S^2ynRE: Two-stage Self-training with Synthetic data for Low-resource Relation Extraction

Anonymous ACL submission

Abstract

Current relation extraction methods suffer from the inadequacy of large-scale annotated data. While distant supervision alleviates the problem of data quantities, there still exists domain disparity in data qualities due to its reliance on domain-restrained knowledge bases. In this work, we propose S^2ynRE, a framework of two-stage Self-training with Synthetic data for Relation Extraction. We first leverage the capability of large language models to adapt to the target domain and automatically synthesize large quantities of coherent, realistic training data. We then propose an accompanied two-stage self-training algorithm that iteratively and alternately learns from synthetic and golden data together. We conduct comprehensive experiments and detailed ablations on popular relation extraction datasets to demonstrate the effectiveness of the proposed framework. Specifically under low resource settings, S^2ynRE brings up to 17.18% absolute improvements and 12.63% on average across all datasets.

1 Introduction

Relation extraction systems aim at discovering relational knowledge between entities by reading from unrestricted texts (Cardie, 1997). Although neural methods, especially pre-trained language models, have greatly advanced the state-of-the-art relation extraction capability (Zeng et al., 2014; Wu and He, 2019), they still require large quantities of training data (Han et al., 2020). However, annotated instances of high quality are usually time-consuming and labor-intensive to obtain in many real-world scenarios, thus leaving it a major challenge to build competent relation extractors with limited resources.

Distant supervision (Mintz et al., 2009), which automatically annotates relational statements by aligning entities with an existing knowledge bases (Bollacker et al., 2008; Vrandečić and Krötzsch, 2014), has been widely explored as an effective way to construct large scale relational dataset. To better exploit such available resources, several recent works propose to first pre-train a relational encoder on distant data using contrastive pretext tasks, then finetune it on downstream tasks (Baldini Soares et al., 2019; Peng et al., 2020; Qin et al., 2021). Although this line of methods have seen certain improvements, they still inevitably raise the concern that the distantly annotated data can vary considerably from downstream tasks both in target schema and in context distributions, thus may not be able to offer optimal transferability. For instance, due to the reliance on existing knowledge bases, current works mostly resort to Wikidata as the source of relational triples and Wikipedia (Vrandečić and Krötzsch, 2014) as the corpus for distant supervision. This circumscribes distant data to only factual knowledge between world entities, while downstream tasks may be of other special interests involving various domains, ranging from semantic relation between nominals (Hendrickx et al., 2009) to chemical-protein interactions (Kringelum et al., 2016).

Meanwhile, recent advances in large-scale pretrained language models (LLM) (Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020) have demonstrated their great potential in generating realistic texts of various domains, including news article, commodity reviews, dialogs, etc (Radford et al., 2019). Accordingly, several very recent works have explored the possibility to exploit LLM as an alternative training data pool (Schick and Schütze, 2021; Vu et al., 2021). However, these works are confined to natural language inference (NLI) task where the training data comply with the plain-text format and the label semantics can be clearly distinguished into fixed categories of neutral, entailment, and contradiction.

In this paper, we study the construction of synthetic data for relation extraction tasks to simultaneously address both training data scarcity in low resource scenarios and domain disparity in distant supervision. We employ LLM to estimate and adapt to the target domain distribution with only a few training instances, and synthesize a large amount of ones accordingly. Different from the NLI task, relational data exhibit specific structure involving an entity pair within the context. Besides, relation labels often entail more abstractive semantics, making it difficult to accurately synthesize label-specific instances. To resolve these two emerging challenges, we advocate two key designs: 1) we linearize relational statements into natural language sequences where entity pairs are indicated by special marker tokens; 2) we resort to unconditional generation instead of label-conditioned ones, which relaxes the requirements for strict label-semantic correspondence but increases sample availability and diversity. In general, it is observed that with only a few accessible samples, we are able to successfully synthesize a large amount of domain-customized training data with satisfactory quality.

To effectively learn from such synthetic data, we novelly advocate a two-stage self-training algorithm. The approach in general follows the self-training framework (Yarowsky, 1995; Xie et al., 2020), which is widely employed to exploit unlabeled data. Typically, such methods iteratively annotate and learn pseudo labels for unlabeled data to bootstrap the model's performance. Distinctively, we make a two-stage adaptation where in each of the iterations, the model is firstly trained on synthetic instances, then on golden ones. Such sequential training procedure favors golden data with more importance since they are introduced in the latter stage of the training curriculum. Besides, we formulate synthetic data training as a knowledge distillation process using soft labels instead of assigning them with hard labels. In general, we show that the proposed two-stage self-training algorithm contributes significant improvements by taking the benefit of synthetic data while mitigating its noise impact.

We refer to our method as S^{2}ynRE, a framework of two-stage **S**elf-training with **Syn**thetic data for **R**elation **E**xtraction. To demonstrate its effectiveness, we conduct comprehensive experiments on popular relation extraction datasets, including SemEval 2010 Task 8 (Hendrickx et al., 2009), TACRED (Zhang et al., 2017) and two of its rectified versions (Alt et al., 2020; Stoica et al., 2021), as well as ChemProt (Kringelum et al., 2016) in biomedical domain. We show that S^{2}ynRE brings consistent improvements over its baseline, and outperforms existing works that learn from distant data. Besides, in low resource settings, S^{2}ynRE brings much more significant advantages (up to 17.18% absolute improvements, and 12.59% on average across all datasets) benefiting from the availability of large quantities of domain-customized synthetic samples. The contributions of this paper can be summarized as follows:

- We propose to leverage LLM to synthesize large quantities of domain-customized relational instances for relation extraction, which novelly mitigates the problems of both data scarcity and domain disparity, and also outperforms the prevailing distant supervision. We formulate it into an unconditional generation of marked natural language sequence to accomplish a successful synthesis.

- We propose a novel two-stage self-training algorithm to effectively learn from unlabeled synthetic data and golden data together. We demonstrate that this is a non-trivial adaptation that significantly outperforms standard self-training widely employed in semi-supervised learning.

- We provide solid experimental results of S^{2}ynRE on several established relation extraction benchmarks, showing its advantage along with detailed ablations that demonstrate the effectiveness of the entire framework as well as the advantages of each specific component.

2 Related Works

Relation Extraction Relation extraction is one of the fundamental tasks in natural language processing (Cardie, 1997), where lots of research efforts have been made to advance the state-of-the-art methods (Zeng et al., 2014; Zhou et al., 2016; Zhang et al., 2018; Baldini Soares et al., 2019), as well as the low-resource scenario (Han et al., 2018; Sainz et al., 2021; Dong et al., 2021; Chen et al., 2022). One of the most prominent methods is distant supervision (Mintz et al., 2009), which automatically constructs annotated relational data by aligning corpus with existing knowledge base. Many recent works investigate how to learn effectively with such distant data (Baldini Soares et al., 2019; Peng et al., 2020; Ding et al., 2021; Qin et al., 2021). Generally, they propose various pretext tasks that pre-train a model to learn relational representation. We will further explain some of these works for comparison in Section 5.2.

Learning from Synthetic Data Built upon massive corpora, pre-trained language models are promising at producing texts of eligible quality, resulting in a surge of research interests in its usage for data augmentation (Feng et al., 2021). One straightforward way is to introduce mask corruptions in the way language models are pretrained, then collect predictions as augmented data (Kobayashi, 2018; Ng et al., 2020). Later works further developed such technique into conditional augmentation (Wu et al., 2019; Kumar et al., 2020). Nevertheless, these methods are mostly editing existing instances, which limits the diversity and scale of augmented data.

With increasingly powerful LLMs, recent works turn to direct synthesis of new instances (Schick and Schütze, 2021; Wang et al., 2021; Meng et al., 2022; Ye et al., 2022). Different from this work, most of them focus on zero-shot language understanding where no labeled data is available (Schick and Schütze, 2021; Wang et al., 2021; Meng et al., 2022; Ye et al., 2022). They investigate ways to generate label-conditioned data by prompting LLMs, but these methods can hardly be applied to low-resource or full data scenarios while still preserving effectiveness.

With the existence of labeled data, synthetic data needs to be of higher quality to bring further utility. Several works thus propose to finetune the generator (Anaby-Tavor et al., 2020; Vu et al., 2021; He et al., 2021). A major challenge of learning from these synthetic and golden data together is how to further alleviate the noise, existing attempts include threshold-based confidence filtering (Anaby-Tavor et al., 2020), classical semi-supervised learning (He et al., 2021) or restricting the usage of synthetic data within a supplemental intermediate task (Vu et al., 2021).

For structured learning tasks, Ding et al. (2020) similarly formulates NER task data as sequential language. Specifically for relational data synthesis, Papanikolaou and Pierleoni (2020) explore the biomedical domain and Chia et al. (2022) focus on zero-shot setting of triplet extraction. By contrast, Syn^{2}RE distinguishes not only in applied scenario and synthesis strategy, but also in the two-stage learning framework, which is specially designed for improved synthetic data adaptation.

3 Preliminary

This section formulates the task of relation extraction and the baseline models used throughout all experiments.

Task Formulation A typical relation extraction task is defined by a corpus of relational statements and a set of relations, i.e., schema S. Assume the training dataset $\boldsymbol{\mathcal{D}}^{tr} = \{(\mathbf{x}_i, s_i, o_i)\}_{i=1}^N$ and its corresponding labels $\boldsymbol{\mathcal{Y}}^{tr} = \{y_i\}_{i=1}^N$, where $\mathbf{x}_i$ is a sequence of words $\{w_l^i\}_{l=1}^L$, $y_i \in S$, $s_i = [w_{s_{start}} : w_{s_{end}}]$ and $o_i = [w_{o_{start}} : w_{o_{end}}]$ are subject and object entities within the context. The target is to learn a function $f_{\boldsymbol{\theta}}(\mathbf{x}_i, s_i, o_i)$ that predicts the correct relation label y_i.

Baseline Model As S^{2}ynRE is a data-centric framework, we keep the model architecture simple but competitive, which is the vanilla finetuning of pre-trained language models. Instead of autoregressive LMs, we use auto-encoding networks like BERT as they usually perform better on language understanding downstream tasks. Following Baldini Soares et al.'s (2019) comprehensive study of building relation extractors, we inject special marker tokens to the input word sequence:

$$\begin{aligned}\mathbf{x}_{marked} = (\ldots, \texttt{[Sub]}, s, \texttt{[\textbackslash Sub]}, \ldots \\ \ldots, \texttt{[Obj]}, o, \texttt{[\textbackslash Obj]}, \ldots)\end{aligned} \tag{1}$$

After the encoding process of transformer, the representation $\mathbf{h}$ in corresponding positions will be concatenated for classification:

$$\widehat{\mathbf{y}} = \texttt{softmax}(\mathbf{W}^{|S|}[\mathbf{h}_{\texttt{[Sub]}}; \mathbf{h}_{\texttt{[Obj]}}]) \tag{2}$$

where $W^{|S|}$ is a feedforward network and the predicted categorical distribution $\widehat{\mathbf{y}}$ will be trained against y using cross-entropy loss.

4 Methodology

We elaborate on the framework of S^{2}ynRE (see Fig. 1) in this section, including the construction of an LLM-based synthesizer, and the two-stage self-training algorithm.

4.1 Relational Data Synthesis

Training instances of relation extraction task is of specific structure $(\mathbf{x}_i, s_i, o_i)$, i.e., the relational statement is expected to be a sentence containing exact two entities as subject and object. Inspired

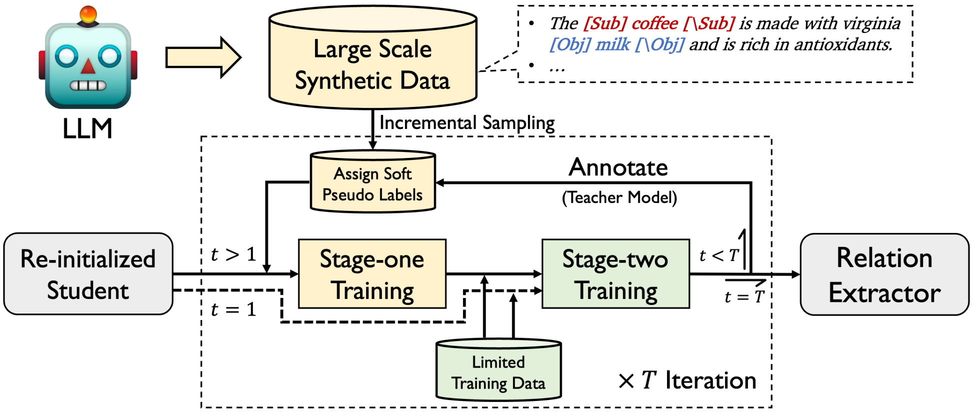

Figure 1: The overall framework of S^{2}ynRE. We iteratively train the student model on both synthetic and golden data via a two-stage self-training strategy. Note that in iteration $t = 1$, stage-two is directly applied. The exemplary instance is sampled from our synthetic data for SemEval.

by Paolini et al. (2021), we linearize relational data into marked natural language sequence as in Eq 1. The synthesizer can be built upon any existing LLMs, e.g., GPT-2. We first finetune it for a few steps in the same autoregressive way as how it is pre-trained:

$$\mathcal{L} = \sum_{l=1}^{L+4} logP(w_l|w_0, ..., w_{l-1}; LLM) \qquad (3)$$

where $\{w_l\} = \mathbf{x}_{marked}$, and a <bos> token is prepended as w_0. Note that we ignore relation labels y in training data and approach it as unconditional generation. This eliminates the noise caused by label-semantic inconsistency, and leaves it to model itself to learn from unlabeled synthetic data.

After the finetuning is completed, we simply prepend the <bos> token to prompt the generation, and repeatedly perform inference using multinomial sampling until we obtain the expected scale of synthetic data $\boldsymbol{\mathcal{D}}^{syn}$. We show in appendix D that these synthetic data are coherent, realistic, and most importantly, customized to the target domain.

4.2 Two Stage Self-training

Self-training is a widely adopted learning algorithm for semi-supervised learning. Typically, to jointly learn from an unlabeled dataset and a labeled dataset, it iteratively samples from the unlabeled set, assigns them with pseudo labels, merges them with the labeled dataset, and re-trains the model. In this paper, we argue that this design of naive merging is built upon a strong assumption that the unlabeled dataset must be in the exact distribution with the labeled ones, for which the synthetic data does not strictly satisfy.

In S^{2}ynRE, differently, we make a two-stage adaptation: where synthetic data and golden data are trained sequentially. We start from a base model initialized using any auto-encoding language models, e.g., BERT (Devlin et al., 2019), and train it on $\boldsymbol{\mathcal{D}}^{tr}$ to produce a teacher model $\boldsymbol{\eta}$, as introduced in Section 3. We first use $\boldsymbol{\eta}$ to annotate the unlabeled synthetic data $\boldsymbol{\mathcal{D}}^{syn}$:

$$\hat{\mathbf{y}}_i^{syn} = \boldsymbol{\eta}(\mathbf{x}_i^{syn}, s_i, o_i) \qquad (4)$$

and we keep $\widehat{\boldsymbol{\mathcal{Y}}}^{syn} = \{\hat{\mathbf{y}}_i^{syn}\}$ as soft pseudo labels of $\boldsymbol{\mathcal{D}}^{syn}$, note that here the ˆ denotes *soft* as we keep the categorical distribution intact instead of keeping its argmax. Inspired by Li and Qian (2021), to further eliminate fluctuations in pseudo labels, we train multiple teachers using different random seeds, and the pseudo labels annotated by k-th teacher is referred to as $\widehat{\boldsymbol{\mathcal{Y}}}_k^{syn}$.

We then re-initialize a new student model $\boldsymbol{\theta}$, and apply a two-stage training strategy. In stage-one training, student $\boldsymbol{\theta}$ is trained on synthetic data using soft pseudo labels:

$$\boldsymbol{\theta}' \leftarrow \mathcal{L}_{KD}(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}^{syn}, \{\widehat{\boldsymbol{\mathcal{Y}}}_k^{syn}\}_{k=1}^{K}) \qquad (5)$$

This can be seen as a distillation procedure that transfers knowledge from $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$ based on synthetic data $\boldsymbol{\mathcal{D}}^{Syn}$. And $\mathcal{L}_{KD}$ is calculated as:

$$\mathcal{L}_{KD} = \frac{1}{K} \sum_{k=1}^{K} D_{KL}(\hat{\mathbf{y}}_i^{syn} \parallel \boldsymbol{\theta}(\mathbf{x}_i^{syn}, s_i, o_i)) \qquad (6)$$

where D_{KL} is the Kullback-Leibler divergence. Then in stage-two training, we take from $\boldsymbol{\theta}'$, and train it on labeled training dataset:

$$\boldsymbol{\theta}'' \leftarrow \mathcal{L}_{CE}(\boldsymbol{\theta}', \boldsymbol{\mathcal{D}}^{tr}, \boldsymbol{\mathcal{Y}}^{tr}) \qquad (7)$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss, and $\boldsymbol{\theta}''$ is the resulting model in this iteration. We then use $\boldsymbol{\theta}''$ as the teacher model $\boldsymbol{\eta}$ for the next iteration to re-annotate $\boldsymbol{\mathcal{D}}^{syn}$, and this procedure is repeated T times. Following the standard practice of self-training, in each iteration, we incrementally sample $1/T$ more synthetic data from $\boldsymbol{\mathcal{D}}^{syn}$ until in iteration T, where $\boldsymbol{\mathcal{D}}^{syn}$ will be running out of new instances. The entire two-stage self-training process can be formulated as Algorithm 1.

Algorithm 1: Two-stage Self-training.

```
Input: Golden training dataset $\boldsymbol{\mathcal{D}}^{tr}$, $\boldsymbol{\mathcal{Y}}^{tr}$,
       synthetic dataset $\boldsymbol{\mathcal{D}}^{syn}$
/* ===== Iteration 1 =====                    */
$t = 1$;
$\boldsymbol{\mathcal{D}}_1^{syn} = \varnothing$;
Initialize $\boldsymbol{\theta}$ from auto-encoding LM;
$\boldsymbol{\theta}_1 \leftarrow Train(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}^{tr}, \boldsymbol{\mathcal{Y}}^{tr})$ ;        // Eq.7
$\boldsymbol{\theta}_1^{Tea} \leftarrow \boldsymbol{\theta}_1$;        // assign teacher model
/* ===== Iteration 2∼T =====                  */
repeat
    $t = t + 1$;
    $\boldsymbol{\mathcal{D}}_t^{syn} = \boldsymbol{\mathcal{D}}_{t-1}^{syn} \cup \boldsymbol{\mathcal{D}}^{syn}[\frac{t-1}{T} : \frac{t}{T}]$;
    $\widehat{\boldsymbol{\mathcal{Y}}}_t^{syn} \leftarrow Annotate(\boldsymbol{\theta}_{t-1}^{Tea}, \boldsymbol{\mathcal{D}}_t^{syn})$;
     // Eq.4
    Re-initialize $\boldsymbol{\theta}$ from auto-encoding LM;
    /* stage-one training                      */
    $\boldsymbol{\theta}_t' \leftarrow Train(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}_t^{syn}, \widehat{\boldsymbol{\mathcal{Y}}}_t^{syn})$;   // Eq.5
    /* stage-two training                      */
    $\boldsymbol{\theta}_t'' \leftarrow Train(\boldsymbol{\theta}_t', \boldsymbol{\mathcal{D}}^{tr}, \boldsymbol{\mathcal{Y}}^{tr})$ ;   // Eq.7
    $\boldsymbol{\theta}_t^{Tea} \leftarrow \boldsymbol{\theta}_t''$;    // update teacher model
until performance converges or t reaches
  maximum iteration limit $T$;
Output: Final model $\boldsymbol{\theta}_t''$
```

5 Experiments

5.1 Experimental Settings

We evaluate S^{2}ynRE on popular datasets including **SemEval 2010 Task 8** (Hendrickx et al., 2009), **TACRED** (Zhang et al., 2017), **TACRED-Revisited** (Alt et al., 2020), **Re-TACRED** (Stoica et al., 2021), and **ChemProt** (Kringelum et al., 2016). Their statistics are given in Table 2 and we refer to detailed introduction in Appendix A.

For each dataset, we set three different prerequisites of resource availability. Respectively, *FULL* for 100% training data, *LIMITED* for 10% training data and *FEW* for 1% training data. To provide robust and convincing conclusions, we run all experiments (including ablation studies) with 5 different random seeds and report their average. With each random seed, we employ grid search to select the best model as well as the teacher model in each iteration. We use only development set for such selection, and report the corresponding test set score as the final results.

For data synthesis, we use GPT-Large as the aforementioned LLM. Specifically, for ChemProt, we use an adapted version of GPT-2 (Papanikolaou and Pierleoni, 2020), which is further trained on 500k PubMed abstracts. When generating, we restrict sequence length to 128, and perform necessary filtering by removing instances that do not conform with the relational structure, i.e., there must exist 4 exact special markers and each start position marker shall appear before its end position marker. The synthesis efficiency is 24.05 instances per second before any filtering. In total, we collect 10,000 samples for *FEW* setting, and 100,000 synthetic samples for *LIMITED* and *FULL* settings.

We use *bert-base-uncased* to initialize the student model. All experiments are conducted on 40GB A100 machines. We leave other hyper-parameters to Appendix B.

5.2 Main Results

We choose competitive baselines and reproduce them under comparable settings to provide more reliable conclusions. These baseline methods are:

BERT We finetune BERT model (Devlin et al., 2019) in a straightforward way for relation extraction as explained in Section 3 and implemented in many existing works. This serves as our re-implemented *Finetune Baseline* and will be referred to in the following figures.

MTB (Baldini Soares et al., 2019) pre-trains a relational encoder using matching the blanks task, which is built on the hypothesis that two relational statements containing the same entity pair should express similar relational representations. Note that this is a weaker reliance than distant supervision as it only aligns entities, and does not need relations.

CP (Peng et al., 2020) proposes a contrastive learning pretext task that encourages sentence representations with the same relation to be similar and different ones to be disparate.

ERICA (Qin et al., 2021) further extends distant supervision to document-level corpus, and design similar pretext task that discriminates relational representations across sentences.

Method	SemEval	TACRED	TACRED-Revisited	Re-TACRED	ChemProt
			FULL (100% training data)		
BERT	88.86±0.30	69.27±0.27	79.24±0.37	87.75±0.22	81.66±0.79
MTB	88.95±0.31	69.93±0.40	79.69±0.32	87.67±0.37	81.75±0.86
CP	89.16±0.17	70.16±0.20	**80.08**±0.32	87.95±0.09	81.77±0.97
ERICA	88.62±0.24	68.91±0.75	78.95±0.86	87.73±0.31	81.52±0.43
S^{2}ynRE$_{\text{BERT}}$	**89.20**±0.27	**70.25**±0.47	79.80±0.29	**88.01**±0.24	81.65±0.60
S^{2}ynRE$_{\text{CP}}$	89.04±0.32	70.03±0.27	79.75±0.49	87.98±0.07	**82.15**±0.12
			LIMITED (10% training data)		
BERT	82.38±0.51	59.32±0.35	66.56±0.48	80.51±0.77	68.96±0.97
MTB	82.56±0.27	59.45±0.55	66.48±0.71	81.15±0.59	71.44±1.12
CP	83.80±0.50	62.81±0.39	**70.81**±0.58	83.42±0.41	71.89±1.09
ERICA	82.41±0.55	58.54±0.65	66.65±0.68	80.45±0.77	69.03±1.22
S^{2}ynRE$_{\text{BERT}}$	84.01±0.23	61.26±0.53	68.62±0.15	83.28±0.40	73.62±0.14
S^{2}ynRE$_{\text{CP}}$	**84.64**±0.30	**62.94**±0.45	70.36±0.75	**84.36**±0.32	**75.32**±0.92
			FEW (1% training data)		
BERT	40.81±1.62	30.40±7.74	33.75±8.68	54.75±4.52	39.50±1.47
MTB	45.12±1.23	36.52±2.00	40.69±2.25	58.35±0.93	41.53±2.11
CP	53.29±1.80	49.81±0.59	55.53±0.90	68.03±0.76	43.96±2.62
ERICA	43.62±2.33	34.91±1.40	39.17±1.69	57.14±0.83	40.01±0.86
S^{2}ynRE$_{\text{BERT}}$	57.99±1.08	45.87±1.07	50.61±0.99	62.82±0.52	45.09±0.38
S^{2}ynRE$_{\text{CP}}$	**68.03**±0.46	**51.91**±0.68	**58.48**±0.29	**70.21**±0.81	**46.23**±0.73

Table 1: Main results. Best performances are **bold**, and the second bests are underlined. We report Accuracy for Chemprot, and Micro-F1 for other datasets. Results for all baseline methods are reproduced with identical hyper-parameter searches for fair comparison[1].

Dataset	Train	Dev	Test	**1% Train**	Relation
Semeval	6507	1493	2717	73	19
TACRED	68124	22631	15509	703	42
TACRED-Revisited	68124	22631	15509	703	42
Re-TACRED	58465	19584	13418	570	40
ChemProt	4169	2427	3469	49	13
Wiki80	39200	5600	11200	400	80

Table 2: Numbers of instances in train, dev, test splits and low resource settings.

Dataset	**Resource Usage**	**Domain**	**External Requirements**	
			KB Entities	**KB Relations**
MTB	6,000,000 sent pairs	Wiki	✓	No Requirements
CP	867, 278 sents	Wiki	✓	✓
ERICA	1,000,000 docs	Wiki	✓	✓
S^{2}ynRE	100,000 sents	Customized	No Requirements	

Table 3: Comparison of external resource usage and requirements for different methods.

We provide an overview of these works regarding various resource usage and requirements in Table 3. The main results are shown in Table 1. Under all three settings across five datasets, S^{2}ynRE outperforms the BERT finetune baseline. Specifically for the *FEW* setting, improvements are much more significant, respectively **+17.18**, **+15.47**, **+16.86**, **+8.07** and **+5.59**. We further employ CP as a stronger base model to initialize the students, and the performances are even better. This implies that the improvements of S^{2}ynRE are mostly orthogonal with those of the distantly pre-trained methods. In general, S^{2}ynRE$_{\text{CP}}$ achieves a new state-of-the-art for low resource relation extraction tasks.

5.3 Ablation Study

We investigate the advantages of S^{2}ynRE via comprehensive ablations. In accordance with the main claim, all experiments are conducted under the low-resource (*FEW*) setting unless otherwise stated.

Synthetic Data Instead of Distant Data Distant supervision has long been the prevailing solution to automatically construct relational data. We make its comparison against the proposed synthetic data in table 4. We keep the two-stage self-training algorithm intact, only replace the synthetic data with distant data[2]. On 5 investigated datasets, distant data can provide appreciable improvements ranging from **+2.06** to **+13.25**, however, synthetic data brings much more significant improvements ranging from **+5.59** to **+17.18**, which clearly demonstrates the superiority of being

[1] We obtain MTB and CP checkpoints from `https://github.com/thunlp/RE-Context-or-Names` and ERICA checkpoint from `https://github.com/thunlp/ERICA`

[2] The distant data is produced and released by Peng et al. (2020), we randomly sample 100,000 instances out of it

Dataset	NA	Distant	Synthetic
SemEval	40.81	49.36 (+ 8.55)	**57.99** (+17.18)
ChemProt	39.50	41.56 (+ 2.06)	**45.09** (+ 5.59)
TACRED	30.40	42.43 (+12.03)	**45.87** (+15.47)
Re-TACRED	54.75	62.34 (+ 7.59)	**62.98** (+ 8.23)
TACRED-Revisited	33.75	47.00 (+13.25)	**50.61** (+16.86)
Wiki80	63.08	**66.77** (+ 3.69)	65.52 (+ 2.44)

Table 4: Comparison between synthetic data and distant data. Inside the parentheses are absolute improvements, red means the higher one.

domain-customized for target tasks. We further include another dataset Wiki80 (Han et al., 2019), which very closely follows identical distribution of distant data as both are constructed using distant supervision on wikipedia and wikidata. Result shows that synthetic data provides competitive improvements but no longer outperforms distant ones. This verifies the importance and advantage of domain-customized data from an opposite perspective. Nevertheless, real-world scenarios mostly involve distribution beyond the scope of wikipedia, and only the proposed synthetic approach can offer such advantage.

Two Stage Self-training Typical self-training algorithms merge the pseudo-labeled data into existing labeled data in each iteration, and minimize the model's empirical loss on a mixture of both. We refer to such classical implementation as mixed self-training as opposed to the proposed two-stage self-training. Fig. 2 compares these two approaches. The transparent blue in the background denotes iterations. In each iteration, there will be one evaluation for mixed self-training (**blue** curve), but two evaluation for Two-stage Self-training (**teal** for stage one, **Red** for stage two). We observe that in stage-one training, the performance might drop a few compare to its previous iteration, however, it effectively provides a better initialization where the model can further learn from the golden data. Overall, the model can continually bootstrap its performance by learning from synthetic and golden data iteratively and alternately. While in mixed self-training, the golden data are treated equally as synthetic ones, and the model is overwhelmed by large amounts of the latter. Therefore, the improvement quickly saturates to a limited plateau. We also provide illustrations of the bootstrapping performance over iterations on other datasets in Appendix C.

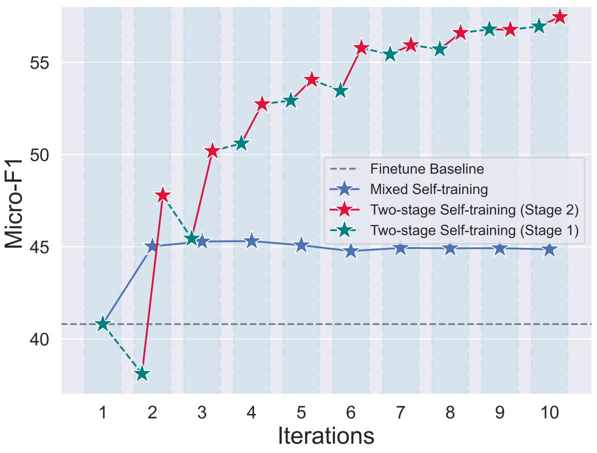

Figure 2: Performance illustration for two-stage self-training compared to classical mixed self-training. Analyzed on SemEval.

Method	SemEval	TACRED
MetaSRE	80.09$_{\pm 0.78}$	56.95$_{\pm 0.34}$
GradLRE	81.69$_{\pm 0.57}$	58.20$_{\pm 0.33}$
S^{2}ynRE w/ Golden	**84.11**$_{\pm 0.27}$	59.07$_{\pm 0.54}$
S^{2}ynRE w/ Synthetic	84.01$_{\pm 0.23}$	**61.26**$_{\pm 0.53}$

Table 5: Comparison to state-of-the-art methods for semi-supervised setting, including (Hu et al., 2021a) and GradLRE (Hu et al., 2021b). w/ Golden means unlabeled set are sampled from 50% of the golden training data and their original labels are removed accordingly.

Comparison Under Semi-supervised Setting Standard semi-supervised setting also investigates low-resource relation extraction by joint learning from both labeled data and unlabled data. However, they make a strong assumption of identical distribution between unlabeled data and labeled ones, and most existing works actually directly sample from the golden training data and remove the labels to construct the unlabled set. We provide comparison with state-of-the-art methods of semi-supervised learning in Table 5 (under the *LIMITED* setting). Results show that 1) the proposed two-stage self-training outperforms other semi-supervised learning algorithms, and 2) synthetic data demonstrates better or comparable performance compared to unlabled set constructed from golden training data. We attribute the latter to its domain-customized quality and unlimited large-scale quantity.

Unconditional Generation Although a lot of previous works intuitively resort to conditional synthesis, we show that this is not the optimal choice for relation extraction task. We finetune the synthesizer by prepending label-specific prompts: *"write*

Dataset	NA	Conditional Syn	Unconditional Syn
SemEval	40.81	45.26 (+4.45)	57.99
TACRED	30.40	33.34 (+2.94)	45.87
Re-TACRED	54.75	53.03 (-1.72)	62.98
TACRED-Revisited	33.75	37.60 (+3.85)	50.61

Table 6: Comparison between conditional and unconditional synthesis. Inside the parentheses denote the effectiveness comparing to Finetune Baseline.

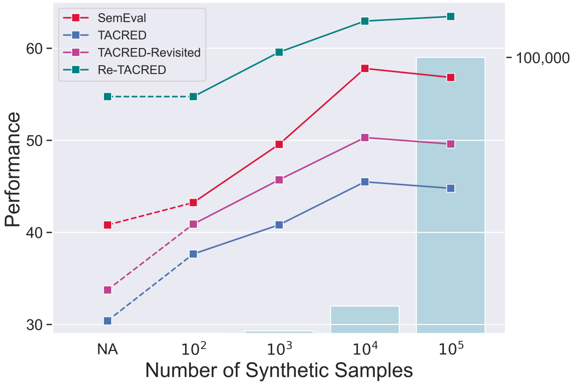

Figure 3: Performances w.r.t. different scales of synthetic data usage.

a sentence describing relation $V(r)$: ", where $V(r)$ is the verbalizer for each relation r and we directly use corresponding label strings, e.g., *Component-Whole(e2,e1)*. We synthesize each relation class proportional to its original distribution in golden dataset. As conditional generation provides already labeled data, we can directly finetune the student model instead of self-training. We still train synthetic and golden data sequentially as we empirically found it a better choice. The results show that conditional generation only brings minimum or no benefits. We attribute this to the difficulty of preserving required label semantics for highly abstractive tasks like relation extraction. As a consequence, while these extra amounts of data can still provide certain usability, they also most likely cause considerable distractions.

Scale of Synthesizer Model We test S^{2}ynRE with a different scale LLM, i.e., GPT-Small with 117M parameters. The results in Table 7 show that even with such a small size LM, S^{2}ynRE can still bring significant improvements. But in general, larger model unsurprisingly performs better. With the emergence and applicability of increasingly stronger LLMs, we can look forward to further advancement of relation extraction task.

Dataset	NA	GPT-2 Small 117M	GPT-2 Large 774M
SemEval	40.81	49.87	**57.99**
TACRED	30.40	43.95	**45.87**
TACRED-Revisited	33.75	48.35	**50.61**
Re-TACRED	54.75	**63.51**	62.98

Table 7: Performances w.r.t. synthesizer model size.

Scale	Golden	Synthetic
100	98.9	97.8 (- 1.1)
1,000	96.8	88.8 (- 8.0)
10,000	88.6	74.3 (-14.3)

Table 8: Sample diversity (type-token ratio in percentage for 3-grams) of synthetic and golden data w.r.t. different data scales on SemEval.

Scale of Synthetic Samples Figure 3 investigates the scale of synthetic samples. The improvements are approximately increasing in log scale w.r.t. the number of synthetic samples. The best performance is reached at 10,000, after which if we keep adding more samples, the performance saturates. As the synthesis of data is a repeatedly sampling process, we think exploiting too much data will deteriorate the diversity at the same time. We verify this by evaluating its diversity using type-token ratio (Roemmele et al., 2017; Kumar et al., 2020), which is defined as the ratio of unique n-grams out of all n-grams (see Table 8). We can see that the diversity gap between synthetic and golden data is enlarged when increasing the data scale.

6 Conclusion

In this paper, we present S^{2}ynRE, a framework of two-stage self-training with synthetic data for relation extraction. We show that synthetic data generated using LLMs can resolve data scarcity in low-resource scenarios and mitigate domain disparity compared to distant supervision. To enable effective learning from such synthetic data, we then propose a novel two-stage self-training algorithm that continually bootstraps model performance by iteratively and alternately training the synthetic and golden data together. The proposed framework brings substantial improvements and achieves the new state-of-the-art for relation extraction under low-resource scenarios. In the future, we expect new possibilities brought by LLMs and will further explore accompanied techniques to exploit their potential.

Ethical Considerations

Synthetic data generated by language models may involve potential ethical risks regarding fairness and bias (Bommasani et al., 2021; Blodgett et al., 2020), which results in further consideration when they are employed in downstream NLP tasks. Although the scope of this paper remains how to produce and leverage such synthetic data to build an improved relation extraction system, it is worth further investigation to manage the proposed framework in conjunction with well-established methods that can measure (Nadeem et al., 2021) and mitigate (Nadeem et al., 2021; Gupta et al., 2022) such ethical risks.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Claire Cardie. 1997. Empirical methods in information extraction. *AI Magazine*, 18(4):65.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2778–2788, New York, NY, USA. Association for Computing Machinery.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*.

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.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Hai-Tao Zheng, and Rui Zhang. 2021. Prototypical representation learning for relation extraction. In *International Conference on Learning Representations*.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. MapRE: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2694–2704, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, China. Association for Computational Linguistics.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2021. Generate, annotate, and learn: Nlp with synthetic text.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado. Association for Computational Linguistics.

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. 2021a. 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.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021b. Gradient imitation reinforcement learning for low resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2746, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Wanli Li and Tieyun Qian. 2021. From consensus to disagreement: Multi-teacher distillation for semi-supervised relation extraction. *arXiv preprint arXiv:2112.01048*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *CoRR*, abs/2202.04538.

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.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

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

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, Online. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Melissa Roemmele, Andrew S Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *Workshop on Machine Learning for Creativity, at the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining*.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabas Poczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13843–13850.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

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.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *CoRR*, abs/2109.09193.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2361–2364, New York, NY, USA. Association for Computing Machinery.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

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 (EMNLP 2017)*, pages 35–45.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

A Datasets

SemEval 2010 Task 8 (Hendrickx et al., 2009) is a widely used testbed for relation extraction, the schema targets at semantic relations between pairs of nominals, which requires certain level of abstractive capabilities. **TACRED** (Zhang et al., 2017) is a large-scale dataset annotated using Amazon Mechanical Turk crowdsourcing. It was initially created for the TAC knowledge base population and mainly covers common relations between people, organizations, and locations based on the TAC KBP scheme. **TACRED-Revisited** (Alt et al., 2020) is a label-corrected version of the TACRED dataset, which motivates from the unresolved challenging cases in original TACRED dataset. **Re-TACRED** (Stoica et al., 2021) further conducted a more comprehensive analysis and re-annotated the entire dataset. Besides, it made alternations to the schema to make it more clear and intuitive, which greatly improved the dataset quality. **ChemProt** (Kringelum et al., 2016) is a bio-domain dataset that extracts 13 kinds of chemical-protein interactions. It is widely used for evaluating domain-specific model capabilities (Lee et al., 2019; Beltagy et al., 2019).

B Experimental Settings

S^{2}ynRE involves three different training processes, respectively the finetuning of LLM, stage-one training, and stage-two training. Except for training steps or epochs, we do not exhaust further search for other hyper-parameters and set them empirically.

For the finetuning of LLM as synthesizer, we set batch size to 64, learning rate to 3e-5. We found that the quality of generated samples is sensitive to the finetuning steps. Considering that the scale of training samples varies from 73 (SemEval 1%) to 68,124 (TACRED 100%) w.r.t. different datasets and different settings, we search steps within different ranges accordingly. The final choices are listed in Table 9.

For stage-one training, we set batch size to 64, learning rate to 3e-5, and fix the training steps as 1500. We save the checkpoint from 500, 1000, and 1500 steps respectively and select the best one. For stage-two training, we set batch size to 16, learning rate to 3e-5, and the epochs are set as Table 10. These epoch settings are empirically chosen in our pilot study to obtain a competitive baseline performance.

C Performance Over Self-training Iterations

We provide the performance curve w.r.t. iterations in Figure 4. It shows that the iterative training procedure following the classical self-training method

Setting	SemEval	ChemProt	TACRED	TACRED-Revisited	Re-TACRED
Full	256	256	512	1024	2048
Limited	64	256	256	256	512
Few	32	32	128	128	128

Table 9: Finetuning steps for LLM under different settings.

Setting	SemEval	ChemProt	Wiki80	TACRED	TACRED-Revisited	Re-TACRED
Full	{5, 10}	{5, 10}	{5, 10}	2	2	2
Limited	{10, 20}	{10, 20}	{10, 20}	5	5	5
Few	{40, 80}	{40, 80}	{40, 80}	10	10	10

Table 10: Training epochs for stage-two training under different settings.

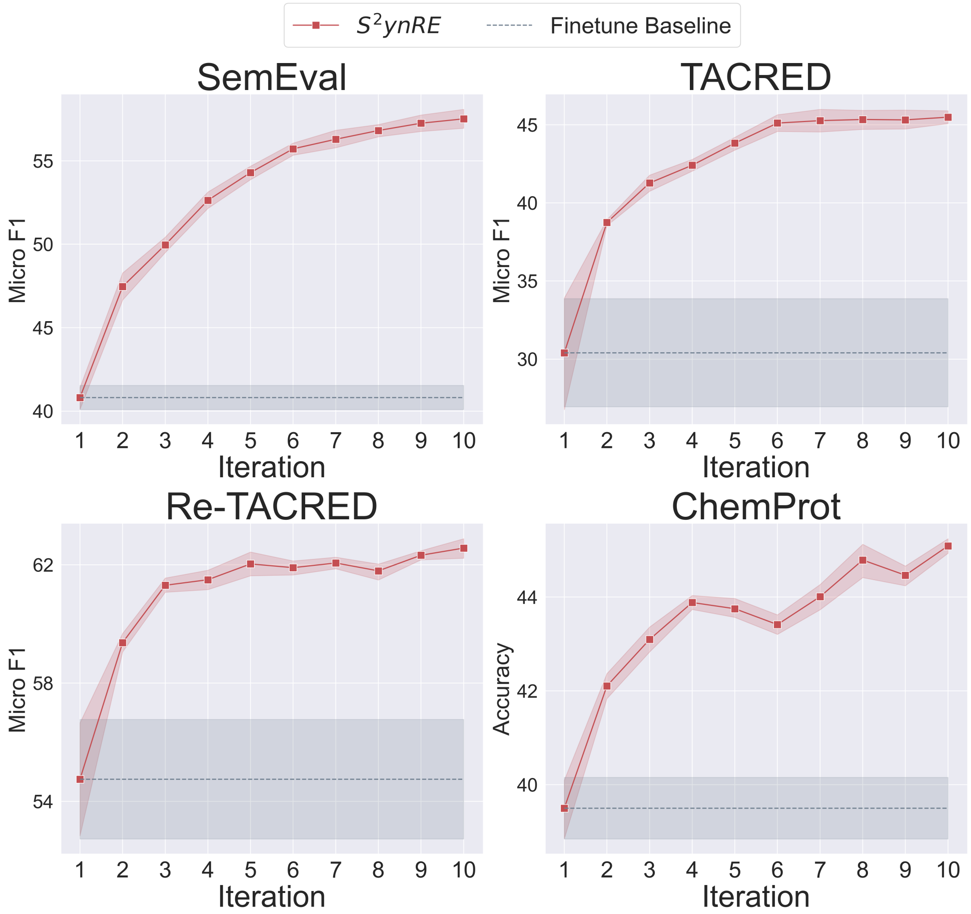

Figure 4: Performance over self-training iterations. Drawn with standard error of mean.

is indeed effective. We simply set iteration to 10 as most of the self-training methods did and find it already a robust choice across different datasets.

D Case Study

We provide randomly sampled case studies of synthetic data for SemEval, TACRED, and ChemProt in Table 11, 12, and 13 respectively. These cases show that LLMs are capable of synthesizing coherent, realistic sentences with relational structure. Most importantly, such synthetic data are customized to target domains with various topics and styles.

Nevertheless, we also notice several limitations, especially in low-resource scenarios where it's still challenging to get a good estimation of the target dataset distribution:

- Lack of diversity. For example, instances 2.1, 2.2, 2.3 all start with *"the marmalade"*.

- Fragmentary structure. For example, instances 2.4 and 2.8 contain atypically lengthy object.

For pseudo labels, most of the time teacher model confidently assigns one specific label with very high probabilities (> 0.95), but for some other cases, it goes for more than one possible label, such as 1.8, 2.8, 4.1, etc. We attribute this to two possible reasons: 1) the limited capability of the teacher model to accurately recognize all relations, and 2) the imperfections of certain synthetic data, i.e., some synthetic instances do not well align with pre-defined schema and are difficult to be assigned exact relation labels. In these cases, forcing the student to learn from hard labels assigned using argmax might introduce severe noise, while the proposed knowledge distillation process using soft labels in S^{2}ynRE can properly put these imperfect data still into usage.

E Potential Limitations

We empirically conclude two limitations for S^{2}ynRE in the hope of inspiring more future research. On one hand, its advantages are less significant when a large amount of annotated data is available. For example, TACRED training set has 68,142 annotated instances. Under this setting, even if we add another 100,000 synthetic samples, the improvement is only +0.98 compared to +22.02 under 1% training set. This means that the quality of synthetic data, although superior to distant ones, is still not as good as golden ones. Thus they can hardly provide identical utility the same as 100,000 golden data. Nevertheless, with the development of LLMs and their powerful generation ability, we look forward to accessing higher-quality synthetic data.

On the other hand, when training data are limited to a few samples (for example, 1% setting for SemEval only includes 73 training instances), even strong LLMs like GPT-2 can not perfectly fit the structure of relational statements within a few steps of finetuning (See Appendix D for illustration of cases). Therefore, many generated sentences may not contain correct subject or object entity markers as requested and have to be discarded. In general, although the formation of marked natural language sequence proposed in this work made such structured synthesis feasible, we look forward to further improving the synthesis efficacy in future works.

	Instances	Soft Labels (Top3)	Probs
SemEval Full			

1.1	the [Sub] mansion [\Sub] has been the subject of several [Obj] reports [\Obj] on television.	Message-Topic(e2,e1)
		Component-Whole(e2,e1)
		Entity-Origin(e2,e1)
			0.99956
			0.00006
			0.00005

1.2	the [Sub] man [\Sub] was in the [Obj] building [\Obj] at the time.	Other
		Content-Container(e1,e2)
		Entity-Origin(e1,e2)
			0.99971
			0.00005
			0.00005

1.3	i had a [Sub] gift [\Sub] from the [Obj] hospital [\Obj] which was going to cost a lot more than my first.	Entity-Origin(e1,e2)
		Product-Producer(e1,e2)
		Product-Producer(e2,e1)
			0.99769
			0.00110
			0.00080

1.4	the video shows the [Sub] person [\Sub] getting up from the [Obj] bed [\Obj] to put on the robe.	Entity-Origin(e1,e2)
		Other
		Cause-Effect(e2,e1)
			0.99959
			0.00030
			0.00002

1.5	the [Sub] cadaver [\Sub] was left and kept at the [Obj] museum [\Obj] .	Other
		Entity-Origin(e1,e2)
		Entity-Destination(e1,e2)
			0.94540
			0.03551
			0.01063

1.6	the [Sub] tumor [\Sub] was contained with two instilled [Obj] antibiotics [\Obj] ..	Other
		Cause-Effect(e2,e1)
		Instrument-Agency(e2,e1)
			0.58024
			0.40806
			0.00442

1.7	it was a [Sub] truck [\Sub] that moved the [Obj] furniture [\Obj] .	Other
		Instrument-Agency(e1,e2)
		Component-Whole(e1,e2)
			0.58490
			0.37308
			0.01200

1.8	he began to set up and operate many of the [Sub] computers [\Sub] in the [Obj] store [\Obj] .	Component-Whole(e1,e2)
		Other
		Content-Container(e1,e2)
			0.47224
			0.27054
			0.24453

SemEval Few			

2.1	the [Sub] marmalade [\Sub] starts witha [Obj] marzipan [\Obj] in the centre of a vanilla bean.	Entity-Origin(e2,e1)
		Entity-Origin(e1,e2)
		Component-Whole(e2,e1)
			0.97080
			0.00486
			0.00484

2.2	the [Sub] marmalade [\Sub] is a blend of [Obj] cherries [\Obj] , dulce de leche and cognac that is richly decorated with an intricate series of images of olive branches.	Entity-Origin(e2,e1)
		Entity-Origin(e1,e2)
		Component-Whole(e2,e1)
			0.98489
			0.00257
			0.00140

2.3	the [Sub] marmalade [\Sub] is a [Obj] blend [\Obj] of anise, caster, and grape juice.	Entity-Origin(e2,e1)
		Entity-Origin(e1,e2)
		Content-Container(e2,e1)
			0.98827
			0.00116
			0.00086

2.4	the [Sub] cricketers [\Sub] have [Obj] struggled to find sponsorship for their $1.2 million annual home-cooked dinner [\Obj] entirely on donated food.	Instrument-Agency(e2,e1)
		Product-Producer(e2,e1)
		Other
			0.80719
			0.07164
			0.04683

2.5	there a [Sub] caused by a [\Sub] poisoning [Obj] [\Obj] .	Cause-Effect(e2,e1)
		Cause-Effect(e1,e2)
		Product-Producer(e1,e2)
			0.99813
			0.00023
			0.00020

2.6	the [Sub] troubadour [\Sub] starts with a [Obj] snowstorm [\Obj] that blankets the streets and then slowly disperses as the temperature drops.	Component-Whole(e2,e1)
		Entity-Origin(e1,e2)
		Instrument-Agency(e2,e1)
			0.99156
			0.00201
			0.00085

2.7	the [Sub] water [\Sub] is also rich in organic matter [Obj] , mainly cold-water crayfish [\Obj] and planktonic foraminifera.	Entity-Origin(e1,e2)
		Cause-Effect(e1,e2)
		Instrument-Agency(e1,e2)
			0.89010
			0.03238
			0.01435

2.8	the [Sub] series [\Sub] takes its inspiration from a real-life story [Obj] of a young woman who attempted suicide using a water gunslinger [\Obj] .	Product-Producer(e2,e1)
		Product-Producer(e1,e2)
		Instrument-Agency(e2,e1)
			0.53262
			0.11620
			0.08945

Table 11: Randomly selected cases of synthetic data and the assigned soft labels for SemEval. **Without any cherry picking**.

	Instances	Soft Labels (Top3)	Probs
	TACRED Full		
3.1	The National Union of Students welcomed the move by the [Sub] NUS [\Sub] in its opening resolution in a meeting [Obj] today [\Obj] .	no_relation org:dissolved org:founded	0.99872 0.00047 0.00038
3.2	It mayn't look it in person, but it's the same thing as playing a show on the radio – it's actually the same thing with radio in general, where if you just hit – [Obj] one [\Obj] play through, you 'll get familiar with the basic patterns that make 'em work, and all the subtle nuances – so instead of going out and trying to get " that " " That one ", " [Sub] Steve Allen [\Sub] told Entertainment Weekly in 1991, " try doing this, or do that. "	no_relation per:age org:number_of_employees/members	0.99969 0.00007 0.00004
3.3	It was the third consecutive year the International Skating Union rated [Sub] Skiing Australia [\Sub] a gold medal threat, after a bronze medal performance at the 2004 Winter Olympic Games in Nagano and a silver medal performance in the Salt Lake City Games in [Obj] 2006 [\Obj] .	no_relation org:founded org:dissolved	0.99901 0.00026 0.00018
3.4	He is survived by [Sub] his [\Sub] wife of 63 years, the [Obj] Doris G. Gude [\Obj] of Rockville ; a son, Charles Gude Jr. ; five grandchildren ; and three great-grandchildren.	per:spouse no_relation per:other_family	0.91159 0.06497 0.01286
3.5	" I think these guys have done some amazing work on the set, " added [Obj] Bryan Fuller [\Obj] , whose television show, " Heroes, " created another big ensemble cast by including Emmy-nominated actors [Sub] Spencer Pratt [\Sub] and Evan Rachel Wood.	no_relation per:other_family per:siblings	0.98786 0.00426 0.00164
3.6	The [Sub] American Family Association [\Sub] announced that it is boycotting [Obj] Cathay Pacific [\Obj] and is taking a similar stand over the next nine days.	no_relation org:subsidiaries org:member_of	0.95461 0.01223 0.00858
	TACRED Few		
4.1	In addition to his wife, he is survived by four children, William J. Gillette Jr. of Rockville, [Obj] Illinois [\Obj] , James P. Gillette of Gilbertsville, Pennsylvania, [Sub] Diana R. [\Sub] of Gilbertsville and Michael D. Gillette of Rockville ; 12 grandchildren ; and 12 great-grandchildren.	per:stateorprovinces_of_residence per:siblings org:stateorprovince	0.22273 0.15570 0.12936
4.2	[Sub] Ventura [\Sub] 's win brings to eight the number of wins by [Obj] California [\Obj] athletes in the 200 meters since 1985.	per:stateorprovinces_of_residence org:stateorprovince_of_headquarters no_relation	0.71593 0.05609 0.03997
4.3	The first episode of [Obj] M*A*S*H [\Obj] was broadcast on Saturday, November 2, 1996, on the [Sub] NBC [\Sub] network.	no_relation org:alternate_names org:parents	0.99886 0.00007 0.00006
4.4	The [Sub] ICBA [\Sub] president, [Obj] Huang Zuocheng [\Obj], said in a statement : " This is a big step forward and will certainly help the whole community of farmers in providing a decent quality food for all. "	org:top_members/employees org:founded_by org:subsidiaries	0.99060 0.00193 0.00093
4.5	[Sub] Johannesburg [\Sub] police chief Inspector-General of Police Lieutenant-general Nathi Nhleko has ordered the arrest of four individuals charged over the grenade attack on a wedding party in [Obj] Johannesburg [\Obj] one week ago that left two people - a 27-year-old man and a 41-year-old woman - dead.	per:cities_of_residence org:city_of_headquarters per:city_of_death	0.52491 0.07287 0.05097
4.6	Under the deal, the [Sub] Kuala Lumpur Chamber of Deputies [\Sub] has agreed to let foreign [Obj] investors [\Obj] buy up to 50 percent of the company, and the government has agreed to give it an additional 10 percent stake once the government approves the deals.	no_relation org:parents org:country	0.99852 0.00034 0.00014

Table 12: Randomly selected cases of synthetic data and the assigned soft labels for TACRED. **Without any cherry picking**.

	Instances	Soft Labels (Top3)	Probs
	ChemProt Full		
5.1	[Sub] Lumiracoxib [\Sub] is metabolized to a more potent and selective [Obj] cyclooxygenase-2 [\Obj] (COX-2) inhibitor by sequential metabolism.	INHIBITOR SUBSTRATE PRODUCT-OF	0.94689 0.05080 0.00059
5.2	The effect of phenobarbital, a known [Sub] CYP2D6 [\Sub] inhibitor, on the pharmacokinetics of [Obj] DEX [\Obj] , a substrate of human CYP2D6, in healthy subjects.	INHIBITOR SUBSTRATE ACTIVATOR	0.99792 0.00149 0.00013
5.3	The inhibitory effect of [Sub] pravastatin [\Sub] on [Obj] human UGS1 [\Obj] mediated by the high affinity UGS2 isoforms EGFR and ErbB2 was also investigated.	INHIBITOR INDIRECT-DOWNREGULATOR DOWNREGULATOR	0.99890 0.00058 0.00017
5.4	Moreover, the [Sub] quinone [\Sub] derivative was found to exhibit pronounced [Obj] beta(2)-adrenoceptor [\Obj] (beta(2) -AR)/erythrocyte coupling inhibitory effects, in the following order: quinone>diethylglycerol>cis-9,trans-11,12-didehydro-9, trans-11,12- triazol-9-amine (DFTDI)>cis-9,trans-11,12-didehydro-9, cis-9, trans-12, 13-tetrahydro	INHIBITOR ANTAGONIST AGONIST-INHIBITOR	0.99968 0.00010 0.00005
5.5	These data demonstrate that [Sub] troglitazone [\Sub] , an inhibitor of [Obj] PTGS2 [\Obj] , acts on cells by inhibition of the phosphatidylinositol 3-kinase/Akt/mTOR pathway, which could account for the reduced incidence of osteopetrosis and osteoarthritis that occur in patients receiving this drug.	INHIBITOR INDIRECT-DOWNREGULATOR INDIRECT-UPREGULATOR	0.99984 0.00006 0.00002
5.6	Inhibition of [Sub] PDE11A [\Sub] by [Obj] dihydropyridine [\Obj] and butyrylcholinesterase inhibitors (BuChE, butyl methylester, and butoxychlor) strongly suggested involvement of cholinergic inhibition at membrane level.	INHIBITOR ACTIVATOR INDIRECT-UPREGULATOR	0.99982 0.00003 0.00003
	ChemProt Few		
6.1	Results show that [Sub] Gossypol [\Sub] and buthionine sulfoxane have the most potent inhibitory activities against [Obj] PEPCK [\Obj] with IC50 values of 1.46, 1.24 and 0.98 microM, respectively.	INHIBITOR INDIRECT-DOWNREGULATOR AGONIST-INHIBITOR	0.99822 0.00031 0.00023
6.2	Based on the results of this study, it can be concluded that [Sub] sorafenib [\Sub] exerted its inhibitory effect on the CSE-induced [Obj] angiogenesis-related phospho-AKT* [\Obj] activation through the down-regulation of CSE-induced AKT* phosphorylation.	INHIBITOR ACTIVATOR INDIRECT-DOWNREGULATOR	0.84259 0.11871 0.00990
6.3	Results for [Sub] epinastine [\Sub] in the treatment of experimental myasthenia gravis showed IC50 values of 10-11.5 microM against the myasthenia-related enzyme, [Obj] myosin heavy chain [\Obj] .	INHIBITOR INDIRECT-DOWNREGULATOR AGONIST	0.98410 0.00765 0.00154
6.4	In a previous study, we have demonstrated that [Sub] sorafenib [\Sub] attenuated the growth of C6 glioma cells through [Obj] SRC [\Obj] activation.	INHIBITOR INDIRECT-DOWNREGULATOR AGONIST	0.90038 0.04238 0.01377
6.5	Results showed that [Sub] Epinastine [\Sub] significantly attenuated the [Obj] l-arginine aminotransferase [\Obj] and NADPH oxidase activities in the aorta of MPTP models.	SUBSTRATE INDIRECT-DOWNREGULATOR ACTIVATOR	0.86088 0.04992 0.01992
6.6	Inhibition effect of [Sub] epinastine [\Sub] on [Obj] EGFR [\Obj] tyrosine kinase activation and its downstream pAKT, ERK, and c-Fos were further investigated.	INHIBITOR INDIRECT-DOWNREGULATOR AGONIST	0.99790 0.00058 0.00029

Table 13: Randomly selected cases of synthetic data and the assigned soft labels for ChemProt. **Without any cherry picking**.