang et al. 2018). These methods form positive bags of sentences for a relation following the expressed-at-least-once assumption. Then the learner is trained on sentence bags instead of individual sentences. However, Feng et al. (2018) first report that bag-level methods struggle in sentence-level prediction, which is also verified by Jia et al. (2019) and Ma et al. (2019).

Another line of approaches adopts various sentence-level denoising strategies. For example, pattern mining methods have been shown to be effective in reducing false positives in training data (Li et al. 2018a; Qu et al. 2018; Zheng et al. 2019; Jia et al. 2019). However, pattern-based methods tend to have low recall since pattern matching is a restricted process. Some methods apply reinforcement learning to automatically recognize false positive samples (Feng et al. 2018; Qin, Xu, and Wang 2018; He et al. 2020b). Ma et al. (2021) employ negative training to denoise and use a relabeling mechanism to iteratively train RE models.

2.2 Zero-Shot Relation Extraction

Zero-shot RE methods apply indirect supervision and convert RE tasks to other NLP tasks (Levy et al. 2017; Obamuyide and Vlachos 2018; Sainz et al. 2021; Lu et al. 2022). For the zero-shot setting, human annotators do not label any samples but are usually asked to generate relation paraphrase templates. For example, Levy et al. (2017) use crowdsourcing to generate question templates. Sainz et al. (2021) ask human annotators to generate verbalization templates for at most 15 minutes per relation. These methods show that the few-shot setting performs significantly better than the zero-shot setting, but requires some human efforts for annotating training samples.

In this paper, we adopt NLI-based indirect supervision for two reasons: 1) relation verbalization templates can be directly obtained from textual patterns mined from distantly labeled corpus; 2) the inference step from NLI to RE is straightforward.

3 Preliminary

3.1 RE Task Definition

We formalize the RE task as follows. Let $\boldsymbol{x} = [x_1, ..., x_n]$ denote a sentence, where x_i is the i-th token. An entity pair $(\boldsymbol{e}_{subj}, \boldsymbol{e}_{obj})$, referring to the subject and object entities, respectively, is identified in the sentence, where $\boldsymbol{e}_{subj} = [x_{ss}, ..., x_{se}]$ and $\boldsymbol{e}_{obj} = [x_{os}, ..., x_{oe}]$ are two non-overlapping consecutive spans. Given an instance, which includes the sentence $\boldsymbol{x}$, and the specific positions and entity types of $\boldsymbol{e}_{subj}$ and $\boldsymbol{e}_{obj}$, the goal is to predict the relation $r \in \mathcal{R} \cup \{\mathrm{NA}\}$ that holds between $\boldsymbol{e}_{subj}$ and $\boldsymbol{e}_{obj}$, where $\mathcal{R}$ is a pre-defined relation set, and NA indicates that no relation from $\mathcal{R}$ is expressed between them.

3.2 Distant Annotation

To construct distantly annotated training data, named entity mentions are first recognized from the corpus by named entity recognition (NER) methods (Meng et al. 2021; Zhou, Li, and Li 2022). Then, by exact string matching, the named entity mentions are linked to the entities in the knowledge base (e.g., Freebase) that covers the relations of interest (i.e., $\mathcal{R}$). If a sentence contains two entity mentions that have a relation of interest in the knowledge base, then a corresponding instance will be generated and labeled as the relation type r. Otherwise, an instance labeled as NA will be generated.

3.3 NLI for RE

Since our framework will utilize the reformulation of RE into NLI to obtain indirect supervision, here we briefly introduce the reformulation schema proposed in Sainz et al. (2021). Given a textual premise and a hypothesis, NLI, also

known as textual entailment, is to determine whether the premise entails, or is neutral to, or contradicts the hypothesis. The reformulation requires three sub-processes: relation verbalization, NLI input generation, and relation inference.

Relation verbalization is to verbalize a relation by a simple paraphrase of only a few tokens. Such paraphrases are called verbalization templates. For example, {subj} was founded by {obj} is a verbalization template for the relation /business/company/founders from Freebase, where {subj} and {obj} are placeholders. Note that a relation can have multiple verbalization templates. For example, {obj} is a founder of {subj} can be another template for founders relation. For a relation $r \in \mathcal{R}$, Sainz et al. (2021) give 15 minutes to human annotators to generate several templates and construct a verbalization template set $\mathcal{T}_r$, and $|\mathcal{T}_r| \geq 1$.

NLI input generation is to generate premise-hypothesis pairs for each sentence. These pairs will be taken as input by an NLI model. Given a template $\boldsymbol{t} \in \mathcal{T}_r$ and a sentence $\boldsymbol{x}$ mentioning two entities $\boldsymbol{e}_{subj}$ and $\boldsymbol{e}_{obj}$, NLI input generation yields a premise-hypothesis pair $(\boldsymbol{x}, \boldsymbol{h})$, where $\boldsymbol{h} = hyp(\boldsymbol{t}, \boldsymbol{e}_{subj}, \boldsymbol{e}_{obj})$ with $hyp(\cdot)$ substituting placeholders in the template with actual entities.

Relation inference is to infer the relationship expressed by two entity mentions in a sentence from the outputs of an NLI model. Taking a premise-hypothesis pair as input, an NLI model with a softmax output layer yields a probability distribution $\boldsymbol{P}_{NLI}(\boldsymbol{x}, \boldsymbol{h})$ over entailment (E), neutrality (N), and contradiction (C). Then, the probability of $(\boldsymbol{e}_{subj}, \boldsymbol{e}_{obj})$ expressing relation r in sentence $\boldsymbol{x}$ is computed by:

$$P_r(\boldsymbol{x}, \boldsymbol{e}_{subj}, \boldsymbol{e}_{obj}) = \delta_r \max_{\boldsymbol{t} \in \mathcal{T}_r} P^E_{NLI}(\boldsymbol{x}, \boldsymbol{h}),$$

where P^E_{NLI} is the probability of entailment and δ_r is an indicator function to tackle entity type constraints. Occasionally, a template can verbalize more than one relation, which will bring ambiguity in later inference. For example, {subj} was born in {obj} can verbalize both country_of_birth and city_of_birth. δ_r considering NER type information can tackle this issue. $\delta_r(\boldsymbol{e}_{subj}, \boldsymbol{e}_{obj}) = 1$ if $\text{NER}(\boldsymbol{e}_{subj}, \boldsymbol{e}_{obj}) \in \mathcal{E}_r$, otherwise 0, where $\mathcal{E}_r$ is a set of NER type constraints for relation r. For example, the former requires $\text{NER}(\cdot) \in$ {PERSON:COUNTRY}, while the latter requires $\text{NER}(\cdot) \in$ {PERSON:CITY}. Then, the final predicted relation $\hat{r}$ of $(\boldsymbol{e}_{subj}, \boldsymbol{e}_{obj})$ in $\boldsymbol{x}$ is given by:

$$\hat{r} = \arg\max_{r \in \mathcal{R}} P_r(\boldsymbol{x}, \boldsymbol{e}_{subj}, \boldsymbol{e}_{obj}).$$

Note that if $\max_{r \in \mathcal{R}} P_r(\boldsymbol{x}, \boldsymbol{e}_{subj}, \boldsymbol{e}_{obj}) < \tau$, then $\hat{r} = \text{NA}$, where τ is a threshold that is a hyperparameter.

Sainz et al. (2021) propose two settings: zero- and few-shot. Zero-shot directly uses a pretrained NLI model, while few-shot uses training data with ground truth labels to fine-tune the pretrained NLI model. Zero-shot shows a significant gap in performance compared with few-shot.

4 Methodology

This section introduces the proposed framework DSRE-NLI. Figure 2 illustrates the architecture. We use the distant annotation process in Section 3.2 to provide distant supervision (DS) and the NLI-based zero-shot RE in Section 3.3 to provide indirect supervision (IS). The proposed SARV generates high-quality textual patterns from the distantly annotated corpus for each relationship. Then the patterns together with one human-generated general template are used for the NLI model to provide indirect supervision. Finally, the distant annotations are consolidated with the results of the NLI model and used for the multi-class RE model training.

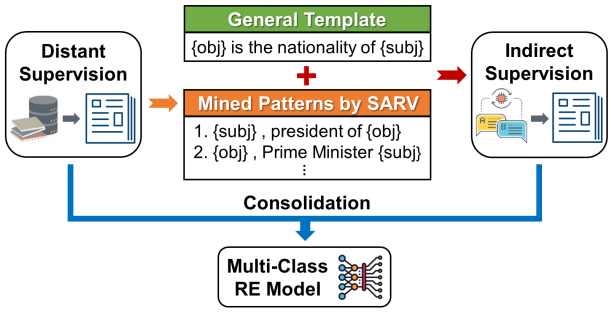

Figure 2: DSRE-NLI framework. Template and pattern examples are from relation nationality.

4.1 Semi-Automatic Relation Verbalization

The performance of the NLI-based zero-shot RE is highly dependent on the quality and diversity of verbalization templates, where template quality reflects if a template accurately conveys the semantic meaning of the relationship, and template diversity reflects the semantic coverage of templates. As introduced in Section 3.3, the reformulation of RE into NLI requires human-written templates.

However, human-written templates are usually prone to semantic generalization and duplication, and may not fit the writing style of the specific corpus, leading to poor performance and computational inefficiency in the zero-shot setting. For example, in Figure 2 the human-written template {obj} is the nationality of {subj} may be too general to capture the specific expressions of relation nationality in the corpus.

Furthermore, a pretrained NLI model determines text entailment differently from humans, making it harder for humans to propose diverse useful templates. For example, an annotator may provide both {subj} is the parent of {obj} and {obj} is the son of {subj} for relation children, believing the two templates increase the semantic diversity than a single one. However, a pretrained NLI model may determine that the latter actually strongly entails the former, indicating that the latter has no contribution to the diversity because of an observed transition property that if sentence $\boldsymbol{x}$ entails template $\boldsymbol{t}_1$, and $\boldsymbol{t}_1$ entails template $\boldsymbol{t}_2$, then $\boldsymbol{x}$ entails $\boldsymbol{t}_2$. On the other hand, an annotator may think {subj} was founded by {obj} is similar to {obj} is a co-founder of {subj} for relation founders and thus omit one. However, a pretrained NLI model may not determine that they strongly entail each other, which means that the two templates may

improve the semantic coverage. Therefore, without prior knowledge of the corpus and the NLI model behavior, it is hard for an annotator to propose proper and diverse templates for specific relationships.

To overcome the template fitness and diversity issues, we propose a novel semi-automatic relation verbalization (SARV) method by considering the content of the distantly annotated corpus, so that an annotator can efficiently select templates of higher quality and higher diversity for each relationship.

We propose to generate template candidates by textual pattern mining and grouping. Pattern mining from the distantly annotated corpus can discover various expression styles in the specific corpus. Pattern grouping can discard semantically duplicated templates, and thus improve computational efficiency. Moreover, pattern grouping can accumulate the pattern frequency in individual groups so as to highlight truly useful semantic patterns.

Pattern Mining and Grouping For each relationship, we collect all the distantly labeled instances of it and conduct a simple pattern mining that directly takes the token sequence between subject and object entities, which is easy and efficient. To ensure the quality of candidate patterns, we use three criteria to filter patterns: 1) pattern frequency should be higher than a threshold; 2) pattern length should be shorter than a threshold; 3) a pattern must contain at least one non-stop-word. Advanced pattern mining techniques can also be applied, such as PATTY (Nakashole, Weikum, and Suchanek 2012), mining patterns from the shortest dependency path between a pair of entities, and MetaPAD (Jiang et al. 2017), mining patterns from entire sentences.

Previous pattern mining methods use shallow features to group patterns into semantic clusters, such as pattern lexicon and extraction overlaps (Li et al. 2018b; Nakashole, Weikum, and Suchanek 2012; Jiang et al. 2017) and pattern embeddings (Li et al. 2018a). These methods do not group patterns with semantic understanding of the patterns. Therefore, we propose an NLI-based pattern grouping method.

We first define semantic duplication from the perspective of the NLI task as follows.

Definition. Given patterns $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ of relation r, and $|\boldsymbol{p}_1| \geq |\boldsymbol{p}_2|$, if $\boldsymbol{p}_1$ strongly entails $\boldsymbol{p}_2$ (i.e., $P^E_{NLI}(\boldsymbol{p}_1, \boldsymbol{p}_2) \geq \tau$) by an NLI model, then we say $\boldsymbol{p}_1$ is semantically duplicated to $\boldsymbol{p}_2$.

Pattern grouping aims to reduce semantic duplication and only maintain the semantically distinct representative patterns for each relation. To do so, we first rank the initial patterns from the mining results by their length. Then starting from the shortest pattern, we recursively use a pretrained NLI model to determine which patterns from the rest longer ones are semantically duplicated to the pattern. If $|\boldsymbol{p}_j| \geq |\boldsymbol{p}_i|$ and $\boldsymbol{p}_j$ is duplicated to $\boldsymbol{p}_i$, then $\boldsymbol{p}_j$ is grouped with $\boldsymbol{p}_i$, and the frequency of $\boldsymbol{p}_j$ will be added to that of $\boldsymbol{p}_i$. After the grouping process, the shortest (semantically dense) patterns of individual groups will be the final representative patterns for the specific relation, and they will be ranked by their group frequencies (i.e., accumulated frequencies of all patterns in the group) for the further manual screening.

Algorithm 1: SARV

Input: Distantly labeled corpus $\mathcal{C}$ and relation label r
Output: Template set $\mathcal{T}_r$

1 mine an initial pattern set $\mathcal{P}_r^{initial}$ from $\mathcal{C}$;
2 sort $\mathcal{P}_r^{initial}$ by pattern length $|\boldsymbol{p}|$ in increasing order;
3 $\mathcal{P}' \leftarrow \emptyset$;
4 **for** $\boldsymbol{p}_i$ **in** $\mathcal{P}_r^{initial}$ **do**
5 $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{\boldsymbol{p}_i\}$;
6 **for** $\boldsymbol{p}_j$ **in** $\mathcal{P}_r^{initial} - \mathcal{P}'$ **do**
7 **if** $f_{\boldsymbol{p}_i} > 0$ **and** $f_{\boldsymbol{p}_j} > 0$ **and** $\boldsymbol{p}_j \Rightarrow \boldsymbol{p}_i$ **then**
8 $f_{\boldsymbol{p}_i} \leftarrow f_{\boldsymbol{p}_i} + f_{\boldsymbol{p}_j}$;
9 $f_{\boldsymbol{p}_j} \leftarrow 0$;
10 $\mathcal{P}_r^{grouped} \leftarrow \mathcal{P}_r^{initial} - \{\boldsymbol{p} \mid f_{\boldsymbol{p}} = 0\}$;
11 request a general template $\boldsymbol{t}_r$, and add into $\mathcal{T}_r$;
12 **for** $\boldsymbol{p}_i$ **in** $\mathcal{P}_r^{grouped}$ **do**
13 **if** $\boldsymbol{p}_i \Rightarrow \boldsymbol{t}_r$ **then**
14 $f_{\boldsymbol{p}_i} \leftarrow 0$;
15 $\mathcal{P}_r^{grouped} \leftarrow \mathcal{P}_r^{grouped} - \{\boldsymbol{p} \mid f_{\boldsymbol{p}} = 0\}$;
16 sort $\mathcal{P}_r^{grouped}$ by pattern frequency, then select high-quality patterns from $\mathcal{P}_r^{grouped}$, and add into $\mathcal{T}_r$;
17 **return** $\mathcal{T}_r$;

Template Generation and Selection Since a pre-defined relation label usually conveys the semantic meaning of the relation, an annotator can easily propose one semantically general verbalization template as a general template for each relation even without much domain knowledge. This can guarantee there is at least one template of high quality. It is important especially for long-tail relationships where the patterns are sparse. With the general template, we can continue shrinking the pattern candidates of each relation by removing patterns that are semantically duplicated to it. From the results, human annotators further select high-quality patterns as additional templates.

Algorithm 1 summarizes the SARV process, where '$\Rightarrow$' denotes 'semantically duplicated to', and $f_{\boldsymbol{p}}$ denotes the frequency of $\boldsymbol{p}$. Line 1 is conducting pattern mining. Lines 2-10 are conducting pattern grouping. Lines 11-16 are performing template generation and selection.

4.2 Training Data Consolidation

Recall from Section 3.2, the distant supervision can annotate a set of instances for each relation $r \in \mathcal{R} \cup \{\mathrm{NA}\}$. Our first consolidation strategy is to use a pretrained NLI model to filter all distantly annotated sets. For the set of instances annotated as r by distant supervision, if an instance is not predicted as r by the NLI model, it will be removed. That is, the intersection of the results from DS and IS. We call this strategy as IPIN (**I**ntersection of **P**ositives and **I**ntersection of **N**egatives).

Another strategy is to use only IS to construct the set of instances for relation $r \in \mathcal{R}$ and use the intersection of DS and IS to construct the set of instances for NA. The reason is that the intersection of positive instances may reduce the size of positive instances significantly and impact the learn-

Dataset		NYT10.1	NYT10.2	TACREV
# relation		10	11	41
Train	# total inst	376,355	373,643	68,124
	# pos inst	95,519	92,807	13,012
Dev	# total inst	2,379	4,569	22,631
	# pos inst	338	973	5,300
Test	# total inst	2,164	4,482	15,509
	# pos inst	330	1,045	3,123

Table 1: Statistics of used datasets.

ing effect, especially for long-tail relationships, and IS (i.e., the NLI model) can generally provide predictions with high precision. For NA, since both DS and IS can produce large sets, the intersection of sets can filter out false negatives and still maintain sufficient instances. We call this strategy NPIN (**N**LI **P**ositives and **I**ntersection of **N**egatives).

4.3 Multi-Class RE Model

Using the consolidated training data, we can train a multi-class RE model. We adopt the architecture proposed by Zhou and Chen (2021) using entity mask as the entity representation strategy. Although in their experiments, the typed entity marker strategy performed better, for the DSRE task, we find that entity mask is more tolerant to the label noise. This is because the pretrained language model (BERT in our case) has strong prior knowledge about entities and the noise in training can strengthen the reliance on prior knowledge. Without the surface names of entities, the model learns from the context of the sentences instead and thus is more robust to label noise.

5 Experiments

5.1 Experiments on Real Distantly Annotated Datasets

Datasets and Evaluation Metrics We conduct experiments to evaluate the proposed framework on the widely-used public dataset: New York Times (NYT), which is a large-scale distantly labeled dataset constructed from NYT corpus using Freebase as the distant supervision (Riedel, Yao, and McCallum 2010). Recently, Jia et al. (2019) manually labeled a subset of the data as the testing set for a more accurate evaluation and constructed two versions of the dataset: NYT10.1 and NYT10.2[1]. The latter version is released after their paper publication. The statistics of the two datasets are summarized in Table 1, and more details about the instance generation can be found in Technical Appendix Section 1. We report the evaluation results in terms of precision, recall, and F1 score. For all metrics, the higher the better.

Baseline Methods We compare the proposed DSRE-NLI with three categories of methods including normal RE models, DSRE methods, and zero-shot RE methods. For normal RE models, we consider two representative models BiLSTM (Zhang et al. 2015) and $\text{BERT}_{\text{EntityMask}}$ (Zhou and Chen 2021). They both use the position information of entity mentions. For DSRE methods, we consider two recent methods ARNOR (Jia et al. 2019) and SENT (Ma et al. 2021). ARNOR achieves the SOTA performance on NYT10.2, while SENT achieves the SOTA performance on NYT10.1. For zero-shot RE methods, we consider the NLI-based RE method (Sainz et al. 2021), only using the human written general template for each relation.

DSRE-NLI Setups For the setup of **SARV**, one of the authors constructs one general template for each relation without reading any example sentences from the corpus. We initially choose the top 10% most frequent mined patterns and only keep those that consist of less than 10 tokens and at least one non-stop-word token. For computational efficiency, we retain at most 50 patterns for each relation for the following pattern grouping. After pattern grouping, patterns with a frequency of at least 10 are eligible for manual screening. One author is presented with a pattern and one example sentence at a time and is asked if this pattern can induce the target relation.

For the setup of **NLI model**, we use the pretrained DeBERTa v2 model (He et al. 2020a) to implement the NLI-based RE method and set the entailment probability threshold τ to 0.95, which is empirically suggested by Sainz et al. (2021) based on their study on a different dataset. This setting is used for both pattern grouping and IS. Note that in the following sections 'genr' denotes the NLI model using the general template for each relation, and 'genr+patt' denotes using extra patterns given by SARV. For the setup of **multi-class RE model**, we use $\text{BERT}_{\text{EntityMask}}$ model from Zhou and Chen (2021).

We run all methods using one Tesla V100S GPU (32G). We train DSRE-NLI for 2 epochs on both NYT10.1 and NYT10.2 training variants.

Main Results We summarize the comparison results in Table 2 on dev and test sets of NYT10.1 and NYT10.2, respectively. Note that DSRE-NLI does not use the dev set to tune any hyperparameter. The first category of baseline methods treats the distant annotations as ground truth labels, and the results show that they suffer from low precision. It validates that for the distantly annotated data, high false positive rate is the major issue. With the denoising process, the DSRE methods clearly improve the precision and achieve significantly higher F1 scores. The zero-shot method (i.e., $\text{NLI}_{\text{DeBERTa}}$-genr) obtains high precision, either the best or the runner-up among all methods, but suffers from low recall. It implies that the general template for each relation has considerably low semantic coverage.

The results clearly show that the proposed DSRE-NLI framework significantly outperforms previous state-of-the-art methods. DSRE-NLI with the two consolidation strategies either achieves the best or the runner-up overall performance (F1). $\text{DSRE-NLI}_{\text{NPIN}}$ outperforms on NYT10.1 test set by a margin of 6.56% of F1 compared with the best baseline ($\text{SENT}_{\text{BiLSTM+BERT}}$), and $\text{DSRE-NLI}_{\text{IPIN}}$ outperforms on NYT10.2 test set by a margin of 7.73% of F1 comparing with the best baseline (ARNOR). Comparing the two

[1]https://github.com/PaddlePaddle/Research/tree/master/NLP/ACL2019-ARNOR

Method	NYT10.1-Dev			NYT10.1-Test			NYT10.2-Dev			NYT10.2-Test		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM†	36.71	66.46	47.29	35.52	67.41	46.53	41.46	70.17	52.12	44.12	71.12	54.45
$\text{BERT}_{\text{EntityMask}}$	44.71	77.81	58.38	45.05	77.27	56.92	45.04	78.93	57.36	48.75	81.91	61.12
ARNOR†	62.45	58.51	60.36	65.23	56.79	60.90	**78.14**	59.82	67.77	**79.70**	62.30	69.93
$\text{SENT}^{\dagger}_{\text{BiLSTM+BERT}}$	69.94	63.11	66.35	**76.34**	63.66	69.42	-	-	-	-	-	-
$\text{SENT}_{\text{BiLSTM}}$	58.97	47.63	52.70	58.53	45.76	51.36	55.69	56.83	56.26	56.89	58.47	57.67
$\text{NLI}_{\text{DeBERTa}}$-genr	**74.21**	55.33	63.39	70.20	52.12	59.83	73.43	51.70	60.68	75.24	51.77	61.34
$\text{DSRE-NLI}_{\text{IPIN}}$	71.93	79.59	**75.56**	68.03	80.61	73.79	69.26	79.65	**74.09**	73.90	81.82	**77.66**
$\text{DSRE-NLI}_{\text{NPIN}}$	67.06	**84.91**	74.94	68.80	**84.85**	**75.98**	66.23	**82.63**	73.53	68.59	**84.40**	75.68

Table 2: Results are in %, where the bests are in bold, and the runner-ups are underlined. † cites results from referenced papers.

Method	NYT10.1-(D+T)	NYT10.2-(D+T)
	F1	F1
$\text{NLI}_{\text{DeBERTa}}$		
-genr	61.63	61.02
-genr+patt	73.20 (+11.57)	70.63 (+9.61)
DSRE-NLI		
-DS	57.65	59.28
-IPIN(DS, IS(genr))	70.50 (+12.85)	74.58 (+15.30)
-IPIN(DS, IS(genr+patt))	**74.67** (+17.02)	**75.92** (+16.64)

Table 3: Results of two categories of ablation studies.

strategies, $\text{DSRE-NLI}_{\text{NPIN}}$ consistently achieves higher recall because NPIN strategy obtains bigger and possibly noisier training data for positive classes.

Ablation Study We conduct ablation studies to investigate the contributions of DS, IS, and pattern mining and grouping to the overall DSRE performance. Table 3 summarizes the results on combined dev and test data because DSRE-NLI does not use the dev set to tune any hyperparameter.

The first category of studies examines the importance of pattern mining and grouping for IS only. With the additional chosen patterns, the F1 score increases 11.57% and 9.61% on the two datasets, respectively, compared with using general templates only. The results indicate that SARV increases the semantic diversity of the relation verbalization. The only additional manual effort is to select patterns from candidates, which requires much less effort than designing diverse templates from scratch. On average, the pattern mining method can find around 44 patterns per relation initially (Table 4), but after pattern grouping, most patterns are merged, resulting in just around 5 patterns per relation for manual screening. The human annotator chose around 2.5 patterns from them. More details can be found in Tables 11 and 12 in Technical Appendix.

The second category of studies examines the impact of IS consolidation for the DSRE task. It is clear that using DS directly for training, the multi-class RE model performs poorly. Using IS with general templates for data consolidation (e.g., IPIN), we can see that the overall performance boosts significantly, gaining 12.85% and 15.30% of F1 scores on the two datasets, respectively. With selected patterns mined from the corpus, the performance further improves for 4.17% and 1.34%, respectively.

Dataset	# init patt	# patt after group	# selected patt
NYT10.1	45.60	5.20	2.50
NYT10.2	43.00	5.00	2.45

Table 4: Average number of patterns over all relations.

Pattern	After		Before	
	Rk	Fq	Rk	Fq
{subj} , president of {obj} (✓)	1	192	16	7
{obj} , Prime Minister {subj} (✓)	2	89	19	7
{subj} , who led {obj} (✗)	3	87	27	4
{obj} 's foreign minister , {subj} (✓)	4	48	3	45
{obj} 's former prime minister , {subj} (✓)	5	23	12	9
{obj} President {subj} (✓)	6	21	35	4
{obj} named {subj} (✗)	7	20	49	3
{obj} 's acting prime minister , {subj} (✓)	8	19	5	19

Table 5: Patterns of `nationality` generated by SARV.

Case Study We use `nationality` from NYT10.1 dataset as an example relation to illustrate the process of DSRE-NLI. Table 5 illustrates the relation patterns generated by SARV. The patterns are sorted by the frequency after pattern grouping, and the check marks after the patterns indicate if the pattern is selected. We can see that the ranks and the frequencies before and after pattern grouping are very different, indicating the grouping can merge semantically similar patterns. It is interesting to see that most patterns imply that the person is the leader of the country, which indeed implies the person's nationality. Humans may find it hard to construct such patterns without reading the corpus.

The relation extraction performance on `nationality` is shown in Table 6. We can see that both IPIN and NPIN strategies can significantly reduce the training instances of this relation, but the performance is improved, indicating that the removed instances are likely to be noise.

Fine-tuning NLI Models with Distantly Annotated Training Data Since fine-tuning NLI models using training data with ground-truth labels and human-written relation templates demonstrates significant improvement for RE tasks in the previous work (Sainz et al. 2021), here we examine if fine-tuning the NLI model with the distant annotations can also bring the improvement for DSRE tasks. We follow the fine-tuning process in Sainz et al. (2021) and fine-tune the NLI model using the distant annotations and the

NYT10.1-Train		NYT10.1-(D+T)		
Training data	# inst	P	R	F1
DS	8,355	58.46	**92.68**	71.70
IPIN(DS, IS(genr+patt))	2,850	**73.33**	80.49	76.74
NPIN(DS, IS(genr+patt))	3,591	72.92	85.37	**78.65**

Table 6: Results of `nationality` in various settings.

Method	NYT10.2-(D+T)		
	P	R	F1
$\text{NLI}_{\text{DeBERTa}}$-genr (no FT)	74.36	51.73	61.02
-FT with DS	45.32	78.05	57.35
-FT with IPIN(DS, IS(genr))	53.64	**83.55**	65.34
DSRE-NLI			
-IPIN(DS, IS(genr))	**76.08**	73.14	**74.58**

Table 7: Performance comparison between fine-tuning (FT) the NLI model and the proposed method.

consolidated annotations of IPIN strategy, respectively, with the general templates. The results on NYT10.2 are summarized in Table 7. Compared with the zero-shot setting (i.e., no FT), the fine-tuning with distant annotations (i.e., with DS) causes a significant drop in precision, leading to a lower F1. The fine-tuning with the consolidated annotations of IPIN strategy brings improvement but still performs significantly worse than the proposed $\text{DSRE-NLI}_{\text{IPIN}}$ with the multi-class RE model due to its lower precision. We conclude that using fine-tuned NLI model in replacement of the multi-class RE model is not recommended in DSRE tasks due to the noise in training data used for fine-tuning.

5.2 Experiments on Simulated Distantly Annotated Datasets

Since there is no human annotation for the training data of NYTs, we cannot directly evaluate the effectiveness of DSRE-NLI in improving training data quality. Therefore, we simulate the distant annotations on TACREV dataset (TACRED with revised dev and test sets) (Zhang et al. 2017; Alt, Gabryszak, and Hennig 2020), a human-annotated relation extraction dataset, to quantitatively evaluate DSRE-NLI on improving training data. The statistics of the original TACREV dataset can be found in Table 1.

Simulation Process To simulate the effect of distant annotation, we introduce both false positive (FP) errors and false negative (FN) errors in training data. We manipulate the original training instances from the perspective of entity pairs. To add FN errors, we first define long-tail entity pairs: if an entity pair is mentioned by n sentences and n is less than a threshold, then it is of long-tail. We relabel the instances mentioning long-tail entity pairs as NA to simulate the effect of limited coverage of knowledge bases. We empirically set the threshold so that FN rate is about 5% based on our estimation from NYT datasets. To add FP errors, we follow the distant annotation process: if an entity pair participates in a relation, then all sentences in the training corpus mentioning the entity pair are labeled as positive instances of that relation. The statistics for the simulated training dataset, TACREV-S, can be found in Table 8.

Training data	# TP	# FP	# TN	# FN
TACREV	13,012	-	55,112	-
TACREV-S	10,256	9,838 (48.96%)	49,090	2,756 (5.32%)
IPIN(S-DS, IS)	4,895	898 (15.50%)	47,519	1,499 (3.06%)
NPIN(S-DS, IS)	6,419	3,331 (34.16%)	47,519	1,499 (3.06%)

Table 8: Training data quality in different settings.

Method	TACREV-Test		
	P	R	F1
$\text{BERT}_{\text{EntityMask}}$-S-DS	53.75	**69.48**	60.62
$\text{SENT}_{\text{BiLSTM}}$-S-DS	64.69	37.78	47.71
$\text{NLI}_{\text{DeBERTa}}$			
-temp†	80.02	49.25	60.97
-genr	<u>83.82</u>	39.48	53.68
-genr+patt	78.75	49.95	61.13
DSRE-NLI			
-IPIN(S-DS, IS(genr+patt))	**84.59**	47.97	<u>61.22</u>
-NPIN(S-DS, IS(genr+patt))	75.95	<u>55.62</u>	**64.21**

Table 9: Results on TACREV test set.

Main Results Table 8 also shows the training data statistics obtained after IPIN(S-DS, IS(genr+patt)) and NPIN(S-DS, IS(genr+patt)) processes, where S-DS means simulated DS. It is clear that both strategies reduce FP and FN rate significantly, especially IPIN strategy.

Table 9 demonstrates the RE performance on the test set of TACREV using different TACREV-S training data. Similar to NYT datasets, the normal RE model ($\text{BERT}_{\texttt{EntityMask}}$) encounters a significant performance drop compared to using ground-truth annotations (79.06 of F1). $\text{SENT}_{\text{BiLSTM}}$ improves the precision but experiences significantly low recall, indicating that this method may be too aggressive in denoising. In the zero-shot category, temp† uses all human-written templates provided by Sainz et al. (2021), where each relation has multiple human-written templates. We also compare our designed general templates (genr) and general templates with pattern enrichment (genr+patt). We can see that patterns generated by SARV are comparable with the human-written templates given by Sainz et al. (2021). $\text{DSRE-NLI}_{\text{NPIN}}$ achieves the best results on this dataset.

6 Conclusion

In this work, we propose a novel DSRE-NLI framework considering indirect supervision given by pretrained NLI models in DSRE tasks. We also design a novel SARV method to reduce the template design effort required by NLI-based zero-shot RE methods. With two simple and effective data consolidation strategies, the quality of training data is substantially improved. Extensive experiments demonstrate that the proposed framework significantly improves the SOTA performance on DSRE benchmark datasets.

Acknowledgments

The work is supported in part by NIFA grant no. 2022-67015-36217 from the USDA National Institute of Food and Agriculture.

References

Alt, C.; Gabryszak, A.; and Hennig, L. 2020. TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1558–1569.

Distiawan, B.; Weikum, G.; Qi, J.; and Zhang, R. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 229–240.

Feng, J.; Huang, M.; Zhao, L.; Yang, Y.; and Zhu, X. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Hachey, B. 2009. Multi-document summarisation using generic relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 420–429.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020a. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

He, Z.; Chen, W.; Wang, Y.; Zhang, W.; Wang, G.; and Zhang, M. 2020b. Improving neural relation extraction with positive and unlabeled learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7927–7934.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 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*, 541–550.

Jia, W.; Dai, D.; Xiao, X.; and Wu, H. 2019. ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1399–1408.

Jiang, M.; Shang, J.; Cassidy, T.; Ren, X.; Kaplan, L. M.; Hanratty, T. P.; and Han, J. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 877–886.

Jiang, T.; Liu, J.; Lin, C.-Y.; and Sui, Z. 2018. Revisiting distant supervision for relation extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 333–342.

Li, Q.; Jiang, M.; Zhang, X.; Qu, M.; Hanratty, T. P.; Gao, J.; and Han, J. 2018a. Truepie: Discovering reliable patterns in pattern-based information extraction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1675–1684.

Li, Q.; Wang, X.; Zhang, Y.; Ling, F.; Wu, C. H.; and Han, J. 2018b. Pattern discovery for wide-window open information extraction in biomedical literature. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 420–427. IEEE.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2124–2133.

Lu, K.; Hsu, I.; Zhou, W.; Ma, M. D.; Chen, M.; et al. 2022. Summarization as Indirect Supervision for Relation Extraction. *arXiv preprint arXiv:2205.09837*.

Ma, F.; Li, Y.; Zhang, C.; Gao, J.; Du, N.; and Fan, W. 2019. Mcvae: Margin-based conditional variational autoencoder for relation classification and pattern generation. In *The World Wide Web Conference*, 3041–3048.

Ma, R.; Gui, T.; Li, L.; Zhang, Q.; Huang, X.-J.; and Zhou, Y. 2021. SENT: Sentence-level Distant Relation Extraction via Negative Training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6201–6213.

Meng, Y.; Zhang, Y.; Huang, J.; Wang, X.; Zhang, Y.; Ji, H.; and Han, J. 2021. Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10367–10378.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 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*, 1003–1011.

Nakashole, N.; Weikum, G.; and Suchanek, F. 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*, 1135–1145.

Obamuyide, A.; and Vlachos, A. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 72–78.

Qin, P.; Xu, W.; and Wang, W. Y. 2018. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2137–2147.

Qu, M.; Ren, X.; Zhang, Y.; and Han, J. 2018. Weakly-supervised relation extraction by pattern-enhanced embed-

ding learning. In *Proceedings of the 2018 World Wide Web Conference*, 1257–1266.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 148–163. Springer.

Sainz, O.; de Lacalle, O. L.; Labaka, G.; Barrena, A.; and Agirre, E. 2021. Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1199–1212.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 455–465.

Yu, M.; Yin, W.; Hasan, K. S.; dos Santos, C.; Xiang, B.; and Zhou, B. 2017. Improved Neural Relation Detection for Knowledge Base Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 571–581.

Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb): 1083–1106.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1753–1762.

Zhang, S.; Zheng, D.; Hu, X.; and Yang, M. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, 73–78.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 35–45.

Zheng, S.; Han, X.; Lin, Y.; Yu, P.; Chen, L.; Huang, L.; Liu, Z.; and Xu, W. 2019. DIAG-NRE: A Neural Pattern Diagnosis Framework for Distantly Supervised Neural Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1419–1429.

Zhou, K.; Li, Y.; and Li, Q. 2022. Distantly Supervised Named Entity Recognition via Confidence-Based Multi-Class Positive and Unlabeled Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7198–7211.

Zhou, W.; and Chen, M. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.