DuRE: Dual Contrastive Self Training for Semi-Supervised Relation Extraction

Yuxi Feng[1] and Laks V.S. Lakshmanan[1]
[1]The University of British Columbia, Vancouver, Canada
{fyx14, laks}@cs.ubc.ca

Abstract

Document-level Relation Extraction (RE) aims to extract relation triples from documents. Existing document-RE models typically rely on supervised learning which requires substantial labeled data. To alleviate the amount of human supervision, Self-training (ST) has prospered again in language understanding by augmenting the fine-tuning of big pre-trained models whenever labeled data is insufficient. However, existing ST methods in RE fail to tackle the challenge of *long-tail relations*. In this work, we propose DuRE, a novel ST framework to tackle these problems. DuRE jointly models RE classification and text generation as a dual process. In this way, our model could construct and utilize both pseudo text generated from given labels and pseudo labels predicted from available unlabeled text, which are gradually refined during the ST phase. We proposed a contrastive loss to leverage the signal of the RE classifier to improve generation quality. In addition, we propose a self-adaptive way to sample pseudo text from different relation classes. Experiments on two document-level RE tasks show that DuRE significantly boosts recall and F1 score with comparable precision, especially for long-tail relations against several strong baselines.

1 Introduction

Relation Extraction (RE) from unstructured data sources is a key component of building large-scale knowledge graphs (KG) (Noy et al., 2019; Lehmann et al., 2015). Among all the RE tasks, Document-level RE (Zhou et al., 2021) extracts subject-relation-object triples from documents, which remains daunting due to the significant challenges in modeling long text spans and obtaining high-quality supervision signals. Current document-level relation extraction methods (Zhou et al., 2021; Ru et al., 2021) can discover the semantic relation that holds between two entities under supervised learning. However, these methods typically require lots of manually labeled data for model training, which could be labor-intensive to obtain.

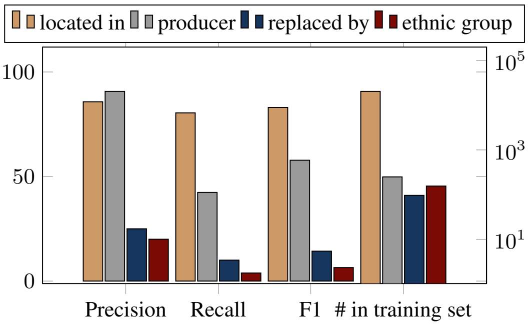


Figure 1: Testing set performance (Precision, Recall, F1, and # of examples in training set) of baseline ATLOP (Zhou et al., 2021) trained on Re-DocRED (Tan et al., 2022b) dataset.

On the other hand, since a large amount of in-domain text is usually accessible, we can tackle document-level RE using semi-supervised learning (Chapelle et al., 2006). There has been substantial work on exploring how to alleviate the amount of human supervision required for RE. Mintz et al. (2009) makes use of distant supervision which leverages external knowledge bases to obtain annotated triples. Since distant supervision makes a strong assumption that the relation between entity pairs should not depend on the context, it usually leads to context-agnostic label noises and sparse matching results.

Alternatively, self-training (ST) (Scudder, 1965; Yarowsky, 1995), a classic semi-supervised learning paradigm, has been proposed in relation extraction (Tan et al., 2023; Hu et al., 2021; Yu et al., 2022). ST minimizes the prohibitively expensive human labeling by iteratively pseudo-annotating unlabeled data with a classifier which is then retrained with the augmented labels. In this way, ST benefits from a vast number of unlabeled instances and extends the generalization bound (Wei et al.,

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 540–555
June 16-21, 2024 ©2024 Association for Computational Linguistics

2021; Zhang et al., 2022b).

A significant challenge of ST is inadequate training data for long-tail relations. As shown in Fig. 1, current document-level RE systems (Zhou et al., 2021) do not perform well on long-tail relations, which hardly appear in the training data. For example, the F1-score for class *located in* is 83.02 while *ethnic group* is only 6.45. The reason could be that the amount of training data is vastly different (20k vs. 155). Assuming that training data and unlabeled data have the same distribution, we cannot expect these long-tail relations to appear sufficiently often in the unlabeled text corpus. To address this, Tan et al. (2023) propose to re-sample training set and to assign more weight to the classes that have high precision and low recall. However, this method does not bring new information to the relation classifier. As a result, these self-training methods might not be able to improve the RE performance on these rare relations.

In order to solve the above issue of long-tail relations, we propose a novel method – Dual contrastive self training for semi-supervised Relation Extraction (DuRE). Unlike previous ST methods (Hu et al., 2021; Tan et al., 2023), we simultaneously train a controllable text generator, generating diverse outputs given specific relation triples. To improve the controllability of the generator, we leverage the signal of the trained RE classifier to label positive and negative generated sequences, and then apply a ranking calibration loss (Zhao et al., 2023) to contrast the positive and negative sequences to improve generation quality. In addition, we propose a self-adaptive way to sample pseudo text from different relation classes. We add noise by increasing generation temperature for relations with higher precision, which introduces diversity to the training set and helps reduce overfitting. Besides, we sample more examples from relations with lower recall. Since long-tail relations usually have a low recall (Fig. 1), they are more likely to be sampled, and thus their recall can be increased through training.

The contributions of this work are as follows:

- We dig into the problem of document-level extraction of long-tail relations and propose to simultaneously train a controllable text generator to address the limitation of previous self-training methods (Tan et al., 2023) that only leverage pseudo-labeling.
- We propose a contrastive loss to control the quality of generated pseudo text, improving the generation quality and thus helping to enhance the classification performance of the relation classifier.
- Comprehensive experiments show that our model significantly improves F1-score in different RE benchmarks on general and biomedical domains, especially on long-tail relations.

2 Related Work

Relation Extraction: Deep neural models have proven to be successful in sentence-level and document-level relation extraction. Zhang et al. (2017) proposed position-aware attention to improve sentence-level RE and published TACRED, which became a widely used RE dataset. Yamada et al. (2020) developed LUKE, which further improved the SOTA performance with entity pretraining and entity-aware attention. Papanikolaou and Pierleoni (2020) proposed to use GPT-2 to generate pseudo data for improving sentence-level RE performance. However, most relations in real-world data can only be extracted based on inter-sentence information. To extract relations across sentence boundaries, recent studies began to explore document-level RE. As previously mentioned, Yao et al. (2019) proposed the popular benchmark dataset DocRED for document-level RE. To address the multilabel problem of Document-level RE, Zhou et al. (2021) proposed using adaptive thresholds to extract all relations of a given entity pair. Zhang et al. (2021) developed the DocUNET model to reformulate document-level RE as a semantic segmentation task and used a U-shaped network architecture to improve the performance of DocRE. Tan et al. (2022a) proposed the use of knowledge distillation and focal loss to denoise the distantly supervised data for DocRE. Wang et al. (2022) proposed a positive-unlabeled learning algorithm under incomplete annotation scenario. However, *the methods above were not designed to tackle the challenge of long-tail relations.*

Self-training: Recently, Self-training has flourished again by iteratively generating pseudo labels and augmenting the tuning of data-hungry language models, showing great advantages in further enhancing NLU (Meng et al., 2020; Vu et al., 2021; Du et al., 2021; Bhat et al., 2021; Chen et al., 2021) and Relation Extraction (RE) (Hu et al., 2021; Yu et al., 2022; Xu et al., 2023), where massive unlabeled input text exists. To tackle the issue of confirmation bias in self-training, Wei et al. (2021) re-samples pseudo-labels based on the frequencies of training examples. Tan et al. (2023) samples different numbers of pseudo-labeled data based on the development set performance. However, it ignores the relations where both precision and recall are low. *All the above ST methods for RE apply ST only in generating pseudo-labels.* Feng et al. (2023b) were the first to propose dual self-training to improve controllable text generation by introducing two kinds of noise. However, they do not study the effectiveness of dual self-training in classification problems (such as relation extraction), and in fact the strategy of adding noise sometimes harms the classification performance.

Unlike all the above ST methods, we are the first to apply dual self-training on document-level relation extraction, generating both **pseudo-labeled data** using RE classifier and **pseudo texts** given specific relation triples using a text generator. The design of contrastive loss and self-adaptive generation for different relation classes further improves the performance, especially on long-tail relations.

3 Method

3.1 Problem Formulation

Document-level relation extraction Given a document x and a set of entities $\mathbf{e} = \{e_j\}_{j=1}^{m}$, the task of document-level relation extraction is to predict a subset of relations from $\mathcal{R} \cup \{N_A\}$ between entity pairs $(e_h, e_t)_{h,t=1...m,h\neq t}$, where $\mathcal{R}$ is a predefined set of relations, N_A represents no relations between given entities, and e_h, e_t are identified as head and tail entities, respectively. At the test time, the model needs to predict the labels of all entity pairs in document x.

Semi-supervised document-level relation extraction Let $\mathbf{x}$ be the text, $\mathbf{e}$ be the entities mentioned in $\mathbf{x}$, and $\mathbf{y} = \{e_{h_j}, r_j, e_{t_j}\}_{j=0}^{n_x}$ be the existing relations in $\mathbf{x}$, $D_L = \{\mathbf{x}_i, \mathbf{y}_i, \mathbf{e}_i\}$ be a labeled dataset with paired text and its corresponding relation sets, and $D_U = \{\mathbf{x}_i, \mathbf{e}_i\}$ be an unlabeled dataset from the same domain. In reality, we do not obtain $\mathbf{e}_i$ for unlabeled text corpus. However, we can use tools of name entity recognition (NER) and coreference resolution (CR) to get the entity list in advance. Since we focus on relation extraction only, we assume we have already obtained the entity list for simplicity.

Controllable text generation given relation triples Given relation triples $\mathbf{y} = \{e_{h_j}, r_j, e_{t_j}\}_{j=1}^{n_x}$, where n_x is the number of relations, the task is to generate document x that contains these relation triples.

3.2 Methodology

We aim to jointly learn an attribute-controllable generator $\mathcal{G} = P_\theta(\mathbf{x}|\mathbf{y})$ parameterized by θ (*e.g.*, a large PLM) to generate, in an auto-regressive manner, high-quality text $\mathbf{x} \sim P_\theta(\mathbf{x}|\mathbf{y})$ containing the given relations $\mathbf{y}$. We also endow our model with the ability to produce pseudo extractions for $\{\mathbf{x}_i, \mathbf{e}_i\} \in D_U$ through jointly learning a Document-RE classifier $\mathcal{C} = P_\phi(\mathbf{y}|\mathbf{x}, \mathbf{e})$. We simultaneously model and optimize $\mathcal{G}$ and $\mathcal{C}$ with a shared PLM as a dual process.

We train our relation classifier using Adaptive Thresholding (AT) loss (Zhou et al., 2021). AT loss aims to learn a threshold class (TH), which are entities-dependent threshold values. The definition of AT loss is as follows:

$$\begin{aligned}
\mathcal{L}_P &= -\sum_{r\in\mathcal{P}_\mathbf{e}} \log\left(\frac{e^{\text{logit}_r}}{\sum_{r'\in\mathcal{P}_\mathbf{e}\cup\{\text{TH}\}} e^{\text{logit}_{r'}}}\right), \\
\mathcal{L}_N &= -\log\left(\frac{e^{\text{logit}_{\text{TH}}}}{\sum_{r'\in\mathcal{N}_\mathbf{e}\cup\{\text{TH}\}} e^{\text{logit}_{r'}}}\right), \\
\mathcal{L}_C &= \mathcal{L}_P + \mathcal{L}_N \qquad (1)
\end{aligned}$$

where positive classes $\mathcal{P}_\mathbf{e} \subseteq \mathcal{R}$ are the relations that exist between the entities in $\mathbf{e}$, negative classes $\mathcal{N}_\mathbf{e} \subseteq \mathcal{R}$ are the relations that do not exist between the entities, logit_r and logit_{TH} are the predicted logits for class r or threshold TH by classifier $\mathcal{C}$. AT loss is a sum of losses of two parts. The first part $\mathcal{L}_P$ pushes the logits of all positive classes to be higher than the TH class. The second part $\mathcal{L}_N$ is a categorical cross-entropy loss with TH class being the true label and pushes the logits of negative classes to be lower than the TH class.

For the generation side, we use cross-entropy loss for auto-regressive generation.

$$\mathcal{L}_G = -\frac{1}{N}\sum_{(\mathbf{x},y)\in D}[\sum_{j=1}^{L}\log P_\theta(\mathbf{x}^j|\mathbf{x}^{<j}, y)], \quad (2)$$

where $\mathbf{x}^j$ means the j-th token in $\mathbf{x}$, L is the length of $\mathbf{x}$, D is the training set with N samples. We will show later how to construct D for different training phases. Finally, we compute a weighted sum of

Algorithm 1: Training Process of DuRE

```
Input: Labeled set $D_L$, unlabeled set $D_U$,
       relation set $\mathcal{R}$.
Jointly train base model $\mathcal{G}, \mathcal{C}$ on $D_L$ by
  optimizing Eq.(3), store the best $\mathcal{G}_0, \mathcal{C}_0$.
for $epoch \leftarrow 1$ to $MaxEpoch$ do
    for $\mathbf{x}_i, \mathbf{e}_i$ in $D_U$ do
        $\hat{y}_i = \mathcal{C}_{epoch-1}(\mathbf{x}_i, \mathbf{e}_i)$
    end
    Build pseudo label set:
      $D_{PL} = \{\mathbf{x}_i, \mathbf{e}_i, \hat{y}_i\}$
    for $r_j$ in $\mathcal{R}$ do
        Sample $n_j$ triples $\{\mathbf{y}\}^{n_j} \subset D_L$ with
          relation $r_j$ following Eq.( 5).
        for $k \leftarrow 0$ to $n_j$ do
            Generate $m$ pseudo texts:
              $\{\mathbf{x}^k\}^m = \{\mathcal{G}_{epoch-1}(\mathbf{y}^k)\}^m$
            Select the entities $\mathbf{e}^k$ by parsing
              $\{\mathbf{x}^k\}$
            Compute pseudo labels with
              $\mathcal{C}_{epoch-1}(\mathbf{x}^k, \mathbf{e}^k)$
            Select positive example
              $\mathbf{x}^+, \mathbf{y}^+, \mathbf{e}^+$ and negative
              example $\mathbf{x}^-, \mathbf{y}^-$
        end
    end
    Build pseudo text:
      $D_{PT} = \{\mathbf{x}^+, \mathbf{y}^+, \mathbf{e}^+, \mathbf{x}^-\}$
    Train $\mathcal{G}_{epoch-1}, \mathcal{C}_{epoch-1}$ on
      $\{D_{PT}, D_{PL}, D_L\}$ by optimizing
      Eq.(3) and Eq.(4), update the
      parameters to $\mathcal{G}_{epoch}$ and $\mathcal{C}_{epoch}$.
end
```

classification and generation loss, where λ_C and λ_G are tunable hyperparameters.

$$\mathcal{L} = \lambda_C \mathcal{L}_C + \lambda_G \mathcal{L}_G. \tag{3}$$

3.3 Self-adaptive pseudo text generation

The full DuRE method is described in Alg. 1. Following the practice of self-training in NLU (Vu et al., 2021), we start ST from a strong base model tuned on D_L and use the full unlabeled D_U to produce pseudo labels, rather than select part of the data with certain criteria as in (Tan et al., 2023). In addition to pseudo-labeled data, we also use our generator to generate pseudo-text given entity-relation triples. To better improve the RE performance, we propose to use Contrastive Loss (CL) and Self-Adaptive Generation (SAG) methods, which are illustrated below.

Contrastive Loss (CL) Following Zhao et al. (2023), we use Sequence Likelihood Calibration (SLiC) to align a language model's sequence likelihood, $P_\theta(x|y)$, over decoded sequences according to their similarity to reference sequences. Given $\mathbf{y}_j = \{e_{h_j}, r_j, e_{t_j}\}$, we generate a batch of m generations $x_j^k = \mathcal{G}(\mathbf{y}_j)$. Then the ranking calibration loss contrasts a positive sequence x^+ and a negative sequence x^-, encouraging the model to assign more probability mass to positive compared to negative sequences, and thus enhancing generation quality. Thanks to the dual learning framework, we can use the classifier $\mathcal{C}$ to evaluate the confidence of whether generation x contains relation triple $\{e\}$. More specifically, $x^+ = \text{argmax}_i logit^i_{r_j}$, and $x^- = \text{argmin}_i logit^i_{r_j}$. Thus the contrastive loss is computed as follows:

$$\mathcal{L}_G = \max\left(0, \beta - \log P_\theta(x^+|y) + \log P_\theta(x^-|y)\right) \tag{4}$$

where $\beta > 0$ is a hyperparameter to control the margin between positive and negative examples.

Self-Adaptive Generation (SAG) Another advantage of using pseudo text is that we can tune the distribution of relations and generate more pseudo texts for long-tail relations. Compared to traditional methods like bootstrapping (Dupret and Koda, 2001) and class-rebalanced self-training (Wei et al., 2021) where duplicated samples are chosen for a specific class, the generator can produce more diverse training examples.

Here we propose a self-adaptive pseudo-text generation strategy. For different relation classes with different development set performances, we generate pseudo text accordingly. For a relation class r, if its recall R_r is low, then we should improve the recall by generating more pseudo examples. If its precision P_r is high, we believe this class is well-predicted. Then the purpose of pseudo-text generation is to augment the training set with noisy data to enhance generalization. A good way to add noise is to generate data with higher temperature (Feng et al., 2023b). Otherwise, if P_r is low, we use a lower temperature to sample a more certain output to ensure the generation quality. In summary, we define the sampling probability ϕ_r and generation temperature temp_r for relation r as fol-

Dataset	Train	Dev	Test	Unlabeled
Re-DocRED				
# Documents	3053	500	500	108000
Avg. # Entities	19.4	19.4	19.6	19.4
Avg. # Triples	28.1	34.6	34.9	-
Avg. # Sentences	7.9	8.2	7.9	8.0
# NA rate	94.3%	93.1%	93.1%	-
CDG (5%)				
# Documents	3847	1480	523	72697
Avg. # Words	196.9	236.5	235.6	197.0
Avg. # Entities	7.4	8.8	10.0	7.6
Avg. # Triples	2.1	2.2	2.6	-
Avg. # Sentences	12.6	14.0	13.2	12.6
# NA rate	96.8%	97.7%	93.8%	-

Table 1: Dataset statistics.

lows.

$$\begin{aligned} \phi_r &\sim (1 - R_r), \\ \text{temp}_r &= \alpha + P_r. \end{aligned} \tag{5}$$

where $\alpha > 0$ is a hyperparameter to control the noise level of generation.

Our SAG method has the following major differences from Tan et al. (2023). (1) Data resampled: Tan et al. (2023) resample the training set while we sample generated pseudo text to generate diverse outputs given specific relation triples. (2) Sampling Strategy: Tan et al. (2023) resample the training set with probability $P_r(1 - R_r)$, which causes relations with low precision to be ignored. On the other hand, we resample with probability $(1 - R_r)$, and encourage the classes with higher precision to generate more diverse outputs (sampling with high temperature). For the relations with low precision, we sample with low temperature to ensure its faithfulness.

4 Experiments

4.1 Datasets

We experimented with our method on two datasets. For general-domain document level RE, we use Re-DocRED (Tan et al., 2022b), a high quality dataset. We use the distantly-labeled set of Re-DocRED as unlabeled set, only keeping named entities information and coreference information but ignoring the distant labels. As a second dataset, we tested our method on the biomedical document-level RE dataset ChemDisGene (CDG) (Zhang et al., 2022a). Since there is no unlabeled set in CDG, we only leveraged part (e.g., 5%) of training set as labeled data and kept aside the rest as unlabeled data. Our models are evaluated on the test sets of Re-DocRED and CDG. Both of the test sets are human-annotated and have high quality. The statistics of the datasets can be found in Table 1.

4.2 Experimental Settings

We use ATLOP (Zhou et al., 2021) as our base RE classifier and an encoder-decoder framework as our pseudo text generator. Following Feng et al. (2023b), we share the encoder part of RE classifier and pseudo text generator to save the total number of training parameters. For Re-DocRED dataset, we use Flan-T5-base (Chung et al., 2022) as the base encoder-decoder model. For CDG dataset, we use a version of Flan-T5-base that is pretrained on Pubmed dataset[1]. We use AdamW (Loshchilov and Hutter, 2019) with learning rate = 5e-5, warm-up rate = 0.06, $\lambda_g = 1$, $\lambda_c = 5$, $\alpha = 0.7$, $\beta = 0.3$, and batch size = 8 for optimization across all tasks. As is common practice (Holtzman et al., 2019), we use the top-p sampling method with $p = 0.9$ for decoding. For the generation task, we add a prompt sentence at the beginning of the input: *Generate text given the following relation triples*. The number of pseudo text is the same as the number of labeled data, i.e., 3053 for Re-DocRED and 3847 for CDG. More implementation details are provided in the Appendix.

4.3 Evaluation Metrices

Following Tan et al. (2023), we used micro-averaged F1 score as the evaluation metric. We also evaluate the F1 score for frequent classes and long-tail classes, denoted as Freq_F1 and LT_F1, respectively. For the Re-DocRED dataset, the frequent classes include the top 10 most popular relation types in the label space; the rest of the classes are categorized as long-tail classes. We also use an additional metric Ign_F1 on the DocRE task. This metric represents the F1 score, calculated for the triples that do not appear in the training data.

4.4 Baselines

We compare our model with the following strong document-level RE baselines, including both supervised and semi-supervised.

Supervised Approaches (1) **ATLOP** (Zhou et al., 2021) A vanilla baseline model for document-level RE. (2) **ATLOP-Flan** Only use Flan-T5

[1] The model can be found at `https://huggingface.co/gubartz/ssc-flan-t5-base-pubmed`.

Model	P	R	F1	Ign_F1	Freq_F1	LT_F1
Bert-base-cased						
ATLOP[†]	**86.70**	62.46	72.61	71.86	75.92	67.46
NS[†]	77.63	69.17	73.16	72.92	77.28	67.59
VST[†]	72.77	**75.55**	74.14	72.48	78.47	68.13
SSR-PU	76.78	71.46	74.33	72.91	78.41	68.32
CREST[†]	75.94	72.47	74.17	72.77	77.93	68.68
CAST[†]	76.59	72.84	74.67	73.32	78.53	69.34
Flan-T5-base						
ATLOP-Flan	86.40	61.78	72.05	71.32	76.15	65.45
ATLOP-Dual	85.17	61.93	71.72	70.95	76.07	64.70
DuRE	79.01	73.84	**76.84**	**75.32**	**79.81**	**72.88**

Table 2: Relation classification results on Re-DoCRED dataset. [†]Results are obtained from Tan et al. (2023).

Model	P	R	F1
PubMedBERT on CDG (100%)			
ATLOP [†]	**76.17**	29.70	42.73
NS [†]	71.54	35.52	47.47
SSR-PU	54.27	43.93	48.56
CREST [†]	59.42	42.12	49.28
CAST [†]	66.68	45.48	54.03
Flan-T5-base on CDG (5%)			
ATLOP-Dual	46.40	21.78	32.05
DuRE	52.01	48.84	50.38
Flan-T5-base on CDG (50%)			
ATLOP-Dual	47.67	51.24	49.39
DuRE	54.91	**59.43**	**57.08**

Table 3: Results on CDG dataset. [†]Results are obtained from Tan et al. (2023).

(Chung et al., 2022) Encoder to train the RE classifier without generator. (3) **ATLOP-Dual** Simultaneously train RE classifier given input documents and text generator given relation triples but without self-training. (4) **Negative Sampling (NS)** (Li et al., 2021): randomly select partial negative samples in training to alleviate the detrimental effect of the false negative problem.

Semi-supervised Approaches All the baselines below leverage ATLOP as their backbone. (1) **Vanilla Self-Training (VST)** (Peng et al., 2019; Jie et al., 2019): a variant of simple self-training where models are trained with N folds, and all pseudo-labels are directly combined with the original labels. (2) **SSR Positive Unlabeled Learning (SSR-PU)** (Wang et al., 2022): SSR-PU utilizes positive unlabeled learning and a shift-and-squared ranking (SSR) loss to accommodate the distribution shifts for the unlabeled examples. (3) **Class Re-balancing Self-Training (CREST)**(Wei et al., 2021): This algorithm re-samples the pseudo-labels generated by models based on the frequencies of the training samples. (4) **Class-Adaptive Self-Training (CAST)** (Tan et al., 2023): this method calculates the precision and recall scores of each class on the development set and uses the calculated scores to compute the sampling probability of each class to alleviate confirmation bias caused by erroneous pseudo labels.

4.5 Results

The experimental results on the test set of Re-DocRED (Table 2) demonstrate that our DuRE achieves consistent performance improvements in terms of F1 scores over all baselines. The F1 difference between the best baseline CAST and our DuRE is 2.17 (76.84 vs. 74.67). We also found that simply adding an additional RE-controlled generation task does not improve the relation classification performance (ATLOP-Flan vs. ATLOP-Dual), where F1 scores decreased slightly (72.05 vs. 71.72). In addition, training on an encoder-decoder framework (Flan-T5-base) does not outperform a single encoder framework (Bert-base) in document-level RE tasks. However, we notice that the gap between CAST and ATLOP(bert) is 2.06, while the gap between ATLOP-DuRE and ATLOP-Dual is 5.12. This indicates that our proposed dual self-training method improves the RE classification quality significantly more compared to the backbone model. We also notice a considerable improvement (+3.54 F1) especially in long-tail relations, showing that self-training on a rebalanced pseudo text is better than simply doing bootstrapping in the existing training set. The reason is that our trained generator can generate more diverse examples, which helps reduce overfitting on training data.

Table 3 presents the experiments on biomedical RE (CDG dataset). Our DuRE model achieves a performance comparable to the baselines with only 5% of the training data. When trained with 50% of the training data, we get the best performance, outperforming the CAST baseline with +3.05 F1. Based on the results of RE experiments in general and biomedical domains, self-training-based methods aim to improve recall and consistently improve overall performance. However, our DuRE maintains a better balance between increasing recall and maintaining high precision.

Model	P	R	F1	Freq_F1	LT_F1
DuRE	80.01	**73.84**	**76.84**	**79.81**	**72.88**
−SAG	79.51	73.24	76.24	79.71	70.68
−SAG−CL	80.44	71.43	75.67	79.24	69.90
−PT	80.04	70.35	74.88	79.14	67.88
−PL	82.71	66.53	73.75	77.81	67.20
−PL−PT	**85.17**	61.93	71.72	76.07	64.70

Table 4: Ablation study on Re-DocRED dataset. Here − means removing components from DuRE. −SAG: remove the self-adaptive generation strategy and sample pseudo texts for different classes in the same setting. −CL: do not generate positive/negative examples and sample random examples. −PT/−PL: do not use pseudo text/labels.

4.6 Ablation Study

We conduct an ablation study on the Re-DocRED dataset. As shown in Table 4, we can see that (i) Self-Adaptive Generation benefits the F1 score for both frequent and long-tail relations; (ii) Contrastive loss further enhances all F1 scores in relation extraction; and (iii) Self-training on pseudo-labeled text leads to an improvement in recall but has relatively low precision, which shows that self-training is able to balance precision and recall. This observation is also consistent with Tan et al. (2023). Full version of Table 4 is included in the Appendix.

4.7 Analysis

Effect of Self-Training We compare our model with a variant (−**Dual**) where we use the base generation model $\mathcal{G}_0$ to generate pseudo text and do not update it through self-training. As depicted in Fig. 2, classification F1 reaches its maximum quickly and then stops increasing. On the other hand, thanks to the simultaneously optimized generator, DuRE keeps improving generation quality and refining pseudo labels, and thus iteratively improves the quality of the relation classifier.

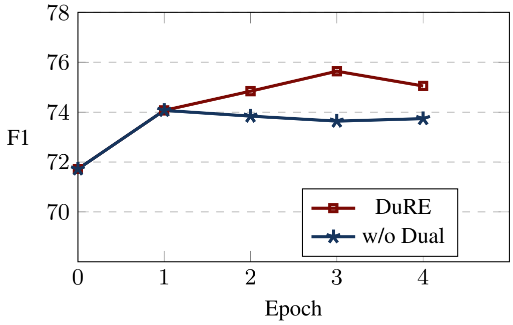


Figure 2: F1 score over the number of training epochs on Re-DocRED.

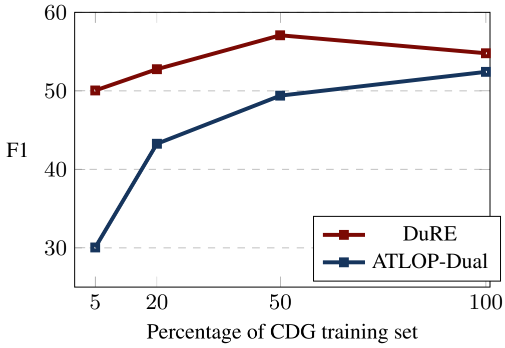
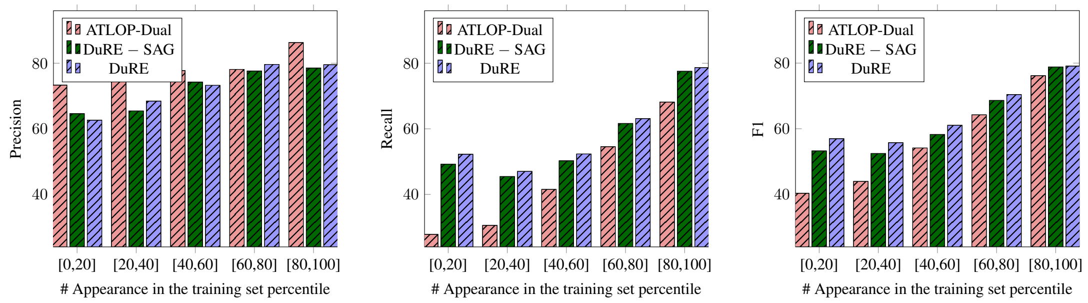


Figure 3: Test F1 score with different numbers of labeled data on CDG dataset. All methods use FLAN-T5 as the base model.

Number of labeled data We also assess our model with varying numbers of labeled training instances, with the remaining instances treated as unlabeled data. We observe consistent superiority of DuRE to ATLOP-Dual model. Indeed, even DuRE (50%) outperforms ATLOP-Dual (100%), showing that our method can work with scenarios with fewer training data. However, we notice a drop in DuRE (100%). We conjecture that the reason is that the training set of CDG itself is distantly labeled, which is noisy and incomplete. Through self-training on pseudo-labeled data, the model can figure out incompletely labeled relations. However, as for training on the full distantly-labeled dataset, DuRE only benefits from generated pseudo texts compared to ATLOP-Dual, which limits the improvement of the classification model on pseudo-labeled data. Notice that the generated pesudo texts might still have higher quality than simply distantly labeled data.



(a) Precision (b) Recall (c) F1

Figure 4: Precision, recall, and F1 score on the development set over different percentile of appearances in the Re-DocRED training set.

Input	* The Invisible Man*, lyrics by, * Roger Taylor*.
Positive	" * The Invisible Man* " is a rock song originally **written by** * Roger Taylor* and * Mark Smith* in * 1942*. The song was sung at the * Royal Academy of Music* (Royal Academy of Music) in * 1952*. The lyrics and compositions of " The Invisible Man " were originally a single with * Taylor*'s voice in a voodoo voice, a version of * Ralph Ellis*'s " * The Man Who Walks Away* ". The medley is a ballad, and is not a traditional ballad. " * The Invisible Man* " is often performed in * England* and * France*, but it was originally recorded as a concert song for the * Royal Academy of Music*, in * 1936*, and in * Japan* (including the final song).
Negative	* The Invisible Man* is the sixth **album by** * Roger Taylor*. It was released in October 1966. The album features a different version of " * The Man In the Sun* " and a different theme. *** The Man In the Sun*** is set in a mystical (and heavily symbolic) future. **It** was **composed by** * Roger Taylor*. The album is very similar to the * The Black Swan* album. It is not necessarily a remake of the album, but rather a remastered version of the earlier album. The album contained a new riff and a new orchestration which was intended to provide additional background music for the final section of the album.

Table 5: Example of generation on Re-DocRED dataset.

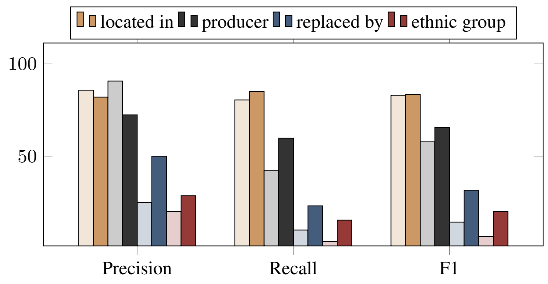


Figure 5: Comparison of testing set performance of baseline ATLOP (light) and DuRE (dark) trained on Re-DocRED dataset.

Performance of relations with different frequency We sorted the relations by their frequency of appearance in the training set from low to high and grouped them into five groups by their percentile. Fig. 4 plots the development set performance of relations with different frequencies. We can see that original ATLOP tends to have much higher precision than recall, while our DuRE consistently improves recall and F1 scores over relations across all frequency groups, at the price of a modest drop in precision. We also notice a tendency that our method improves F1 score more for less frequent relations thanks to the self-adaptive generation, indicating the effectiveness of our methods on long-tail relations.

Improvement on rare relations We plot the precision and recall scores of DuRE and CAST for different relation classes in Fig. 5, where the experimental results are obtained by training with the Re-DocRED dataset. We found that DuRE significantly improves the recall scores of long-tail classes (*producer*, *replaced by*, and *ethnic group*) and thus improves F1 scores, while maintaining the F1 scores of frequent classes (*located in*). We also notice that the improvements in recall scores are accompanied by a decline in precision scores for frequent classes due to our strategy of learning from noisy pseudo texts for frequent classes. Learning from these augmented noisy texts would decrease the threshold and thus improve recall but decrease precision.

Further analysis of contrastive loss Table 5 illustrates a case study of how contrastive loss improves generation quality and thus improves relation classification. Both positive and negative examples are generated under the same decoding strategy. We can see that the negative example does not entail the relation (*The Invisible Man, lyrics by, Roger Taylor*), given that there are similar relations (*The Invisible Man, album by, Roger Taylor*) and (*The Man In the Sun, composed by, Roger Taylor*). This example shows that our learned text generator could sometimes generate text that does not entail the given prompt. Thanks to the contrastive loss, the learned text generator could learn from the signal from the RE classifier, maximize the margin of positive and negative samples, and be more faithful to the given prompt. Detailed analysis of generation quality can be found in Appendix.

5 Conclusion and Future Work

We propose a novel DuRE method to apply Self-training to semi-supervised document-level relation extraction. DuRE (1) jointly optimizes generation and classification via a dual learning framework to leverage both pseudo text and pseudo labels, (2) incorporates contrastive loss to improve the quality of pseudo texts, and (3) applies self-adaptive generation to reduce overfitting of well-predicted relation classes and to improve the performance of long-tail relations. Given that the pseudo data is generated in an auto-regressive manner, which takes longer training time, we plan to explore ways to accelerate the self-training process in the future.

6 Limitations

Though DuRE works well, it has the following limitations:

- Decelerated training process. Like all other Self-training methods, DuRE also needs to reproduce pseudo labels and pseudo text at each ST iteration. Since the pseudo text is generated in an auto-regressive manner, which is hard to be done in parallel, it takes longer training time. Feng et al. (2023a) proposed to use a non-autoregressive generator in self-training, which could be possibly used to accelerate the training speed of DuRE.

- Bias introduced. Bias can be introduced by various forms of supervision, including human annotators or machine synthetic data. We acknowledge that potential bias exists in the pseudo text generated by our DuRE method since DuRE shifts the distribution of training data based on the performance (precision and recall) on the development set. Analyzing the impact of bias found in semi-supervised learning algorithms (like self-training) is an interesting study for future work.

- Reliance on unlabeled in-domain text. As we discussed in Sec. 4, though our proposed contrastive loss and self-adaptive generation bring non-trivial improvement, the overall performance of all ST methods still relies on pseudo labels from unlabeled text. When unlabeled text is extremely inadequate or even unavailable (e.g., low-resource scenarios), how to better utilize pseudo text for further improvement is an open challenge.

- Task generalization and scalability. We mainly investigate document-level RE in this work, in which ST actually acts as a kind of regularization and smoothing. How to apply this paradigm to super large language models (*e.g.*, LLaMa-2 (Touvron et al., 2023) and GPT4 (OpenAI, 2023)) and for tasks beyond RE is also an open question.

7 Ethics Statement

The generative part of our model may be utilized to generate fake information when the input triples are not factual, which could possibly be utilized to produce and propagate disinformation. Also, the generated pseudo-text may contain some socially biased, offensive, or politically sensitive expressions. However, these generated texts are designed to be used as pseudo data in data augmentation to improve the robustness of the relation extraction model and to improve the performance of long-tail relations.

Acknowledgement

Lakshmanan's research was supported in part by a grant from NSERC (Canada) and Feng's research was supported in part by a grant from UBC Data Science Institute.

References

Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee. 2021. Self-training with few-shot rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10702–10712, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. 2006. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Georges Dupret and Masato Koda. 2001. Bootstrap re-sampling for unbalanced data in supervised learning. *European Journal of Operational Research*, 134(1):141–156.

Yuxi Feng, Xiaoyuan Yi, Laks V.S. Lakshmanan, and Xing Xie. 2023a. Kest: Kernel distance based efficient self-training for improving controllable text generation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5049–5057. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yuxi Feng, Xiaoyuan Yi, Xiting Wang, Laks Lakshmanan, V.S., and Xing Xie. 2023b. DuNST: Dual noisy self training for semi-supervised controllable text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8760–8785, Toronto, Canada. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. 2021. Semi-supervised relation extraction via incremental meta self-training. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 487–496, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 729–734, Minneapolis, Minnesota. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yangming Li, lemao liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: Lessons and challenges. *Communications of the ACM*, 62 (8):36–43.

OpenAI. 2023. Gpt-4 technical report.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1239–1250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2023. Class-adaptive self-training for relation extraction with incompletely annotated training data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8630–8643, Toronto, Canada. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4123–4135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. 2021. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynRE:

Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Junjie Yu, Xing Wang, Jiangjiang Zhao, Chunjie Yang, and Wenliang Chen. 2022. STAD: Self-training with ambiguous data for low-resource relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2044–2054, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. 2022a. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. 2022b. How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. In *International Conference on Learning Representations*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Detailed Setting

A.1 Implementation Details

We use pre-trained Flan-T5-base (Chung et al., 2022) as the encoder and decoder of our DuRE model, more suitable for our joint classification and generation schema.

We tuned all hyperparameters only on the held-out development set. In self-training phase, we tuned $\lambda_c \in \{1, 5, 10\}$, $\alpha \in \{0.5, 0.7, 1.0\}$, and $\beta \in \{0.1, 0.3, 0.5\}$ in Re-DocRED dataset to obtain the reported results. Finally, we set $\lambda_c = 5$ and $\lambda_g = 1$ in base-model training phase, while $\lambda_c = 5$, $\lambda_g = 1$, $\alpha = 0.7$, and $\beta = 0.3$ in the self-training phase. We tuned the hyperparameters in Re-DocRED dataset and applied them to all tasks. We use AdamW (Loshchilov and Hutter, 2019) as an optimizer. The training batch size is 8, and the learning rate is $5e-5$. We apply linear warmup to the optimizer, and the warmup ratio is 0.06.

We implement DuRE on Huggingface Transformers (Wolf et al., 2020) library of v4.31.1 and use four NVIDIA Tesla V100 nodes to train our model. The total number of training hours is around 39.62h for Re-DocRED and 34.54h for CDG(5%). The number of parameters of our model is 242.91M. In the generation phase, we use top-p sampling ($p = 0.9$) as the decoding method. Other configuration of the generator includes a length penalty to be 1.0 and a repetition penalty to be 1.0 for all baselines. All experimental results are trained and tested in a single run with fixed random seeds.

For the number of pseudo text generated, following (Feng et al., 2023b), we choose the same number as the labeled training set. The prompt for generation is *Generate text given the following relation triples*. An example of generation input could be *Generate text given the following relation triples. * The Invisible Man*, lyrics by, * Roger Taylor**. For each input triple, we sample 8 sentences and apply the RE classifier to select positive and negative pseudo texts among these samples.

A.2 Generation Evaluation Metric Details

We set the minimum generation length to 100. We evaluate NLG quality on the following metrics:

Fluency: We evaluate generation fluency by the perplexity of generated text measured by GPT2-XL (Radford et al., 2019), *i.e.*, **Output PPL**.

Controllability: We evaluate the control accuracy through classification performance (accuracy (**Acc**)) on the generated text by our DuRE RE classifier.

Diversity: To evaluate the diversity of generated text, we consider **Dist-n** (Li et al., 2016): the percentage of distinct n-grams on generated samples. We evaluate on $n = 1, 2, 3, 4$ and compute the geometric mean as **Dist**. **Dist** emphasizes the amount of novel n-grams within every generation.

Among all the above metrics, F1 and Dist-n are reported as 100 times their original value for convenience.

A.3 Baseline Details

We compare our model with the following strong document-level RE baselines, including both supervised and semi-supervised.

ATLOP (Zhou et al., 2021) A vanilla baseline model for document-level RE. We use bert-base-cased(Devlin et al., 2019) encoder in Re-DocRED dataset and Pubmed-Bert-base(Gu et al., 2021) encoder in CDG dataset.

ATLOP-Flan Only use Flan-T5(Chung et al., 2022) Encoder to train the RE classifier without generator.

ATLOP-Dual Simultaneously train RE classifier given input documents and text generator given relation triples but without self-training.

Negative Sampling (NS) (Li et al., 2021) This method tackles the incomplete annotation problem through negative sampling. To alleviate the effects of false negatives, this method randomly selects partial negative samples for training. Such an approach can help to alleviate the detrimental effect of the false negative problem.

Vanilla Self-Training (VST) (Peng et al., 2019; Jie et al., 2019) VST is a variant of simple self-training. In this approach, models are trained with N folds, and all pseudo-labels are directly combined with the original labels. Then, a new model is trained on the dataset with combined labels.

SSR Positive Unlabeled Learning (SSR-PU) (Wang et al., 2022) This method applies a positive unlabeled learning algorithm for document-level RE under the incomplete annotation scenario. SSR-PU utilizes a shift-and-squared ranking (SSR) loss to accommodate the distribution shifts for the unlabeled examples.

Class Re-balancing Self-Training (CREST) (Wei et al., 2021) This algorithm re-samples the pseudo-labels generated by models based on the frequencies of the training samples.

Class-Adaptive Self-Training(CAST) (Tan et al., 2023) This method calculates the precision and recall scores of each class on the development set and uses the calculated scores to compute the sampling probability of each class to alleviate confirmation bias caused by erroneous pseudo labels.

B Additional experimental results

Method	Samples	PPL ↓	Acc ↑	DIST ↑
	Testing set	17.80	-	53.42
ATLOP-Dual	random	17.31	71.94	49.84
	positive	17.04	80.96	45.09
	negative	18.98	60.81	50.82
DuRE	random	17.03	79.54	50.24
	positive	16.93	82.28	49.92
	negative	18.32	67.43	52.88

Table B1: Performance of learned generator on Re-DocRED dataset.

B.1 Performance of Generation Model

To evaluate the quality of generated pseudo text, we measure generation fluency (perplexity, PPL), faithfulness (accuracy of generated text followed by given prompts, Acc), and diversity (number of distinct n-grams, DIST). Details of generation metrics are described in the appendix. We measured different types of samples: random, positive, and negative. The result is shown in Table B1. We found that our generator improves all metrics through dual self-training. Also, the generation fluency is comparable to the testing set. Besides, the diversity of DuRE also improves thanks to the use of more unlabeled text. However, the diversity of generation is still worse than the testing set due to the restricted model size.

B.2 Full Ablation Study Results

Full table of ablation study at Tab. B2.

C More Examples of Generation

We sample some generated texts given long-tail relations in Re-DocRED dataset. Table C1 reports the positive and negative examples generated by our DuRE model. We can find out that even if the relations are long-tail and might not be recognized by our RE classifier, positive examples still contain the relation triples, while some negative examples no longer contain the given relation triples. Learning the contrastive loss helps the generator learn more faithful examples, and thus improves the quality of pseudo training data for RE classifier. This can explain why generated pseudo text for long-tail relations can still be helpful in improving the F1 scores.

Model	**P**	**R**	**F1**	**Ign_F1**	**Freq_F1**	**LT_F1**
DuRE	80.01	73.84	76.84	75.32	79.81	72.88
DuRE −SAG	79.51	73.24	76.24	75.02	79.71	70.68
DuRE −CL	80.44	71.43	75.67	74.04	79.24	69.90
DuRE −PT	80.04	70.35	74.88	73.72	79.14	67.88
DuRE −PL	82.71	66.53	73.75	72.79	77.81	67.20
DuRE −PL−PT	85.17	61.93	71.72	70.95	76.07	64.70

Table B2: Full ablation study results on Re-DocRED dataset. Here − means removing components from DuRE. −SAG: remove the self-adaptive generation strategy and sample pseudo texts for different classes in the same setting. −CL: do not generate positive/negative examples and sample random examples. −PT/−PL: do not use pseudo text/labels.

Input	* Christianity*, seperated from, * Judaism*.
Positive	The * Holy Spirit* (* Synchrony of the Holy Spirit*) is the most effective means of establishing and strengthening the religious adherence to * Judaism*. As * Christianity* continues to be a branch of * Judaism*, it is crucial that the doctrines of the * Holy Spirit* and of * Judaism* remain with us. It is the aim of the * Holy Spirit* to establish and strengthen the character of the religious adherence of * Judaism* through systematic means of reforming the system of religious practices, the decentralization of religious ceremonies, and the re-establishment of unity among the * Judaic* and * Christian* communities. The most successful form of this is the systematic observance of the * Bible*, in which the primary source of the faith is the knowledge of * Allah* (" Father, Lord, High, Father, High)... In modern terms, the * Holy Spirit* focuses primarily on the advancement of religious adherence, which has been known as the most effective form of a religious system in * Judaism*. The * Holy Spirit* emphasizes the observance of the * Bible*, which has its roots in the * Bible*. The * Holy Spirit* is focused on the expansion of * Judaism*, which has a tendency to create a sense of unity among all * Judaism*. Its success in teaching and interpreting * Judaism* is demonstrated by its use of simple language and the presentation of the * Bible* as a unified unit.
Negative	* Christianity* and * Judaism* are two forms of the * Western Church* which originated around * 1500*. * Christianity* and * Judaism* had a strong influence on the * Eastern Church* and in * Europe* during the second half of the 19th century, with significant influence over * Germany* and * Poland*, which had a relatively neo - western influence. * Christianity* was characterized by a strong influence on the * Eastern Church* and in * European* countries. It is not clear when these influences actually began to manifest in the * Western Church*. However, * Judaism*'influence spread to other areas of the world such as * Africa*, * Asia*, and * the Americas*, and was influenced by this influence.
Input	* Chipilo*, sister city, * Segusino*
Positive	The * Segusino* (or * Segusino* in * Japanese*) was a city in the western part of * Asia*. It was a part of the * Western Asian region* that was unified by the * East Asian region* and * North American region*. It was originally the location of the * Pacific* coast resort town of * Chipilo*, and then the * Mid - East* area. It was a " hamlet ", also called a " sarket ", a kind of rubber cement used in transport. In * Japan*, the town of * Chipilo*, located north of * Tokyo*, was a hamlet that was the home of * Segusino*'s samurai army and the * Japanese* colonel's troops.
Negative	* Segusino* is the second largest city in the island of * Chipilo*. The city has a population of around * 1.27 million* as of the year * 2010*. Its urban population has declined in recent years due to urban development. It is home to the * Segusino Opera House*, a branch of the * Segusino Music Theatre*. The capital of the city is * Segusino*, and the city is surrounded by a number of historic sites, such as the * Segusino Cathedral* and the * Rialto Hall*. A major site dedicated to historic architecture is * Seiji*'s city hall. The first major buildings were built on * Segusino*'s main mountain, * Yatsuda*. The second large building was built on * Seiji*'s main mountain, * Yatsuda*. It was built by * Yatsuda* in * 1955*, but was demolished in * 1964*.
Input	* Russian*, seperated from, * Soviet Union*
Positive	The * Russian* military in the late 18th and early 19th centuries was characterized by systematic, largely self - destructive attacks on civilians, in which large numbers of soldiers and civilians were transported to other parts of the world, the " concentration camp ", or at least the most isolated concentration camp in the world. These, however, were mostly civilians who were not included in the camp's security measures. Most of them were forced to abandon their lives because of armed and destructive attacks, or because they did not comply with the orders of a military commander. The * Russian* military system consisted of a government known as the " * General Army of the Russian People* " (<unk>, * Military Provisional Army of the Soviet People*), which was created for the defense of civilians and was separated by the * Soviet Union* at the beginning of the nineteenth century. * Soviet* military operations began in the late * 19th century*, after the dissolution of the * Soviet Union*. The * General Army of the Soviet People* was, under the command of * Alexander Nevsky*, one of the most successful units in the * Soviet* army, in the early * 20th century*.
Negative	The * Russian* and * German* * Civil War* of * World War I*, fought between * 1940* and * 1941*, was a period of political uncertainty in the post - Communal era. Initially, the crisis erupted in * World War II* (and possibly the reunification of the * German* and * Polish* - * Soviet Union*) when the * German* and * Russian* military governments were faced with the imposition of a series of massive counter - Communal invasions and repressions. The conflict also marked the end of the * German* war, when * France* and * Germany* were both forced to declare war. In the aftermath of the war, * German* and * Russian* nationalists became the major force involved, often with the help of * German* troops. This increased tension, as the war progressed, the use of armies, and a series of armed rebellions.

Table C1: More examples of generation on Re-DocRED dataset on long-tail relations.