Summarization as Indirect Supervision for Relation Extraction

Keming Lu[†], **I-Hung Hsu**[†], **Wenxuan Zhou**[†], **Mingyu Derek Ma**[‡] and **Muhao Chen**[†]
[†]University of Southern California
[‡]University of California, Los Angeles
{keminglu,ihunghsu,zhouwenx,muhaoche}@usc.edu; ma@cs.ucla.edu

Abstract

Relation extraction (RE) models have been challenged by their reliance on training data with expensive annotations. Considering that summarization tasks aim at acquiring concise expressions of synoptical information from the longer context, these tasks naturally align with the objective of RE, i.e., extracting a kind of synoptical information that describes the relation of entity mentions. We present SuRE, which converts RE into a summarization formulation. SuRE leads to more precise and resource-efficient RE based on indirect supervision from summarization tasks. To achieve this goal, we develop sentence and relation conversion techniques that essentially bridge the formulation of summarization and RE tasks. We also incorporate constraint decoding techniques with Trie scoring to further enhance summarization-based RE with robust inference. Experiments on three RE datasets demonstrate the effectiveness of SuRE in both full-dataset and low-resource settings, showing that summarization is a promising source of indirect supervision signals to improve RE models.[1]

1 Introduction

Relation extraction (RE) aims at extracting relations between entity mentions from their textual context. For example, given a sentence *"Steve Jobs is the founder of Apple"*, an RE model would identify the relation *"founded"* between mentioned entities *"Steve Jobs"* and *"Apple"*. RE is a fundamental natural language understanding task and is also the essential step of structural knowledge acquisition for constructing knowledge bases. Hence, advanced RE models is crucial for various knowledge-driven downstream tasks, such as dialogue system (Liu et al., 2018; Zhao et al., 2020), narrative prediction (Chen et al., 2019), and question answering (Yasunaga et al., 2021; Hao et al., 2017).

Given sentences with detected pairs of entity mentions, most recent studies formulate RE as multi-class classification (Zhou and Chen, 2022; Yamada et al., 2020; Baldini Soares et al., 2019). Models presented in these studies employ pre-trained language models (PLM) equipped with classification heads and are finetuned with a cross-entropy loss. Although such methods have achieved enhanced performances on several benchmarks (Zhang et al., 2017; Stoica et al., 2021; Alt et al., 2020), they fall short of capturing the semantic meanings of the relations. This shortage hinders PLMs from effectively matching the sentential context with the relations that are merely converted as logits. On the other hand, obtaining high-quality annotations for RE is often costly due to the difficulty for annotators to recognize and mutually agree on such structural information. This represents another challenge for RE models that have relied on direct supervision from sufficient end-task training data. Existing literature finds that classification models have drastically degraded performance under low-resource scenarios (Sainz et al., 2021), showing that label efficiency is a vital issue when adopting prior methods in real application scenarios. To combat this issue, we aim at investigating an indirectly supervised method for RE, which allows the use additional supervision signals that are not specific to RE without solely relying on direct RE annotations.

This study proposes SuRE (Summarization as Relation Extraction), which reformulates and addresses RE as a summarization task.[2] Summarization seeks to acquire concise expressions of synoptical information from longer context (El-Kassas et al., 2021), which aligns well with the objective of RE if we consider the relation between entities as

[1]Our code is public available at https://github.com/luka-group/SuRE

[2]In this paper, we specifically consider abstractive summarization instead of extractive summarization. Since relation labels are often not directly expressed in sentences, extractive summarization does not always support the inference of RE.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6604 - 6623
December 7-11, 2022 ©2022 Association for Computational Linguistics

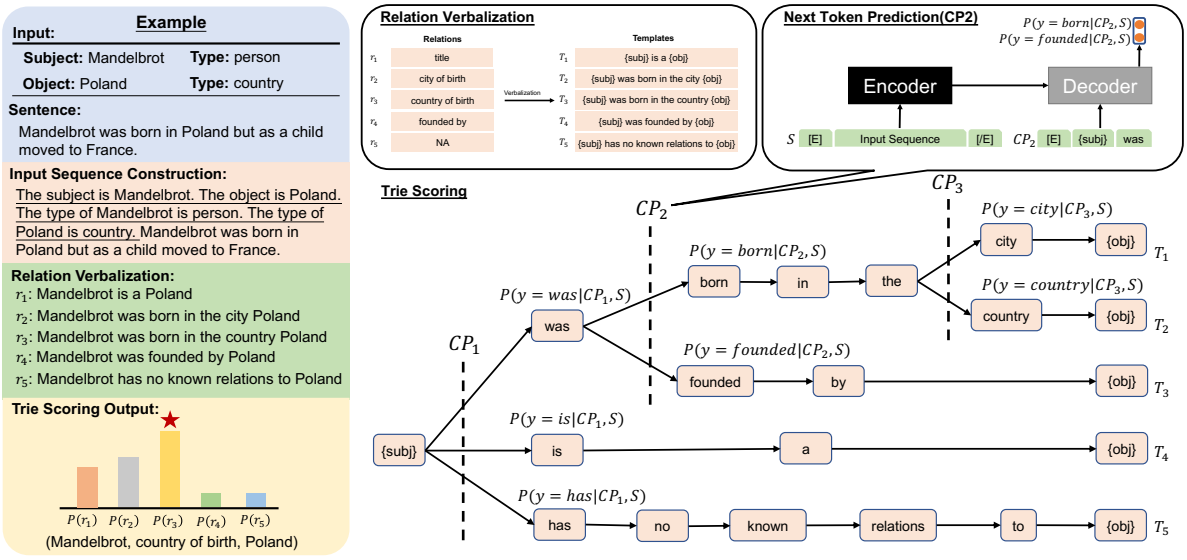

Figure 1: Overview of SURE inference with an example. SURE constructs input sequences and verbalizes candidate relations with semantic templates (Relation Verbalization subfigure). Then we use a summarization model to find the best prediction by Trie scoring technique (Trie scoring subfigure). This model is pretrained on summarization tasks and then simply finetuned with input sequence target verbalized ground-truth relation. For each common prefix (CP), we will calculate the probability of each relation candidate, which is obtained using a trained summarization model, as shown in the Next Token Prediction subfigure. The "{subj}" and "{obj}" are two placeholders representing subject and object entity in each sample.

one aspect of synoptical information in the sentential context. Such an affinity of task objectives naturally motivates us to leverage indirect supervision from summarization tasks to improve RE models. In comparison to a multi-class classifier, summarizing the relation information also allows generating a semantically rich representation of the relation. Furthermore, unlike existing RE models that rely on costly manual annotations of structural information on training sentences, summarization tasks allow training with considerably richer and unannotated parallel text corpora.[3] Hence, summarization tasks can bring in abundant indirect supervision signals, and can potentially lead to label-efficient models under scenarios without much task-specific annotation.

Fig. 1 illustrates the structure of SURE. Specifically, SURE transforms RE to a summarization task with relation and sentence conversion techniques (§3.2), and applies constrained inference for relation prediction (§3.4). We deploy an entity information verbalization technique to highlight the sentential contexts with entity information, and verbalize relations into template-style short summaries. In that way, the converted inputs and outputs of RE naturally suit a summarization model. Then, we adapt a summarization model to the RE task by finetuning it on the converted RE data (§3.3). During inference, a Trie scoring technique is designed to infer the relations (§3.4). In this way, SURE fully utilizes indirect supervision from summarization, allowing a precise RE model to be obtained even in low-resource scenarios.

The contributions of this work are two-folds. First, to the best of our knowledge, this is the first study on using indirect supervision from summarization for RE. Since the objective of summarization naturally aligns with RE, it allows precise RE models to be trained without solely relying on direct task annotations, and benefits with robust RE under low-resource scenarios. Second, we investigate input conversion techniques that effectively bridge the formulation of summarization and RE tasks, as well as constraint techniques that further enhance the inference of summarization-based RE. Our contributions are verified with experiments on three widely used sentence-level RE datasets, TACRED, TACREV, and SemEval, as well as three low-resource settings of TACRED. We observe that SURE outperforms various baselines, especially in the low-resource setting with 10% TACRED train-

[3]Particularly, summarization corpora are constructed in the scales from hundred thousands (Nallapati et al., 2016; Narayan et al., 2018) to million scales (Yin et al., 2021), and may be rapidly augmented in large scales from easy-to-consume data sources (e.g., community Q&A platforms (Mishra et al., 2021) and scientific paper abstracts (Cohan et al., 2018)).

ing data. SuRE also achieves SOTA performance with 75.1% and 83.5% in micro F1 on TACRED and TACREV, respectively. We also perform comprehensive ablation studies to show the effectiveness of indirect supervision from summarization and the best options of input conversion for SuRE.

2 Related Work

Relation extraction. Recent studies on (sentence-level) RE typically formulate the task as multi-class classification tasks by finetuning pretrained language models (Wu and He, 2019; Hsu et al., 2022a; Lyu and Chen, 2021) or developing pretraining objectives for RE (Baldini Soares et al., 2019; Peng et al., 2020). For example, Wu and He (2019) enrich the contextual representation of the PLM with marked-out subject and object entity mentions. Lyu and Chen (2021) propose a model-agnostic paradigm that introduces mutual restrictions of relations and entity types into relation classifiers. On the basis of PLMs, some studies further improve RE with external knowledge from knowledge bases (KBs) (Yamada et al., 2020; Peters et al., 2019; Zhang et al., 2019). More studies have been introduced to improve entity pair representations and classifiers for RE such that we cannot exhaust them in this short summary. We refer readers to the recent benchmarking study (Zhou and Chen, 2022).

Another threads of recent effort in RE introduce several reformulations of RE with prompt learning (Han et al., 2021; Chen et al., 2022). Specifically, Han et al. (2021) propose prompt tuning methods for RE by applying logic rules to construct hierarchical prompts. Chen et al. (2022) leverage prompt tuning for RE by injecting semantics of relations and entity types. Instead of leveraging pretrained masked language models, we use generative approaches to solve RE.

Indirect supervision. Indirect supervision (He et al., 2021) methods often modify the training and inference processes on a task into a different formation, hence allowing the use of additional supervision signals that is not specific to this task. Levy et al. (2017) show that RE can be addressed as answering reading comprehension questions and improved by the training process of a machine reading comprehensive task. Similarly, Wu et al. (2020) also employ QA data to improve model generalization abilities in coreference resolution. Yin et al. (2020) propose a few-shot NLI-based framework to address different tasks, such as question answering and coreference resolution. Li et al. (2022) further improve this strategy by incorporating NLI with learning-to-rank, leading to a robust system for ultra-fine entity typing (Choi et al., 2018). Similar idea of leveraging NLI as indirect supervision signal is applied by Sainz et al. (2021), which focuses on low-resource RE task. As discussed, the objective of RE aligns well with that of a summarization task. While there is no prior study that investigates indirect supervision from summarization, this is exactly the focus of our study.

Generative approaches for discriminative tasks. Formulating discriminative tasks as generation tasks can be an efficient way to guide PLMs to leverage semantics of decision labels (Huang et al., 2021; Hsu et al., 2022b; Huang et al., 2022; Yuan et al., 2022). Instead of predicting classification logits, a common paradigm for these models is to represent the class as a concise structure and employ controlled decoding for generation. Several studies (Zeng et al., 2018, 2020; Ye et al., 2020; Cao and Ananiadou, 2021) use sequence-to-sequence-based models to generate relations written in a triplet format. Paolini et al. (2020) incorporate many structured prediction tasks, including RE, into machine translation. Huguet Cabot and Navigli (2021) simplify RE as expressing relations as a sequence of text to perform end-to-end generation of relations. These works mostly formulate RE as text-to-structure learning instead of generating natural language sentences that is a more natural target to exploit the power of pretrained generative models (Hsu et al., 2022b). Additionally, they do not include indirect supervision of summarization, which is naturally close to the objective of RE and has the potential to benefit RE performance.

3 Method

In this section, we describe SuRE, a model for addressing RE with summarization. We introduce preliminaries (§3.1), how RE data are converted to suit summarization tasks (§3.2), training (§3.3), and constrained inference of SuRE (§3.4).

3.1 Preliminaries

Problem Definition. The input to the sentence-level RE is a sentence s with entity mentions e_1 and e_2 [4] , where their auxiliary entity type information

[4] An entity mention is presented as an *entity name* in text, and is structured as a *mention span* with position information.

t_1, t_2 is given. An RE model aims at inferring the relation r between the subject and object entities e_1 and e_2 from a set of candidate relations $\mathcal{R} = \mathcal{R}_P \cup \{\emptyset\}$, which include positive relations $\mathcal{R}_P$ and a Not-Available (NA) relation $\emptyset$. We also involve type-related candidate relations $\mathcal{R}(t_1, t_2)$, which is a subset of $\mathcal{R}$ which has specific types of head and tail entities.

Overview. Fig. 1 demonstrates the overview of SURE. The summarization task takes a context as the input sequence and a summary target is expected to be generated. To formulate RE as summarization, we first need to hint the summarization model which entity pair is targeted for summarization. To do so, we process the input sentence such that entity mentions and their type information will be highlighted (§3.2). We explore existing entity marking tricks (Zhou and Chen, 2022) and also develop entity information verbalization technique that directly augments entity information as part of the context. The processed sentence will then be fed into SURE. The summary targets for SURE is created via verbalizing existing RE labels to templates, such as the Relation Verbalization subfigure in Fig. 1. In the training process (§3.3), SURE uses pretrained summarization models as a start point, and finetunes them with processed sentences as the input and verbalized relation descriptions as the targets. During inference, we incorporate several constrained inference techniques to help SURE decide the inferred relation (§3.4).

3.2 Relation and Sentence Conversion

Summarization takes text sequences as inputs and outputs. We hereby describe the input sequence construction and relation verbalization, representing two essential techniques for converting RE data to suit the summarization task.

Input sequence construction. Relation extraction focuses on analyzing the interaction between two specific entities, so we need to further process source sentences so that additional information can be involved and captured by summarization models. SURE explores a series of sentence processing techniques that highlight and incorporate entity information, aiming for identifying a technique that suits the summarization task well. Entity information includes entity names, types, and spans, which is useful for inferring the relation. We explore with two strategies for processing the source sentence.

- **Entity typed marker.** Various entity marking techniques are widely adopted in previous multi-class classification RE systems (Zhang et al., 2017, 2019; Wang et al., 2021; Zhou et al., 2021; Zhong and Chen, 2021; Zhou and Chen, 2022). We list all the techniques in Appx. Tab. 9. Our preliminary experiments (Appx. Tab. 10) find that the following typed entity marker technique with punctuation works the best for SURE among these marking methods (inserted typed markers are in blue, while the original text is in black):

  > @ * *person* * Mandelbrot @ was born in Poland but as a child moved to # ∧ *country* ∧ France #.

- **Entity information verbalization.** We develop a simple sentence rewriting technique that directly describes entity information as an augmented part of the linguistic context (in blue):

  > *The subject entity is Mandelbrot. The object entity is France. The type of Mandelbrot is person. The type of France is country.* Mandelbrot was born in Poland but as a child moved to France.

  Although this technique cannot encode entity span information, it keeps the input data close to natural language instead of adding special tokens. This aligns well with the indirect supervision from summarization. Thus, it shows better performance to the entity typed marker technique, as shown in the ablation study (§4.3).

We hereby list all input conversion techniques we experiment in this work in Tab. 9. Tab. 10 shows additional results on the *bart-large-cnn* model, which provides the same conclusion as results on *pegasus-large*. We also compare this mixing technique in the ablation study (§4.3) and find it achieves the best performance in the full training setting.

Relation verbalization. The target of summarization is verbalized by a set of simple *semantic templates*, as shown in the Relation Verbalization subfigure of Fig. 1. Each template contains *{subj}* and *{obj}* placeholders to be filled with subject and object entity mentions in the sentence. The templates seek to form short summaries that describe the relations between two entities, and will be used for models' training and inference. Semantic templates are also leveraged in Sainz et al. (2021), where the templates are used as hypotheses for NLI-based RE. However, their templates are specifically designed for NLI. We adapt minimal additional updates to

their templates so the templates can better fit summarization and less human effort is involved. We let the subject entities always appear in the head of sentences while the objects are in the tail. For example, "*org:parents*" relation is verbalized with templates "*{subj} has the parent company {obj}*". Detailed semantic templates are demonstrates in Appx. Tab. 11 and Tab. 14. Notice that semantic meaning of a relation can be verbalized in various ways, so we also construct alternative semantic templates and discuss how different templates influence model performance in §4.3.

For comparison, we also experiment with *structural templates* that are widely used in existing sequence-to-structure methods (Zeng et al., 2018, 2020). As listed in Appx. Tab. 13, these templates directly concatenate entity names and relations, which are shown by our ablation study (§4.3) to be less effective than the semantic templates.

Discussion. SURE requires manually designed templates of relations for both training and inference. To minimize manual effort and give a fair comparison to prior work, we adopt the same relation verbalization templates from Sainz et al. (2021). They restrict the influence of human effort by limiting the time for creating the templates and build 2 templates in average for each relation. For simplicity, we adopt one template for each relation from their templates, which suggests SURE will need less manual effort for template design.

3.3 Training Process

The aforementioned rewriting and verbalization techniques (§3.2) highlight the sentential contexts with entity information, and convert the extraction as summary. Hence, the converted inputs and outputs of RE naturally suit the summarization task. This allows us to train a summarization model first using large parallel training corpora for abstractive summarization such as XSum (Narayan et al., 2018) or CNN/Dailymail (Hermann et al., 2015), and further adapt it to learn to "summarize" relation. In our experiments, SURE adopts checkpoints of pretrained generative models (Lewis et al., 2020; Zhang et al., 2020) that are pretrained on summarization tasks as starting points. Then, we follow the same finetuning process of seq-to-seq training with the cross entropy loss to finetune the model on converted RE data. In this way, SURE can leverage indirect supervision obtained from the summarization task to enhance RE.

3.4 Inference

The inference process of SURE involves first applying Trie scoring to rank the possibility of each relations, and setting entity type constraints. The score is further calibrated to make selective predictions between known and NA relations.

Trie scoring. Summarization models employ beam search techniques to generate sequential outputs (Zhang et al., 2020), while RE seeks to find out the relation described the input. To support relation prediction using a summarization model, we develop an inference method that will rank each relation candidate by using the summarization model as proxies for scoring. Inspiring by the Trie constrained decoding (Cao et al., 2021), we develop a Trie scoring technique, allowing *efficient ranking* for candidate relation verbalizations. Instead of calculating the probability of whole relation templates for ranking, our method conducts a traverse on the Trie and estimates the probability of each relation candidate as a path probability on the Trie.

Given the set of tokenized templates of all candidate relations as $\mathcal{T} = \{T_i\}_{i=1}^{n_r}$, we build a Trie (Aoe et al., 1992) by combining the prefixes of all templates, as an example in the Trie Scoring subfigure in Fig. 1. A path of a relation template can be described as a sequence of decision processes, which goes from the root to a leaf node. If we denote $\mathcal{N}^f$ as the set of forky nodes (the nodes with more than one child), then the probability of a path can be estimated by continued producting the probability of choosing a specific child in a forky node $n_i^f \in \mathcal{N}^f$. Specifically, we denote the path from root to n_i^f as CP_i, which is the common prefix for all templates in the sub-tree with n_i^f as the root. For example, CP_2 = "{subj} was" in Fig. 1 is the common prefix of templates T_1 and T_2. If we denote the children of n_i^f as a set $\mathbf{C}(n_i^f)$, the prediction probability of relation r_i can be calculated by

$$p(r_i) = \Pi_{\forall n_j^f, c \in \mathbf{C}(n_j^f)} p(c \in T_i | \text{CP}_j, S).$$

$p(c \in T_i|\text{CP}_j, S)$ thereof is the probability for the model to generate the next token c given the previous common prefix CP_j, and c is selected from $C(n_i^f)$ on T_i. This probability is calculated using a seq-to-seq summarization model with input sentence S as encoder input and CP_i as decoder input prefix, such as the illustration in the Next Token Prediction(CP2) subfigure in Fig. 1.

Type constrained inference. Type constrained inference emerges in many recent works (Lyu and Chen, 2021; Sainz et al., 2021; Cohen et al., 2020). By constructing a type-relation mapping from the training set, models can only predict valid relations given entity types, which significantly shrinkages the size of the candidate relation set. Type constrained inference can be easily incorporated into SURE by pruning the templates of invalid entity types from Trie scoring. The merit of type constrained inference will be discussed in Appx. §A.2.

Calibration for NA relation. Considering that it is not possible for a relation ontology to exhaust all possible relations between any entities, the inference of a trained RE model can naturally be exposed to many instances where not a positive relation from $\mathcal{R}_P$ is expressed. Hence, it is particularly important to enforce the model to selectively make a decision between positive relations or predicting abstention. This is realized by a calibration technique in SURE, where a score threshold s is set for NA and is calibrated as below:

$$\hat{r} = \begin{cases} \arg\max_{r_i \in \mathcal{R}_P} p(r_i) \ , & p(\text{NA}) \leqslant s \\ \text{NA} \ , & p(\text{NA}) > s \end{cases}$$

The best threshold s is found on the development set and is used as a fixed threshold in inference.

4 Experiments

In this section, we present the experimental evaluation of SURE for RE under both high- and low-resource setups (§4.1-§4.2). In addition, we also conduct comprehensive ablation studies to investigate the effectiveness of the incorporated techniques in SURE (§4.3).

4.1 Experimental Setup

Datasets. We conduct experiments on three widely used sentence-level RE benchmarks: SemEval 2010 Task 8 (SemEval; Hendrickx et al. 2010), TACRED (Zhang et al., 2017), and TACRED-Revisited (TACREV; Alt et al. 2020).

SemEval is an RE dataset which does not provide entity types, so we simply remove the processing of entity types in sentence conversion (§3.2) to adapt SURE on this dataset. TACRED contains entity pairs drawn from the yearly TAC-KBP challenge. We list our templates for SemEval and TACRED in Appx. Tab. 14 and Tab. 11, respectively. TACREV relabeled develop and test sets of TACRED to correct mislabeled entity types and relations. TACREV shares the same templates with TACRED since they have exactly the same relations. Statistics of these datasets are showed in Appx. Tab. 6. We report macro F1 on SemEval with the official grading script for this benchmark[5], and micro F1 on TACRED and TACREV to keep consistency with previous works (Yamada et al., 2020; Zhou and Chen, 2022).

Baselines. We compare SURE with 8 recent classification-based RE methods: (1) **SpanBERT** (Joshi et al., 2020) is a pre-training method designed to better represent and predict spans of text; (2) **KnowBERT** (Peters et al., 2019) is a PLM embedded multiple KBs; (3) **R-BERT** (Wu and He, 2019) uses a PLM to encode processed sentences where subject and object entities are marked out; (4) **MTB** (Baldini Soares et al., 2019) builds task-agnostic relation representations solely from entity-linked text; (5) **K-Adapter** (Wang et al., 2021) infuses knowledge into pretrained language models with adapters. (6) **LUKE** (Yamada et al., 2020) modifies the PLM with an entity-aware self-attention mechanism; (7) **IRE$_{\text{RoBERTa-large}}$** (Zhou and Chen, 2022) is an improved baseline model incorporated with typed entity markers; (8) **RECENT** (Lyu and Chen, 2021) introduces type constraint (§3.4) and achieve state of the art performance on TACRED. We also compare SURE with two indirect supervision methods, i.e. (9) **NLI$_{\text{DeBERTa}}$** (Sainz et al., 2021) that formulates RE as NLI, and (10) **KnowPrompt** (Chen et al., 2022) that formulates RE as prompt tuning.

Low resource setting. We evaluate the performance of SURE under low-resource scenarios. To do so, we use the same splits of Sainz et al. (2021) to build the TACRED datasets with 1/5/10 percent of training and development samples.

Model configurations. We develop SURE based on two widely pretrained generative models BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020). BART is a denoising autoencoder for pretraining sequence-to-sequence models. We use two summarization models *BART-large-cnn* and *BART-large-xsum* that are finetuned with CNN/Dailymail (Hermann et al., 2015) and XSum (Narayan et al., 2021), respectively. We also consider *BART-large* as a baseline without indirect supervision of summarization. PEGASUS is a sequence-to-sequence model pretrained with a

[5]The metric calculated by the script is the macro F1 on (9+1)-way classification taking directionality into account.

	Dataset	TACRED 1%	TACRED 5%	TACRED 10%	TACRED 100%	TACREV	SemEval
Classification-based	SpanBERT (Joshi et al., 2020)	0.0‡	28.8‡	1.6‡	70.8	78.0	––
	KnowBERT (Peters et al., 2019)	––	––	––	71.5	79.3	89.1
	RoBERTa (Wang et al., 2021)	7.7‡	41.8‡	55.1‡	71.3	––	––
	R-BERT (Wu and He, 2019)	––	––	––	69.4	––	89.3
	MTB (Baldini Soares et al., 2019)	––	––	––	71.5	––	89.5
	K-Adapter (Wang et al., 2021)	13.8‡	51.6‡	56.0‡	72.0	––	––
	LUKE (Yamada et al., 2020)	17.0‡	51.6‡	60.6‡	72.7	80.6	––
	$\text{IRE}_{\text{RoBERTa-large}}$ (Zhou and Chen, 2022)	46.3†	63.6†	67.0†	<u>74.6</u>	<u>83.2</u>	––
	RECENT (Lyu and Chen, 2021)	40.0†	53.3†	54.2†	67.3◇	––	––
Ind Sup	$\text{NLI}_{\text{RoBERTa}}$ (Sainz et al., 2021)	<u>56.1</u>	64.1	67.8	71.0	––	––
	$\text{NLI}_{\text{DeBERTa}}$ (Sainz et al., 2021)	**63.7**	**69.0**	67.9	73.9	––	––
	KnowPrompt (Chen et al., 2022)	51.0†	61.0†	65.2†	72.4	82.4	<u>89.6</u>◇
Proposed	$\text{SURE}_{\text{BART-large}}$	43.6	63.8	67.9	73.3	79.2	86.3
	$\text{SURE}_{\text{BART-large-cnn}}$	50.4	<u>65.3</u>	<u>68.7</u>	73.6	81.0	<u>89.6</u>
	$\text{SURE}_{\text{BART-large-xsum}}$	50.3	64.3	68.0	73.3	81.0	89.1
	$\text{SURE}_{\text{PEGASUS-large}}$	52.0	64.9	**70.7**	**75.1**	**83.3**	**89.7**

† indicates models we re-implement using their official code under the same low-resource setting.
‡ indicates results collected from Sainz et al. (2021).
◇ indicates we reproduce the baseline results (Appx. §A.4).

Table 1: Result of SURE compared with existing methods under both low resource on TACRED and full training scenarios on TACRED, TACREV and SemEval. We report micro F1 on TACRED and TACREV, and report macro F1 on SemEval. Baseline F1 scores on the table without special tags are reported by their original studies. Hyphens indicate unavailable results in prior studies. We report SURE performance with entity information verbalization for consistency. However, SURE achieves better performance (83.5%) with mix technique of entity information verbalization and entity typed marker on TACREV (Tab. 3). We run our models with three different seeds and report the median. The best scores are identified with **bold** and the second best scores are <u>underlined</u>.

Template	TACRED 0%	TACRED 1%	TACRED 5%	TACRED 10%	TACRED full	TACREV full
SEMANTIC1	20.6	**52.0**	64.9	**70.7**	**75.0**	**83.3**
SEMANTIC2	18.5	49.5	**66.9**	69.6	73.5	82.0
STRUCTURAL	18.5	46.8	61.8	69.1	74.4	82.2

Table 2: Analysis of different template designs. The highest scores are highlighted with bold formation. These experiments are conducted with the entity information verbalization technique.

gap sentences generation task, which significantly benefits various summarization downstream tasks. Similarly, we use *PEGASUS-large* as a stronger initial checkpoint than BARTs. We use grid search to find optimal hyperparameters for finetuning summarization models. The best hyperparameters of our experiments and re-implementation of baselines are shown in Appx. §A.4.

4.2 Results

We present our main results on both full training and low-resource settings in Tab. 1. We report the performance of SURE with entity information verbalization technique, which is proved to be the best way of input sequence construction (§3.2) in most settings as shown in our ablation study (§4.3).

Performance comparison. With 1%, 5%, or 10% training data of TACRED, SURE with summarization backbones ($\text{SURE}_{\text{BART-large-cnn}}$, $\text{SURE}_{\text{BART-large-xsum}}$ and $\text{SURE}_{\text{PEGASUS-large}}$) and other baselines with indirect supervision consistently outperform classification-based RE models except IRE, which indicates the benefit of indirect supervision. Although $\text{NLI}_{\text{DeBERTa}}$ significantly outperforms other methods with 1% and 5% training data, SURE has significant improvement on 10% TACRED, which outperform $\text{NLI}_{\text{DeBERTa}}$ for 2.8% and KnowPrompt for 5.5%. Furthermore, the performance of SURE is 1.2% higher than that of $\text{NLI}_{\text{DeBERTa}}$, and 2.7% higher than that of KnowPrompt in full training setting of TACRED, which suggests that SURE achieves the best performance with adequate training samples among all indirect supervision baselines. SURE also achieves the best performance among all baselines, and exceeds the second best model $\text{IRE}_{\text{RoBERTa-large}}$ by 0.5% in F1. Besides, SURE also achieves the best F1 score on TACREV. And it outperforms all classification-based models on SemEval, while its performance is comparable to KnowPrompt.

Dataset	TACRED				TACREV
Split	1%	5%	10%	full	full
Entity information verbalization	**52.0**	**64.9**	**70.7**	**75.1**	83.3
Entity typed marker (punct)	46.3	57.5	59.0	73.3	80.4
Entity information verbalization + Entity typed marker (punct)	47.6	60.2	67.8	75.0	**83.5**

Table 3: Analysis of different input formulation techniques on *PEGASUS-large*

Effectiveness of indirect supervision. We further evaluate SURE based on different pretrained models, as shown in Tab. 1. We observe that models finetuned on summarization tasks (CNN and XSum) generally lead to better performance, especially in the low-resource setting. For example, $\text{SURE}_{\text{BART-large-cnn}}$ outperforms $\text{SURE}_{\text{BART-large}}$ by 6.8% on the 1% split of TACRED, while this improvement diminishes to 0.3% on the full dataset. Besides, pretrained models that perform better on summarization tasks also indicate better performance on RE. Particularly, SURE based on *PEGASUS-large*, which outperforms *BART-large* on summarization tasks, outperforms all other versions under both low-resource and full-dataset setting. Both observations show a strong correlation between the performance in summarization and RE, indicating that indirect supervision from summarization is beneficial for RE models.

4.3 Ablation Study

We provide the following analyses to further evaluate core components of SURE, including different template designs, sentence conversion techniques and Trie scoring. We also report the ablation study on type constrained inference and calibration of NA relation in Appx. §A.2 and A.3, respectively.

Relation template design. Template design is a manual part of SURE. The semantic meaning of a relation can be verbalized in different ways, leading to varied performance. In Tab. 2, we compare three representative types of templates with *pegasus-large* in this ablation study to show how verbalization templates influence the performance of SURE. SEMANTIC1 thereof denotes semantic templates beginning with subject entities and ending with object entities, which is showed in Appx. Tab. 11. SEMANTIC2 are also semantic templates that intuitively describe the relation between two entiteis with a pattern "*The relation between {subj} and {obj} is {relation}*", which is showed in Appx. Tab. 12. STRUCTURAL marks structural templates forming in a triplet structure "{subj} {relation} {obj}", which is showed in Appx. Tab. 13. Furthermore, we also set up a zero-shot setting where the model directly infers on the test set of TACRED after calibration on the development set. The results from different templates are reported on both low-resource and full-training scenarios. First of all, we observe that the two semantic templates consistently outperform structural templates, indicating that semantic templates are more suitable for acquiring indirect supervision from summarization. Besides, comparing two semantic templates, we find that *Semantic1* works better with *pegasus-large*, which suggests that the optimal verbalization may vary. And this difference is more obvious under low resource scenarios. Consequently, zero-shot inference is an effective and efficient method for evaluating manual-designed templates. In future work, we can investigate how to improve this part by prompt tuning.

Input conversion. We conduct experiments to evaluate various input sentence conversion techniques for injecting entity information into source sentences (§3.2). We first conduct experiments on six different input formulations on *bart-large-cnn*, which is listed in Tab. 9 and results are shown in Appx. Tab. 10. This experiment indicates entity typed marker with punctuation is the best technique for SURE among all entity marker techniques. Then, we further evaluate three techniques on *pegasus-large* under both full training and low resource scenarios. Tab. 3 shows entity information verbalization achieves significantly better performance under low resource scenarios compared with marker and mix techniques. This is because entity information verbalization transforms input to better fit the input of summarization, while additional markers need more data to learn their representations. In the full training setting, the mixing technique marginally outperforms entity information verbalization.

Trie scoring. Trie scoring uses teacher forcing to constraint models focusing on candidate templates and have the advantage of efficiency com-

Technique	Example	M[1]	S[1]	T[1]	TS[1]
Entity information verbalization	`Context`[2]. ... {subj} ... {obj} ...	✓	✗	✓	✓
Entity marker (Zhang et al., 2019)	... <e1> {subj} </e1> ... <e2> {obj} </e2> ...	✓	✓	✗	✗
Entity typed marker (Zhong and Chen, 2021)	... <e1-{subj-type}> {subj} </e1-{subj-type}> ... <e2-{obj-type}> {obj} </e2-{obj-type}> ...	✓	✓	✓	✗
Entity typed marker (punct) (Zhou and Chen, 2022)	... @ * {subj-type} * {subj} @ ... # ∧ {obj-type} ∧ {obj} # ...	✓	✓	✓	✓
Entity information verbalization + Entity typed marker	`Context`[2]... @ * {subj-type} * {subj} @ ... # ∧ {obj-type} ∧ {obj} # ...	✓	✓	✓	✓
Entity information verbalization + Entity typed marker (punct)	`Context`[2]... @ * {subj-type} * {subj} @ ... # ∧ {obj-type} ∧ {obj} # ...	✓	✓	✓	✓

[1] Column names are short for mentions, spans, types and type semantics, respectively.
[2] Augmented context are generated with the template: "*The {subj} entity is {subj} . The {obj} entity is {obj} . The type of {subj} is {subj-type} . The type of {obj} is {obj-type}* "

Table 4: Sentence processing techniques for incorporating entity information.

Technique	TACRED			TACREV		
	P	R	F1	P	R	F1
Entity information verbalization	71.8	75.1	73.4	78.4	83.8	81.0
Entity tag	71.3	71.0	71.1	81.0	75.2	78.0
Entity typed tag	73.5	69.7	71.5	79.9	79.6	79.8
Entity typed tag(punct)	70.3	74.0	72.1	81.1	79.3	80.2
Entity infromation verbalization + Entity tag	73.4	71.8	72.6	78.7	81.7	80.2
Entity infromation verbalization + Entity typed tag	70.6	73.5	72.1	82.8	80.0	**81.4**
Entity infromation verbalization + Entity typed tag(punct)	72.6	74.5	**73.6**	82.1	79.8	81.0

Table 5: Analysis of different input formulation techniques on *bart-large-cnn*. We report micro F1 scores on both datasets. The best F1 score is identified with **bold**.

pared with directly comparing likelihoods of all templates. Furthermore, we also make comparisons between Trie scoring and two basic scoring methods on full TACRED with SuRE$_{\text{pegasus-large}}$. The first one is to generate the summary of an example and uses ROUGE-L (Lin, 2004) scores between summary and candidate templates as prediction scores. This method achieves 74.7% on TACRED, which is 0.4% less than that of Trie scoring. Another method is adding a copy mechanism (Zeng et al., 2018) to ensure summaries begin with head entities, which achieves a comparable performance of the previous method (74.8%), which further proves the advantages of Trie scoring.

5 Conclusion

We propose a new method SuRE that leverages indirect supervision from summarization tasks to improve RE. To do so, we verbalize relations with semantic templates, and augments entity information as parts of the linguistic context in the inputs to allow them to suit the formation of summarization. We also incorporate SuRE with constrained inference based on Trie scoring, as well as inference with abstention and entity type constraints. Extensive experiments show that such indirectly supervised RE by SuRE lead to more precise and resource-efficient RE. Future work includes further developing our model on document-level RE tasks and minimizing manual effort in template design with prompt tuning (Li and Liang, 2021).

Acknowledgement

We appreciate the reviewers for their insightful comments and suggestions. This work is partly supported by the National Science Foundation of United States Grant IIS 2105329, and a Cisco Faculty Research Award.

Limitations

SuRE assumes that summarization data and manual-designed verbalization templates of relations are easy to obtain. This assumption is hold in the general domain. However, obtaining such data and templates can still be difficult in specific lower-resource domains. For example, summarization data in other languages are not as rich as those in English. Hence, SuRE may benefit less from indirect supervision signals when it is adapted to multilingual scenarios. Besides, designing templates in specific domains, such as biomedical relation extraction, may require extra involvement of expert

knowledge. Although we put certain manual efforts in template design, automatically optimizing templates are also feasible for SURE and can be explored in future work, as described in the Conclusion section.

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.

Jun-Ichi Aoe, Katsushi Morimoto, and Takashi Sato. 1992. An efficient implementation of trie structures. *Software: Practice and Experience*, 22(9):695–721.

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.

Jiarun Cao and Sophia Ananiadou. 2021. GenerativeRE: Incorporating a novel copy mechanism and pretrained model for joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2119–2126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Jiaao Chen, Jianshu Chen, and Zhou Yu. 2019. Incorporating structured commonsense knowledge in story completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, pages 6244–6251.

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 . *International World Wide Web Conferences (WWW)*.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation classification as two-way span-prediction. *arXiv preprint arXiv:2010.04829*.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231, Vancouver, Canada. Association for Computational Linguistics.

Hangfeng He, Mingyuan Zhang, Qiang Ning, and Dan Roth. 2021. Foreseeing the benefits of incidental supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1782–1800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

I-Hung Hsu, Xiao Guo, Premkumar Natarajan, and Nanyun Peng. 2022a. Discourse-level relation extraction via graph pooling. In *The Thirty-Sixth AAAI Conference On Artificial Intelligence Workshop on Deep Learning on Graphs: Method and Applications (DLG-AAAI)*.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022b. Degree: A data-efficient generative event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kuan-Hao Huang, I-Hung Hsu, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine entity typing with indirect supervision from natural language inference. *Transactions of the Association for Computational Linguistics*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 390–395, Online. Association for Computational Linguistics.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2020. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

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.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word

representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

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.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the Thirty-fifth AAAI Conference on Aritificial Intelligence*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

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*, pages 2361–2364.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. 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.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. 2020. Contrastive triple extraction with generative transformer. *arXiv preprint arXiv:2009.06207*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. *arXiv preprint arXiv:2010.02584*.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9507–9514.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)*.

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*, volume 35, pages 14612–14620.

A Appendix

A.1 Dataset Statistics

Statistics of the RE datasets are listed in Tab. 6.

Dataset	#Train	#Dev	#Test	#Relation
SemEval	8000	-	2717	19
TACRED	68124	22631	15509	42
TACREV	68124	22631	15509	42
TACRED(1%)	682	227		
TACRED(5%)	3407	1133	15509	42
TACRED(10%)	6815	2265		

Table 6: Statistics of datasets

A.2 Analysis of type constrained decoding

In this ablation study, we make comparisons between SURE with and without typed constrained decoding. The results is demonstrated in Tab. 7. Type constraint brings improvement for 0.2% in average in TACRED and has comparable performance on TACREV. Type-relation mapping is inherently involved in training data, so this ablation study proves SURE can learn type-relation mapping from data.

Dataset	TACRED	TACREV
bart-large-cnn	73.6	81.0
- type constraint	73.4	81.0
bart-large-xsum	73.3	81.0
- type constraint	73.1	81.0
pegasus-large	75.1	83.3
- type constraint	75.0	83.3
average gap	-0.2	0

Table 7: Comparison between SURE with and without type constraint decoding. We report micro F1 on TACRED and TACREV.

A.3 Analysis of calibration for NA relation

In this ablation study, we make comparisons between SURE with and without calibration for NA relation. The results is demonstrated in Tab. 8. Calibration brings improvement for 0.3% in average on TACRED and 0.1% in average on TACREV.

A.4 Hyper-parameters and reimplementation

This section details the training and inference processes of baselines and our models. We train and inference all models with PyTorch and Huggingface Transformers on GeForce RTX 2080 or NVIDIA RTX A5000 GPUs. All optimization uses Adam and linear scheduler. A weight decay is used for regularization. We run all experiments on three seeds $[0, 100, 500]$ and report the median. Specifically, the best hyperparameters for full training setting with *pegasus-large* are listed below:

Dataset	TACRED	TACREV
bart-large-cnn	73.6	81.0
- calibration	73.2	80.9
bart-large-xsum	73.3	81.0
- calibration	72.9	80.9
pegasus-large	75.1	83.3
- calibration	74.8	83.2
average gap	-0.3	-0.1

Table 8: Comparison between SURE with and without calibration for NA relation We report micro F1 on TACRED and TACREV.

- learning rate: 1e-4
- weight decay: 5e-6
- epoch number: 20
- max source length: 256
- max target length: 64
- gradient accumulation steps: 2
- warm up steps: 1000

The best hyperparameters for low-resource setting with *pegasus-large* are listed below:

- learning rate: 1e-5
- weight decay: 5e-6
- epoch number: 100
- max source length: 256
- max target length: 64
- gradient accumulation steps: 2
- warm up steps: 0

With the 1/5/10% indices of TACRED provided by (Sainz et al., 2021)[6], we re-implement RECENT (Lyu and Chen, 2021) and test it under both low-resource and full-training scenarios[7]. However, we find the origin evaluation scripts provided by the author has a serious issue which wrongly corrects all false positive samples of the binary classifier as true negative samples. So the recall of NA

[6]Github repository of low-resource indices: https://github.com/osainz59/Ask2Transformers

[7]Github repository of RECENT: https://github.com/Saintfe/RECENT

is always 100% and precision of positive relations is unreasonably high. We correct this issue and the test results significantly differ from origin results reported by previous study. We also test IRE (Zhou and Chen, 2022) and KnowPrompt (Chen et al., 2022) on low resource datasets and search the best hyperparameter with grid searching[8][9]. The previous work of KnowPrompt reports the micro F1 score on SemEval. We train KnowPrompt on SemEval with codes and hyper-parameters provided by authors and re-evaluate it with official macro-F1 scoring method.

A.5 Manual-constructed templates

In this subsection, we display our manual-constructed templates for SemEval (Tab. 14), and three templates designed for TACRED, which are *Semantic1* (Tab. 11), *Semantic2*(Tab. 12), and *Structural*(Tab. 13).

[8]Github repository of IRE: https://github.com/wzhouad/RE_improved_baseline

[9]Github repository of KnowPrompt: https://github.com/zjunlp/KnowPrompt

Technique	Example	M[1]	S[1]	T[1]	TS[1]
Entity information verbalization	Context[2]. ... {subj} ... {obj} ...	✓	✗	✓	✓
Entity marker (Zhang et al., 2019)	... <e1> {subj} </e1> ... <e2> {obj} </e2> ...	✓	✓	✗	✗
Entity typed marker (Zhong and Chen, 2021)	... <e1-{subj-type}> {subj} </e1-{subj-type}> ... <e2-{obj-type}> {obj} </e2-{obj-type}> ...	✓	✓	✓	✗
Entity typed marker (punct) (Zhou and Chen, 2022)	... @ * {subj-type} * {subj} @ ... # ∧ {obj-type} ∧ {obj} # ...	✓	✓	✓	✓
Entity information verbalization + Entity typed marker	Context[2]... @ * {subj-type} * {subj} @ ... # ∧ {obj-type} ∧ {obj} # ...	✓	✓	✓	✓
Entity information verbalization + Entity typed marker (punct)	Context[2]... @ * {subj-type} * {subj} @ ... # ∧ {obj-type} ∧ {obj} # ...	✓	✓	✓	✓

[1] Column names are short for mentions, spans, types and type semantics, respectively.
[2] Augmented context are generated with the template: "*The {subj} entity is {subj} . The {obj} entity is {obj} . The type of {subj} is {subj-type} . The type of {obj} is {obj-type}* "

Table 9: Sentence processing techniques for incorporating entity information.

Technique	TACRED			TACREV		
	P	R	F1	P	R	F1
Entity information verbalization	71.8	75.1	73.4	78.4	83.8	81.0
Entity tag	71.3	71.0	71.1	81.0	75.2	78.0
Entity typed tag	73.5	69.7	71.5	79.9	79.6	79.8
Entity typed tag(punct)	70.3	74.0	72.1	81.1	79.3	80.2
Entity infromation verbalization + Entity tag	73.4	71.8	72.6	78.7	81.7	80.2
Entity infromation verbalization + Entity typed tag	70.6	73.5	72.1	82.8	80.0	**81.4**
Entity infromation verbalization + Entity typed tag(punct)	72.6	74.5	**73.6**	82.1	79.8	81.0

Table 10: Analysis of different input formulation techniques on *bart-large-cnn*. We report micro F1 scores on both datasets. The best F1 score is identified with **bold**.

Relation	Template
org:country_of_headquarters	{subj} has a headquarter in the country {obj}
org:parents	{subj} has the parent company {obj}
per:stateorprovince_of_birth	{subj} was born in the state or province {obj}
per:spouse	{subj} is the spouse of {obj}
per:origin	{subj} has the nationality {obj}
per:date_of_birth	{subj} has birthday on {obj}
per:schools_attended	{subj} studied in {obj}
org:members	{subj} has the member {obj}
org:founded	{subj} was founded in {obj}
per:stateorprovinces_of_residence	{subj} lives in the state or province {obj}
per:date_of_death	{subj} died in the date {obj}
org:shareholders	{subj} has shares hold in {obj}
org:website	{subj} has the website {obj}
org:subsidiaries	{subj} owns {obj}
per:charges	{subj} is convicted of {obj}
org:dissolved	{subj} dissolved in {obj}
org:stateorprovince_of_headquarters	{subj} has a headquarter in the state or province {obj}
per:country_of_birth	{subj} was born in the country {obj}
per:siblings	{subj} is the siblings of {obj}
org:top_members/employees	{subj} has the high level member {obj}
per:cause_of_death	{subj} died because of {obj}
per:alternate_names	{subj} has the alternate name {obj}
org:number_of_employees/members	{subj} has the number of employees {obj}
per:cities_of_residence	{subj} lives in the city {obj}
org:city_of_headquarters	{subj} has a headquarter in the city {obj}
per:children	{subj} is the parent of {obj}
per:employee_of	{subj} is the employee of {obj}
org:political/religious_affiliation	{subj} has political affiliation with {obj}
per:parents	{subj} has the parent {obj}
per:city_of_birth	{subj} was born in the city {obj}
per:age	{subj} has the age {obj}
per:countries_of_residence	{subj} lives in the country {obj}
org:alternate_names	{subj} is also known as {obj}
per:religion	{subj} has the religion {obj}
per:city_of_death	{subj} died in the city {obj}
per:country_of_death	{subj} died in the country {obj}
org:founded_by	{subj} was founded by {obj}"

Table 11: First semantic templates for TACRED, where {subj} and {obj} are the placeholders for subject and object entities.

Relation	Template
no_relation	The relation between {subj} and {obj} is not available
per:stateorprovince_of_death	The relation between {subj} and {obj} is state or province of death
per:title	The relation between {subj} and {obj} is title
org:member_of	The relation between {subj} and {obj} is member of
per:other_family	The relation between {subj} and {obj} is other family
org:country_of_headquarters	The relation between {subj} and {obj} is country of headquarters
org:parents	The relation between {subj} and {obj} is parents of the organization
per:stateorprovince_of_birth	The relation between {subj} and {obj} is state or province of birth
per:spouse	The relation between {subj} and {obj} is spouse
per:origin	The relation between {subj} and {obj} is origin
per:date_of_birth	The relation between {subj} and {obj} is date of birth
per:schools_attended	The relation between {subj} and {obj} is schools attended
org:members	The relation between {subj} and {obj} is members
org:founded	The relation between {subj} and {obj} is founded
per:stateorprovinces_of_residence	The relation between {subj} and {obj} is state or province of residence
per:date_of_death	The relation between {subj} and {obj} is date of death
org:shareholders	The relation between {subj} and {obj} is shareholders
org:website	The relation between {subj} and {obj} is website
org:subsidiaries	The relation between {subj} and {obj} is subsidiaries
per:charges	The relation between {subj} and {obj} is charges
org:dissolved	The relation between {subj} and {obj} is dissolved
org:stateorprovince_of_headquarters	The relation between {subj} and {obj} is state or province of headquarters
per:country_of_birth	The relation between {subj} and {obj} is country of birth
per:siblings	The relation between {subj} and {obj} is siblings
org:top_members/employees	The relation between {subj} and {obj} is top members or employees
per:cause_of_death	The relation between {subj} and {obj} is cause of death
per:alternate_names	The relation between {subj} and {obj} is person alternative names
org:number_of_employees/members	The relation between {subj} and {obj} is number of employees or members
per:cities_of_residence	The relation between {subj} and {obj} is cities of residence
org:city_of_headquarters	The relation between {subj} and {obj} is city of headquarters
per:children	The relation between {subj} and {obj} is children
per:employee_of	The relation between {subj} and {obj} is employee of
org:political/religious_affiliation	The relation between {subj} and {obj} is political and religious affiliation
per:parents	The relation between {subj} and {obj} is parents of the person
per:city_of_birth	The relation between {subj} and {obj} is city of birth
per:age	The relation between {subj} and {obj} is age
per:countries_of_residence	The relation between {subj} and {obj} is countries of residence
org:alternate_names	The relation between {subj} and {obj} is organization alternate names
per:religion	The relation between {subj} and {obj} is religion
per:city_of_death	The relation between {subj} and {obj} is city of death
per:country_of_death	The relation between {subj} and {obj} is country of death
org:founded_by	The relation between {subj} and {obj} is founded by"

Table 12: Second semantic templates for TACRED, where {subj} and {obj} are the placeholders for subject and object entities.

Relation	Template
no_relation	{subj} no relation {obj}
per:stateorprovince_of_death	{subj} person state or province of death {obj}
per:title	{subj} person title {obj}
org:member_of	{subj} organization member of {obj}
per:other_family	{subj} person other family {obj}
org:country_of_headquarters	{subj} organization country of headquarters {obj}
org:parents	{subj} organization parents {obj}
per:stateorprovince_of_birth	{subj} person state or province of birth {obj}
per:spouse	{subj} person spouse {obj}
per:origin	{subj} person origin {obj}
per:date_of_birth	{subj} person date of birth {obj}
per:schools_attended	{subj} person schools attended {obj}
org:members	{subj} organization members {obj}
org:founded	{subj} organization founded {obj}
per:stateorprovinces_of_residence	{subj} person state or provinces of residence {obj}
per:date_of_death	{subj} person date of death {obj}
org:shareholders	{subj} organization shareholders {obj}
org:website	{subj} organization website {obj}
org:subsidiaries	{subj} organization subsidiaries {obj}
per:charges	{subj} person charges {obj}
org:dissolved	{subj} organization dissolved {obj}
org:stateorprovince_of_headquarters	{subj} organization state or province of headquarters {obj}
per:country_of_birth	{subj} person country of birth {obj}
per:siblings	{subj} person siblings {obj}
org:top_members/employees	{subj} organization top members or employees {obj}
per:cause_of_death	{subj} person cause of death {obj}
per:alternate_names	{subj} person alternate names {obj}
org:number_of_employees/members	{subj} organization number of employees or members {obj}
per:cities_of_residence	{subj} person cities of residence {obj}
org:city_of_headquarters	{subj} organization city of headquarters {obj}
per:children	{subj} person children {obj}
per:employee_of	{subj} person employee of {obj}
org:political/religious_affiliation	{subj} organization political or religious affiliation {obj}
per:parents	{subj} person parents {obj}
per:city_of_birth	{subj} person city of birth {obj}
per:age	{subj} person age {obj}
per:countries_of_residence	{subj} person countries of residence {obj}
org:alternate_names	{subj} organization alternate names {obj}
per:religion	{subj} person religion {obj}
per:city_of_death	{subj} person city of death {obj}
per:country_of_death	{subj} person country of death {obj}
org:founded_by	{subj} organization founded by {obj}

Table 13: Structural templates for TACRED, where {subj} and {obj} are the placeholders for subject and object entities.

Relation	Template
Other	{subj} is not related to {obj}
Component-Whole(e1,e2)	{subj} is the component of {obj}
Component-Whole(e2,e1)	{subj} has the component {obj}
Instrument-Agency(e1,e2)	{subj} is the instrument of {obj}
Instrument-Agency(e2,e1)	{subj} has the instrument {obj}
Member-Collection(e1,e2)	{subj} is the member of {obj}
Member-Collection(e2,e1)	{subj} has the member {obj}
Cause-Effect(e1,e2)	{subj} has the effect {obj}
Cause-Effect(e2,e1)	{subj} is the effect of {obj}
Entity-Destination(e1,e2)	{subj} locates in {obj}
Entity-Destination(e2,e1)	{subj} is the destination of {obj}
Content-Container(e1,e2)	{subj} is the content in {obj}
Content-Container(e2,e1)	{subj} contains {obj}
Message-Topic(e1,e2)	{subj} is the message on {obj}
Message-Topic(e2,e1)	{subj} is the topic of {obj}
Product-Producer(e1,e2)	{subj} is the product of {obj}
Product-Producer(e2,e1)	{subj} produces {obj}
Entity-Origin(e1,e2)	{subj} origins from {obj}
Entity-Origin(e2,e1)	{subj} is the origin of {obj}

Table 14: Semantic templates for SemEval, where {subj} and {obj} are the placeholders for subject and object entities.