Gradient Imitation Reinforcement Learning for Low Resource Relation Extraction

Xuming Hu[1], **Chenwei Zhang**[2†], **Yawen Yang**[1], **Xiaohe Li**[1], **Li Lin**[1], **Lijie Wen**[1†], **Philip S. Yu**[1,3]
[1]Tsinghua University, [2]Amazon, [3]University of Illinois at Chicago
[1]{hxm19,yyw19,lixh18,lin-l16}@mails.tsinghua.edu.cn
[1]wenlj@tsinghua.edu.cn [2]cwzhang@amazon.com [3]psyu@uic.edu

Abstract

Low-resource Relation Extraction (LRE) aims to extract relation facts from limited labeled corpora when human annotation is scarce. Existing works either utilize self-training scheme to generate pseudo labels that will cause the gradual drift problem, or leverage meta-learning scheme which does not solicit feedback explicitly. To alleviate selection bias due to the lack of feedback loops in existing LRE learning paradigms, we developed a Gradient Imitation Reinforcement Learning method to encourage pseudo label data to imitate the gradient descent direction on labeled data and bootstrap its optimization capability through trial and error. We also propose a framework called GradLRE, which handles two major scenarios in low-resource relation extraction. Besides the scenario where unlabeled data is sufficient, GradLRE handles the situation where no unlabeled data is available, by exploiting a contextualized augmentation method to generate data. Experimental results on two public datasets demonstrate the effectiveness of GradLRE on low resource relation extraction when comparing with baselines. Source code is available[1].

1 Introduction

Relation Extraction (RE) aims to discover the semantic relation that holds between two entities and transforms massive corpus into structured triplets (entity$_{head}$, relation, entity$_{tail}$). For example, from "A *letter*$_{head}$ was delivered to my *office*$_{tail}$...", we can extract a relation Entity-Destination between head and tail entities. Neural RE methods leverage high-quality annotated data or human curated knowledge bases to achieve decent results (Zeng et al., 2017; Zhang et al., 2017). However, these manually labeled data would be labor-intensive to obtain. This motivates a Low Resource Relation Extraction (LRE) task where annotations are scarce.

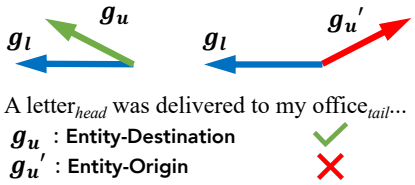

Figure 1: Gradient descent direction on labeled data (g_l) and unlabeled data with correct or incorrect pseudo label (g_u, $g_u^{'}$).

Lots of efforts are devoted to improve the model generalization ability beyond learning directly from existing, limited annotations. Distant Supervision methods leverage facts stored in external knowledge bases (KBs) to obtain annotated triplets as the supervision (Mintz et al., 2009; Zeng et al., 2015). However, these methods should make a strong assumption that two co-occurring entities convey KB relations regardless of specific contexts, which makes model generate relations based on contextless rules and limits the generalization ability. To leverage unlabeled data, Rosenberg et al. (2005) propose to assign pseudo labels on unlabeled data and leverage pseudo labels to iteratively improve the generalization capability of the model. However, during the training process, self-training models suffer from the gradual drift problem (Curran et al., 2007; Zhang et al., 2016) caused by noisy pseudo labels. Hu et al. (2021) alleviate the noise in pseudo labels by adopting a meta-learning scheme during pseudo label generation, then leveraging pseudo label selection and exploitation scheme to obtain high-confidence pseudo labels. However, when limited annotations are directly used during training, the trained models inevitably possesses selection bias towards, if not overfit on, limited labeled data, which impedes LRE models from further generalizing beyond the annotations.

To improve the generalization ability for LRE, we propose to use existing annotations as a guide-

[1]https://github.com/THU-BPM/GradLRE
[†]Corresponding Authors.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2737–2746
November 7–11, 2021.

line instead of having them directly involved in training, as well as introducing an explicit feedback loop when consuming annotations. More specifically, we first encourage pseudo-labeled data to imitate labeled data on the gradient descent directions during the optimization process. We illustrate this idea in Figure 1. g_l represents the average gradient descent direction on labeled data. g_u and $g_u^{'}$ represent the correct and incorrect pseudo labels on unlabeled data, which guides the gradient descent direction in a positive/negative fashion (Du et al., 2018; Sariyildiz and Cinbis, 2019; Yu et al., 2020). Based on how well the pseudo-labeled data mimics the instructive gradient descent direction obtained from limited labeled data, we then design a reward to quantify the behavior and aim to use the reward as an explicit feedback. This learnable setting can be naturally formulated into a reinforcement learning framework, which aims to learn an imitation policy that maximizes the reward through trial and error. When comparing with methods where annotations are directly used in the traditional learning schema, this formulation also allows a feedback mechanism and thus increases generalization ability beyond limited annotations. We name our method as Gradient Imitation Reinforcement Learning in this paper.

We propose a framework called `GradLRE`, which integrates Gradient Imitation Reinforcement Learning and is able to handle two major scenarios in LRE: 1) a typical scenario when limited labeled data and large amounts of unlabeled data are available, and an extreme yet practical scenario where 2) even unlabeled data is absent: only limited labeled data is available. `GradLRE` handles the former scenario via pseudo labeling optimized through Gradient Imitation Reinforcement Learning and tackles the later scenario by using a Contextualized Data Augmentation module.

To summarize, the main contributions of this work are as follows:

- We propose a gradient imitation reinforcement learning method that alleviates the bias from training directly with limited annotation, and encourages the RE model to effectively generalize beyond limited annotations.
- We develop a LRE framework `GradLRE` that handles two low-resource relation extraction scenarios by leveraging both Gradient Imitation Reinforcement Learning and Contextualized Data Augmentation.

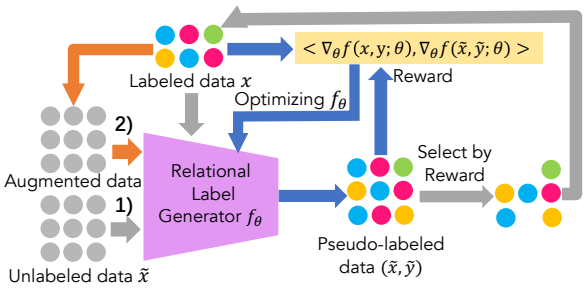

Figure 2: Overview of the proposed `GradLRE` framework for Low Resource Relation Extraction. 1) and 2) represent two LRE scenarios, respectively. Blue arrows represent Gradient Imitation Reinforcement Learning and Orange represents Contextualized Data Augmentation.

- We show that `GradLRE` outperforms strong baselines. Extensive experiments validate the effectiveness of the proposed techniques.

2 Proposed Model

The proposed framework `GradLRE` consists of three modules: Relational Label Generator (RLG), Gradient Imitation Reinforcement Learning (GIRL) and Contextualized Data Augmentation (CDA). As illustrated in Figure 2, two low resource relation extraction scenarios are handled. For the first scenario where limited labeled data and large amounts of unlabeled data are available, the input of RLG is labeled data and unlabeled data. Labeled data consists of sentences and relation mentions: [Sentence, Entity_1, Entity_2, Relation]. For the second scenario where only limited labeled data is available, we adopt CDA to generate unlabeled data and utilize these unlabeled data the same way as in the first scenario.

In a traditional self-training setting, we fine-tune RLG directly using the labeled data, and let RLG assign pseudo labels on unlabeled data as pseudo-labeled data. However, we argue that such learning paradigm suffers from selection bias due to the lack of feedback loops: the bias occurs when a model itself influences the generation of data which is later used for training. In this work, we complete the feedback loop and alleviate such bias by leveraging GIRL to learn a policy that maximizes the likelihood between the expected gradient optimization direction from pseudo labels, and the average gradient optimization direction on labeled data.

2.1 Relational Label Generator

The Relational Label Generator (RLG) aims to obtain contextualized relational features for each in-

put sentence based on the entity pair, and classify the entity pair into specific relations. In this work, we assume named entities in the sentence have been recognized in advance.

For a sequence of words in a sentence x where two entities $E1$ and $E2$ are mentioned, we follow the labeling schema adopted in Soares et al. (2019) and argument x with four reserved tokens to mark the beginning and the end of each entity. We inject the $[E1]$, $[/E1]$, $[E2]$, $[/E2]$ to x as the input token sequence for RLG, for example, "A $[E1]$ letter $[/E1]$ was delivered to my $[E2]$ office $[/E2]$...". Considering that the relational representation between entity pairs are usually contained in the context, we leverage pretrained deep bi-directional transformers networks: BERT (Devlin et al., 2019) to effectively encode entity pairs, along with their context information. We concatenate the outputs corresponding to $[E1]$, $[E2]$ positions as $\mathbf{h} \in \mathbb{R}^{2 \cdot h_R}$ where $\mathbf{h} = [\mathbf{h}_{[E1]}, \mathbf{h}_{[E2]}]$ and h_R is the contextualized relational representation length. The RLG then classifies these representations into specific relations with a fully connected network. We adopt this architecture to generate labels on sentences, and denoted the RLG process as $f_\theta(x, E1, E2)$.

2.2 Gradient Imitation Reinforcement Learning

Generally, we assign pseudo labels via RLG on unlabeled data as pseudo-labeled data, and add the selected pseudo-labeled data into the existing labeled data to iteratively improve RLG. We argue that without a feedback loop measuring the quality of pseudo labels, the model is more likely to suffer from selection bias and is impeded towards a better generalization ability.

We aim to generate pseudo labels with less labeling biases and errors especially with scarce annotations. To achieve this goal, we focus on improving the RLG performance by introducing gradient imitation to define and quantify what an appealing behavior looks like. We define the partial derivatives of the loss function corresponding to RLG parameters on the labeled data as standard gradient descending, and assume that when pseudo-labeled data are correctly labeled in RLG, partial derivatives to the RLG parameters on the pseudo-labeled data would be highly similar to standard gradient descending. Following this assumption, we propose Gradient Imitation Reinforcement Learning (GIRL), which optimizes RLG under a reinforcement learning framework (Williams, 1992). Now we explain the reinforcement learning process in detail.

State: State is used to signal the optimization status. We use $s^{(t)}$ to denote the state. $s^{(t)}$ consists of the updated labeled dataset $\mathcal{D}_l$ at step t, along with a standard gradient direction g_l at step t.

Policy: Our policy is learned to assign correct pseudo label on unlabeled data. The policy network is parameterized by the RLG network f_θ.

Action: The action is to predict relational label on unlabeled data $\widetilde{x}^{(t)}$ as pseudo-labeled data $(\widetilde{x}^{(t)}, \widetilde{y}^{(t)})$ given the State at step t. We consider the relation that corresponds to the maximum probability after softmax as the pseudo label:

$$\widetilde{y}^{(t)} = \text{argmax}(f_\theta(\widetilde{x}^{(t,E1,E2)})). \tag{1}$$

Reward: We use reward to signal labeling biases from the current policy on pseudo-labeled data. Our goal is to minimize the approximation error of the gradients obtained over the pseudo-labeled data. In other words, we maximize the correlation between gradients over the pseudo-labeled data and those over the labeled data.

We define the standard gradient descent direction on the all N labeled data as g_l and the expected gradient descent direction on the pseudo-labeled data as g_p respectively:

$$g_l{}^{(n)}(\theta) = \nabla_\theta \mathcal{L}_l \left(x^{(n)}, y^{(n)}; \theta \right), \tag{2}$$

$$g_p{}^{(t)}(\theta) = \nabla_\theta \mathcal{L}_p \left(\widetilde{x}^{(t)}, \widetilde{y}^{(t)}; \theta \right), \tag{3}$$

where ∇_θ refers to the partial derivatives of the cross entropy loss $\mathcal{L}$ corresponding to Policy f_θ with respect to θ. Considering that the outliers in the labeled data will affect the direction of standard gradient descent, we approximate g_l over all N labeled data and we define $\mathcal{L}_l$ and $\mathcal{L}_p$ as:

$$\mathcal{L}_l = \frac{1}{N} \sum_{n=1}^{N} loss(f_\theta(x^{(n,E1,E2)}), \text{one\_hot}(y^{(n)})), \tag{4}$$

$$\mathcal{L}_p = loss(f_\theta(\widetilde{x}^{(t,E1,E2)}), \text{one\_hot}(\widetilde{y}^{(t)})), \tag{5}$$

where $loss$ is the cross entropy loss function, $f_\theta(x^{(n,E1,E2)})$ returns a probability distribution over all relation categories for the n-th sample and $\text{one\_hot}(y^{(n)})$ returns a one-hot vector indicating the target label assignment.

Since the most important guidance obtained by the gradient vector g_l is its gradient descending direction, so we measure the discrepancy between g_l and g_p for state $s^{(t)}$ by defining their cosine similarity as the reward:

$$R^{(t)} = \frac{g_l(\theta)^{\mathrm{T}} g_p(\theta)}{\|g_l(\theta)\|_2 \, \|g_p(\theta)\|_2}. \tag{6}$$

The range of $R^{(t)}$ is [-1,1]. For those pseudo-labeled data $(\widetilde{x}^{(t)}, \widetilde{y}^{(t)}) \in \mathcal{D}_p$ with $R^{(t)} > \lambda$, $\lambda = 0.5$, we treat them as positive reinforcement to improve the generalization ability of RLG network. We add these selected pseudo-labeled data to the labeled data and correct the standard gradient descending direction:

$$\mathcal{D}_l \leftarrow \mathcal{D}_l \cup \mathcal{D}_p, \tag{7}$$

$$g_l \leftarrow \frac{1}{N+1}(N g_l + g_p). \tag{8}$$

For Eq. (8), we set the weight of the updated gradient direction according to the number of samples, where the standard gradient direction is calculated using all N labeled samples and each pseudo labeled sample. The positive feedback obtained from GIRL via trial and error can attribute the improvement of RLG network (Policy) to assign correct pseudo label for next unlabeled data $\widetilde{x}^{(t)}$ (State).

Reinforcement Learning Loss

We adopt the REINFORCE algorithm (Williams, 1992) and Policy Gradient for optimization. We calculate the loss over a batch of pseudo-labeled samples. The RLG will be optimized by GIRL on each batch according to the following reinforcement learning loss:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} loss\big(f_\theta(\widetilde{x}^{(t,E1,E2)}), \text{one\_hot}(\widetilde{y}^{(t)})\big) * R^{(t)}, \tag{9}$$

where $loss$ is the cross entropy loss function, $R^{(t)}$ is the reward and $\widetilde{y}^{(t)} \sim \pi(\cdot|\widetilde{x}^{(t,E1,E2)}; \theta)$. The π function means Policy in reinforcement learning. In our setting, it is parameterized as f_θ, which is learned to assign pseudo labels on unlabeled data and we minimize $\mathcal{L}(\theta)$ to optimize the θ. T represents a total number of time steps in a reinforcement learning episode and is set to 16, the same number as the batch size. For each high reward $R^{(t)} > \lambda$, $\lambda = 0.5$ pseudo-labeled data, we use it to dynamically update the labeled dataset / standard gradient direction and guide the reinforcement learning process to the next State.

Note that f_θ is first pretrained using all the labeled data in a supervised way. During the process of calculating reinforcement learning loss, our model follows the Markov's decision process and the labeled data $\mathcal{D}_l$ and standard gradient descending direction g_l will be dynamically corrected by the selected pseudo-labeled data $\mathcal{D}_p$, which means that for each State, Policy will be updated over time t. The RLG could solicit positive feedback obtained using GIRL via trial and error.

2.3 Contextualized Data Augmentation

Except the typical LRE scenario where both limited labeled data and large amounts of unlabeled data are available, `GradLRE` handles an extreme yet practical LRE scenario additionally, where only limited labeled data is available. As shown by the orange arrow in Figure 2, we propose to use a contextualized augmentation method, namely CDA, to generate more unlabeled data.

Given a sentence x where two entities $E1$ and $E2$ are mentioned in the labeled data, CDA samples spans of the sentence as `[MASK]` until the masking budget has been spent (e.g., 15% of x) and finally fills the mask with tokens using the pretrained language model. Inspired by Joshi et al. (2020), we sample a span length from a geometric distribution $\ell \sim \text{Geo}(p)$ where $\ell \in [1, 10]$. p will affect the probability of selecting different span lengths. A larger p leads to a shorter span. We follow Joshi et al. (2020) and choose $p = 0.2$. The $\text{Geo}(0.2)$ yields a mean span length of $(\ell) = 3.8$ and shorter spans are more inclined to be chosen. We skip $E1$ and $E2$ as `[MASK]` and also require the starting point of the span must be the beginning of one word which ensures to mask complete words.

For example, we may mask the word DELIVERED TO in "A letter was delivered to my office in this morning." and obtain an augmented sentence "A letter was sent from my office in this morning.". Compared with the original labeled data, the augmented sentence may have a different relation label. We therefore use RLG, which has a strong discriminate power, to assign a correct label to the augmented unlabeled sentence. Since "no relation" has been defined as one valid relation category in the dataset, RLG has the capability to safely assign one augmented sentence as "no relation" when it is out of scope.

Methods / %Labeled Data	SemEval			TACRED		
	5%	10%	30%	3%	10%	15%
LSTM (Hochreiter and Schmidhuber, 1997)	22.65±3.35	32.87±6.79	63.87±0.65	28.68±4.29	46.79±0.99	49.42±0.59
PCNN (Zeng et al., 2015)	41.82±4.48	51.34±1.87	63.72±0.51	40.02±5.23	50.35±3.28	52.50±0.39
PRNN (Zhang et al., 2017)	55.34±1.08	62.63±1.42	69.02±1.01	39.11±1.92	52.23±1.20	54.55±1.92
BERT (Devlin et al., 2019)	70.71±1.24	71.93±0.99	78.55±0.87	40.11±3.88	53.17±1.67	55.55±0.82
Self-Training$_{BERT}$ (Rosenberg et al., 2005)	71.34±1.68	74.25±1.10	81.71±0.79	42.11± 1.04	54.17±0.53	56.52±0.40
Mean-Teacher$_{BERT}$ (Tarvainen and Valpola, 2017)	70.05±3.89	73.37±1.42	80.61±0.81	44.34±1.78	53.08±1.01	53.79±1.38
RE-Ensemble$_{BERT}$ (Lin et al., 2019)	72.35±2.63	75.71±1.39	81.34±0.74	42.78±1.89	54.83±0.95	55.68±1.21
DualRE-Pairwise$_{BERT}$ (Lin et al., 2019)	74.35±1.76	77.13±1.10	82.88±0.67	43.06±1.73	56.03±0.55	57.99±0.67
DualRE-Pointwise$_{BERT}$ (Lin et al., 2019)	74.02±1.68	77.11±1.02	82.91±0.62	43.73±1.60	56.28±0.61	57.72±0.49
MRefG$_{BERT}$ (Li and Qian, 2020)	75.48±1.34	77.96±0.90	83.24±0.71	43.81±1.44	55.42±1.40	58.21±0.71
MetaSRE$_{BERT}$ (Hu et al., 2021)	78.33±0.92	80.09±0.78	84.81±0.44	46.16±1.02	56.95±0.34	58.94±0.36
GradLRE$_{BERT}$ (Ours)	**79.65±0.68**	**81.69±0.57**	**85.52±0.34**	**47.37±0.74**	**58.20±0.33**	**59.93±0.31**
BERT w. gold labels	84.64±0.28	85.40±0.34	87.08±0.23	62.93±0.41	63.66±0.23	64.69±0.29

Table 1: F1 (%) comparisons on the SemEval and TACRED datasets with various amounts of labeled data and 50% unlabeled data.

3 Experiments

We conduct extensive experiments on two datasets to prove the effectiveness of our Gradient Imitation Reinforcement Learning for low resource relation extraction tasks, and give a detailed analysis of each module to show the advantages of GradLRE.

3.1 Datasets

We follow Hu et al. (2021) to conduct experiments on two public RE datasets, including the SemEval 2010 Task 8 (**SemEval**) (Hendrickx et al., 2010), and the TAC Relation Extraction Dataset (**TACRED**) (Zhang et al., 2017). SemEval is a standard benchmark dataset for evaluating relation extraction models, which consists of training, validation, test set with 7199, 800, 1864 relation mentions respectively, with 19 relations types in total (including *no_relation*), of which *no_relation* percentage is 17.4%. TACRED is a large-scale crowd-sourced relation extraction dataset which is collected from all the prior TAC KBP relation schema. The dataset consists of training, validation, test set with 75049, 25763, 18659 relation mentions respectively, with 42 relation types in total (including *no_relation*), of which *no_relation* percentage is 78.7%.

3.2 Baselines and Evaluation metrics

GradLRE is flexible to integrate different contextualized encoders. From Table 1, we first compare several widely used supervised relation encoders with only labeled data: **LSTM** (Hochreiter and Schmidhuber, 1997), **PCNN** (Zeng et al., 2015), **PRNN** (Zhang et al., 2017), **BERT** (Devlin et al., 2019). Among them, BERT achieved the state-of-the-art performance. So we adopt BERT as the base encoder for both GradLRE and other baselines for a fair comparison.

For baselines, we compare GradLRE with other six representative methods: (1) **Self-Training** (Rosenberg et al., 2005) iteratively improves model by predicting unlabeled data with pseudo labels and adds these pseudo label data to labeled data. (2) **Mean-Teacher** (Tarvainen and Valpola, 2017) is jointly optimized by a perturbation-based loss and a training loss to ensure that the model makes consistent predictions on similar data. (3) **DualRE** (Lin et al., 2019) treats relation extraction as a dual task from relations to sentences and combines the loss of a prediction module and a sentence retrieval module. The difference between Pairwise and Pointwise schemes lie in whether the retrieved documents are given scores or a relative order. (4) **RE-Ensemble** (Lin et al., 2019) replaces the retrieval module in the proposed DualRE framework with the same prediction module. (5) **MRefG** (Li and Qian, 2020) semantically connects the unlabeled data to the labeled data by constructing reference graphs, including entity reference, verb reference and semantics reference. (6) **MetaSRE** (Hu et al., 2021) is the state-of-the-art method that generates pseudo labels on unlabeled data by meta learning from the successful and failed attempts on classification module as an additional meta-objective.

Finally, we present another model: **BERT w. gold labels**, which indicates the upper bound of LRE models when all unlabeled data has gold labels during training with labeled data.

For the evaluation metrics, we choose F1 score as the main metric. Note that following Hu et al. (2021), the correct predictions of *no_relation* are ignored.

3.3 Implementation Details

For the two datasets, strictly following the settings used in Hu et al. (2021), we use stratified sampling to divide training set into labeled and unlabeled datasets of various proportions to ensure all subsets share the same relation label distribution. For SemEval, we sample 5%, 10% and 30% of the training set, for TACRED, we sample 3%, 10% and 15% of the training set as labeled datasets. For both datasets, we sample 50% of the training set as unlabeled dataset. As suggested in Hu et al. (2021), we split all unlabeled data into 10 segments. In each iteration, RLG is optimized based on one segment of the data. The RLG gradually improves as we obtain more high-quality pseudo labels one iteration after another. We implement this strategy for our model and the baselines. For the evaluation metrics, we choose F1 score as the main metric.

For RLG, we use the BERT default tokenizer with max-length as 128 to preprocess data. We use pretrained BERT-Base_Cased as the initial parameter to encode contextualized entity-level representation. The fully connected network is defined with layer dimensions of $2h_R$-h_R-label_size, where $h_R = 768$. We use BertAdam with $1e{-}4$ learning rate and warmup with 0.1 to optimize the loss. For GIRL, the total time step T is set to 16, the same number as the batch size. We use AdamW (Loshchilov and Hutter, 2018) with $5e{-}5$ learning rate to optimize the reinforcement learning loss.

3.4 Main Results

Table 1 shows the mean and standard deviation F1 results with 5 runs of training and testing on SemEval and TACRED when leveraging various labeled data and 50% unlabeled data. All methods could gain performance improvements from the unlabeled data when compared with the model that only uses labeled data (BERT), which demonstrates the effectiveness of unlabeled data in the LRE setting. We could observe that `GradLRE` outperforms all baseline models consistently. More specifically, compared with the previous SOTA model MetaSRE, `GradLRE` on average achieves 1.21% higher F1 on SemEval and 1.15% higher F1 on TACRED across various labeled data. When considering standard deviation, `GradLRE` is also more robust than all the baselines.

Considering LRE when labeled data is very scarce, e.g. 5% for SemEval and 3% for TACRED, `GradLRE` could achieve an average 1.27% F1 boost compared with MetaSRE. When more labeled data is available, 30% for SemEval and 15% for TACRED, the average F1 improvement is consistent, but reduced to 0.85%. We attribute the consistent improvement of `GradLRE` to the explicit feedback which GIRL is adopted and learning via trial and error: we use Gradient Imitation as a proxy for the classification loss in optimizing RLG. The guidance from the gradient direction, as a part of the gradient imitation process, is more instructive, explicit, and generalizable than the implicit signals from training directly on labeled data.

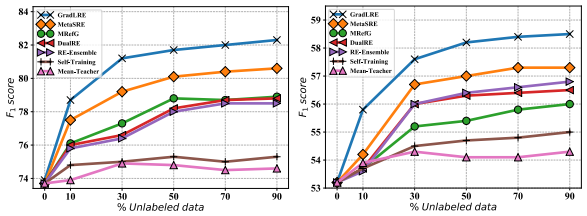

Figure 3: F1 (%) Performance with various unlabeled data and 10% labeled data on SemEval (left) and TACRED (right).

We further vary the ratio of unlabeled data and report performance in Figure 3. F1 performance on a fixed 10% labeled data and 10%, 30%, 50%, 70%, 90% unlabeled data are reported. Note that both labeled data and unlabeled data come from the training set, so we can provide unlabeled data with an upper limit of 90%. We could see that almost all methods have performance gains with the addition of unlabeled data and `GradLRE` achieves consistently better F1 performance, with a clear margin, when comparing with baselines under all different ratios of unlabeled data.

3.5 Analysis and Discussion

Effectiveness of Gradient Imitation Reinforcement Learning

The main purpose of GIRL is to guide RLG to generate pseudo labels with the similar optimization outcomes as labeled data on the unlabeled data. GIRL minimizes the discrepancy between the gradient vectors obtained from the labeled data and generated data. To demonstrate the effectiveness of Gradient Imitation Reinforcement Learning, we first conduct an ablation study in this section. `GradLRE` w/o Gradient Imitation Reinforcement Learning is essentially the same as the Self-Training$_{BERT}$ baseline, which iteratively updates model with the synthetic set containing labeled data and generated data without Gradient Imitation Re-

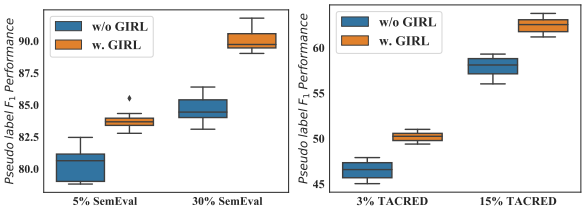

Figure 4: Pseudo label F1 (%) Performance with GIRL based on SemEval (left) and TACRED (right).

My *brother* has entered my *room* without knocking.
Label: **Entity-Destination**
Prediction w/o GIRL: **Other**
Prediction w. GIRL: **Entity-Destination**

The *disc* in a disc *music box* plays this function, with pins perpendicular to the plane surface...
Label: **Content-Container**
Prediction w/o GIRL: **Component-Whole**
Prediction w. GIRL: **Content-Container**

Ditto for his funny turn as a *man* who instigates the *kidnapping* of his own wife in ...
Label: **Cause-Effect**
Prediction w/o GIRL: **Other**
Prediction w. GIRL: **Cause-Effect**

Table 2: Predictions with/without GIRL on SemEval, where *red* and *blue* represent head and tail entities respectively.

inforcement Learning. From Table 1, we observe GradLRE w/o Gradient Imitation Reinforcement Learning (Self-Training$_{BERT}$) gives us 5.38% loss on F1, averaged over all various amounts of labeled data on two datasets.

We identify that the performance gains of GradLRE come from the improved pseudo label quality by adopting GIRL. To validate this, we draw a box plot to show the pseudo label F1. From Figure 4, we could find for the two datasets with different ratios of the labeled data, GIRL could un-

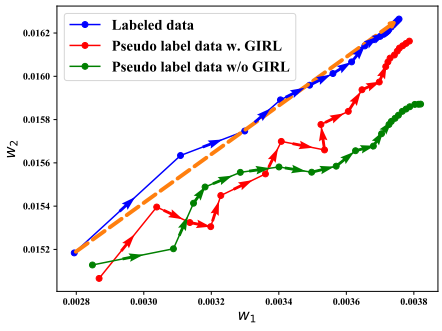

Figure 5: GradLRE gradient descent directions on labeled data and pseudo label data. The dotted line indicates the average gradient direction on labeled data.

	% Labeled Data	L	L + CDA	L + U
SemEval	5%	72.71	75.52	79.65
	10%	73.93	81.47	81.69
	30%	80.55	84.63	85.52
TACRED	3%	41.11	43.34	47.37
	10%	53.23	57.07	58.20
	15%	55.35	58.89	59.93

Table 3: F1 (%) of GradLRE with various percentages of labeled data under different LRE scenarios.

doubtedly improve the F1 performance of pseudo labels. In the case of 30% SemEval and 15% TACRED where labeled data is less scarce, GIRL can obtain more accurate gradient directions based on an increased set of labeled data. As a result, pseudo label performance improvements are more significant.

More specifically, we show the gradient descent direction of GradLRE on labeled data and pseudo label data in Figure 5. Considering the overly-large parameters in RLG, we use Principal Component Analysis (Wold et al., 1987) to reduce the dimension of the parameters to 2, and reflect the direction of gradient descent according to the update of the parameters. Although the optimization direction of pseudo label data fluctuates at the beginning, GIRL is gradually improving and ends up closer to the ideal local minima. When GIRL is not used, the optimization is appealing at the first because of the initial positive gains from the self-training schema. However, the error-prone pseudo labels obtained without instructive feedback gradually push the optimization away from the local minima, which leads to reduced generalization ability.

We further study cases where pseudo labels are improved with GIRL on SemEval, and present them in Table 2. GradLRE w/o GIRL tends to predict the pseudo label as Other with the most occurrences, most likely because Other being the dominating class in the dataset. GradLRE w. GIRL is less sensitive to the label distribution in the data and assigns correct labels. We also observe cases where GIRL is doing better at distinguishing the nuances between similar relations such as Content-Container and Component-Whole.

Handling various LRE scenarios

Considering both labeled/unlabeled data as the resource, we introduce the following LRE scenarios: 1) L+U: Limited labeled data and 50% unlabeled data. 2) L+CDA: Only limited labeled data is available. No unlabeled data is available – we lever-

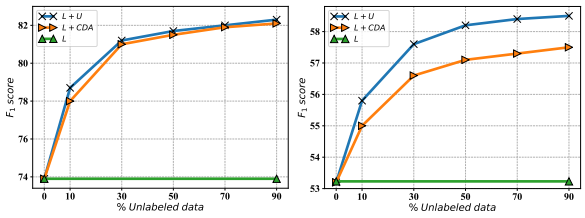

Figure 6: F1 (%) Performance with various unlabeled data and 10% labeled data on SemEval (left) and TACRED (right).

Original: A *letter* was *delivered to* my *office* in ...
Label: **Entity-Destination**
Generated: A *letter* was *sent from* my *office* in ...
Pseudo label: **Entity-Origin**
Original: The *editor* improved the *manuscript* with *his changes*.
Label: **Product-Producer**
Generated: The *editor* improved the *manuscript* with *some improvements*.
Pseudo label: **Product-Producer**
Original: The *suspect dumped* the *dead body* into a local *reservoir*.
Label: **Entity-Destination**
Generated: The *dam bulids* the *human body* into a local *reservoir*.
Pseudo label: **Other**

Table 4: CDA on labeled data to obtain generated data, where *red* and *blue* represent head and tail entities respectively, *cyan* represents the replaced words.

age Contextualized Data Augmentation (CDA) to generate the same amount of data via augmenting the labeled data. 3) L: This is the baseline where the model is trained only on limited labeled data. We present results in Table 3.

Compared to L, L+CDA achieves an average 4.01% improvement in F1, indicating the effectiveness of augmentation. We also observe that L+CDA obtain competitive performance when compared with L+U on SemEval. On a more challenging TACRED dataset, L+CDA achieves only 2.07% less in F1, comparing with L+U when 6.36x less total samples are initially acquired.

We also vary the ratio of unlabeled data (accessible by L+U or augmented using L+CDA). From Figure 6, L+CDA outperforms L consistently, with the ratio of unlabeled data increasing, L+CDA can get more discriminative data and obtain better performance: it can achieve almost the same performance as L+U on SemEval. On TACRED, performance difference is less than 1.53% using various ratio of unlabeled data.

We show some sample generated data produced by CDA in Table 4. BERT Masked Language Model could generate replacement words based on the context information. We find that some part of the sentences with the replaced words could still maintain the original relational information, although the semantic information of another part of the sentence has changed, the RLG can still have the capability to classify the sentence into the most suitable relation.

4 Related Work

Relation Extraction aims to predict the binary relation between two entities in a sentence. Recent literature leverage deep neural network to encode the features among two entities from sentences, and then classify these features into pre-defined specific relation categories. These methods could gain decent performance when sufficient labeled data is available (Zeng et al., 2015; Zhang et al., 2017; Guo et al., 2020; Nan et al., 2020). However, it's labor-intensive to obtain large amounts of manual annotations on corpus.

Low resource Relation Extraction methods gained a lot of attention recently (Levy et al., 2017; Tarvainen and Valpola, 2017; Lin et al., 2019; Li and Qian, 2020; Hu et al., 2021, 2020), since these methods require fewer labeled data and deep neural networks could expand limited labeled information by exploiting information on unlabeled data to iteratively improve the performance. One major method is the self-training work proposed by Rosenberg et al. (2005). Self-training incrementally assigns pseudo labels to unlabeled data and leverages these pseudo labels to iteratively improve the classification capability of the model. However, these methods always endure gradual drift problem (Curran et al., 2007; Zhang et al., 2016; Arazo et al., 2019; Han et al., 2018; Jiang et al., 2018; Liu et al., 2021): during the training process, the generated pseudo label data contains noise and could not been corrected through the model itself. Using these pseudo label data iteratively cause the model to deviate from the global minima. Our work alleviates this problem by encouraging pseudo-labeled data to imitate the gradient optimization direction on the labeled data, and introducing an effective feedback loop to improve generalization ability via reinforcement learning.

Reinforcement Learning is widely used in Nature Language Processing (Narasimhan et al., 2016; Li et al., 2016; Su et al., 2016; Yu et al., 2017;

Takanobu et al., 2019). These methods are all designed with rewards to force the correct actions to be executed during the model training process, so as to improve model performance. Zeng et al. (2019) applies policy gradient method to model future reward in a joint entity and relation extraction task. In our work, we define reward as the cosine similarity between gradient vectors calculated from pseudo-labeled data and labeled data.

Data augmentation methods are leveraged in natural language processing to improve the generalization ability of the model by generating discriminative samples (Kobayashi, 2018; Dai and Adel, 2020; Kumar et al., 2020). Gao et al. (2019) contextually augment data by replacing the one-hot representation of a word by a distribution provided by BERT over the vocabulary. However, they only consider the replacement of a word which limits its capability to expand the sentence semantics(Joshi et al., 2020). In our work, we use [MASK] to replace a span of words and leverage BERT Masked Language Modeling task to fill the [MASK].

5 Conclusion

In this paper, we propose a reinforcement learning framework model `GradLRE` for low resource RE. Different from conventional self-training models which endure gradual drift when generating pseudo labels, our model encourages pseudo-labeled data to imitate the gradient optimization direction in labeled data to improve the pseudo label quality. We find our learning paradigm gives more instructive, explicit, and generalizable signals than the implicit signals that are obtained by training model directly with labeled data. Contextualized data augmentation is proposed to handle the extremely low resource RE situation where no unlabeled data is available. Experiments on two public datasets show effectiveness of `GradLRE` and augmented data over competitive baselines.

Acknowledgments

We thank the reviewers for their valuable comments. The work was supported by the National Key Research and Development Program of China (No. 2019YFB1704003), the National Nature Science Foundation of China (No. 71690231 and No. 62021002), NSF under grants III-1763325, III-1909323, III-2106758, SaTC-1930941, Tsinghua BNRist and Beijing Key Laboratory of Industrial Bigdata System and Application.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *Proc. of ICML*, pages 312–321.

James R Curran, Tara Murphy, and Bernhard Scholz. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *Proc. of PACL*, volume 6, pages 172–180. Bali.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proc. of COLING*, pages 3861–3867.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186.

Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. 2018. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proc. of ACL*, pages 5539–5544.

Zhijiang Guo, Guoshun Nan, Wei Lu, and Shay B. Cohen. 2020. Learning latent forests for medical relation extraction. In *Proc. of IJCAI*, pages 3651–3657. ijcai.org.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, volume 31.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of SemEval*, pages 33–38.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xuming Hu, Fukun Ma, Chenyao Liu, Chenwei Zhang, Lijie Wen, and Philip S Yu. 2021. Semi-supervised relation extraction via incremental meta self-training. In *Proc. of EMNLP: Findings*.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proc. of EMNLP*, pages 3673–3682, Online.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proc. of ICML*, pages 2304–2313.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *TACL*, 8:64–77.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proc. of NAACL-HLT*, pages 452–457.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proc. of CoNLL*, pages 333–342.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proc. of EMNLP*, pages 1192–1202.

Wanli Li and Tieyun Qian. 2020. Exploit multiple reference graphs for semi-supervised relation extraction. *arXiv preprint arXiv:2010.11383*.

Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. In *Proc. of WWW*, pages 1073–1083.

Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021. Noisy-labeled NER with confidence estimation. In *Proc. of NAACL-HLT*, pages 3437–3445.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

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

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proc. of ACL*, pages 1546–1557.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proc. of EMNLP*, pages 2355–2365.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models.

Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. 2019. Gradient matching generative networks for zero-shot learning. In *Proc. of CVPR*, pages 2168–2178.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proc. of ACL*, pages 2895–2905.

Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Lina M Rojas Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proc. of ACL*, pages 2431–2441.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proc. of AAAI*, volume 33, pages 7072–7079.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proc. of AAAI*, volume 31.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *NeurIPS*, 33.

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

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

Xiangrong Zeng, Shizhu He, Daojian Zeng, Kang Liu, Shengping Liu, and Jun Zhao. 2019. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. In *Proc. of EMNLP-IJCNLP*, pages 367–377.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proc. of EMNLP*, pages 35–45.