RTE-GMoE: A Model-agnostic Approach for Relation Triplet Extraction via Graph-based Mixture-of-Expert Mutual Learning

Aziguli Wulamu[1], **Kaiyuan Gong**[1], **Lyu Zhengyu**[1], **Yu Han**[1], **Zhihong Zhu**[2] and **Bowen Xing**[1*]
[1]Beijing Key Laboratory of SMART Traditional Chinese Medicine for Chronic Disease Prevention and Treatment
Beijing Key Laboratory of Knowledge Engineering for Materials Science
School of Computer and Communication Engineering,
University of Science and Technology Beijing
[2]Jarvis Lab, Tencent

Abstract

Relation triplet extraction (RTE) is a fundamental while challenging task in knowledge acquisition, which identifies and extracts all triplets from unstructured text. Despite the recent advancements, the deep integration of the entity-, relation- and triplet-specific information remains a challenge. In this paper, we propose a Graph-based Mixture-of-Experts mutual learning framework for RTE, namely RTE-GMoE, to address this limitation. As a model-agnostic framework, RTE-GMoE distinguishes itself by including and modeling the mutual interactions among three vital task-specific experts: entity expert, RTE expert, and relation expert. RTE expert corresponds to the main RTE task and can be implemented by any model and the other two correspond to the two auxiliary tasks: entity recognition and relation extraction. We construct an expert graph and achieve comprehensive and adaptive graph-based MoE interactions with a novel mutual learning mechanism. In our framework, these experts perform knowledge extractions collaboratively via dynamic information exchange and knowledge sharing. We conduct extensive experiments on four state-of-the-art backbones and evaluate them on several widely-used benchmarks. The results demonstrate that our framework brings consistent and promising improvements on all backbones and benchmarks. Component study and model analysis further verify the effectiveness and advantages of our method.

1 Introduction

Relation Triplet Extraction (RTE) is a fundamental component of knowledge acquisition and information extraction, aiming to extract structured knowledge in the form of triplets (Head-Entity, Relation, Tail-Entity) from unstructured text (Zeng et al., 2018). These triplets provide the building blocks for constructing knowledge graphs, which have been widely used in various applications (Li et al., 2023b; Zhao et al., 2024), including question answering systems, information retrieval, and recommendation engines. For example, given the sentence "*The author of The Three-Body Problem is Cixin Liu.*", an RTE model is expected to extract the triplet *(The Three-Body Problem, author, Cixin Liu)*. This structured representation bridges the gap between unstructured textual data and machine-readable knowledge bases, enabling downstream tasks to perform more effectively (Han et al., 2019).

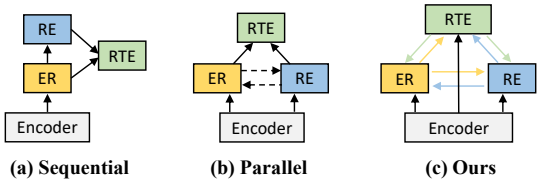

Figure 1: paradigm comparisons between our paradigm - (c) and previous ones - (a) and (b). The dashed arrows in (b) denote some works achieve the interactions.

Despite its utility, RTE is a challenging task due to the usual complex semantics patterns of natural language. It requires models to simultaneously perform entity recognition (ER) and relation extraction (RE), which are closely intertwined. Previous works are generally based on two paradigms, as shown in Fig. 1 (a) and (b). paradigm (a) denotes sequential models, where the RTE process relies on two separate components: ER and RE (Zhang et al., 2022; Ning et al., 2023; Cheng et al., 2025; Wang et al., 2024b). The two sequential tasks operate in the isolation manner and in a predefined order. For example, ER is followed by RE and ER's predictions are directly fed into RE. Although simple to implement, sequential models are highly dependent on the correctness of the first task, making them vulnerable to cascading errors. As a result, the error propagation issue usually leads to suboptimal

*Corresponding author, bwxing714@gmail.com

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7489–7500
November 4-9, 2025

results.

In paradigm (b), ER and RE are performed in a parallel manner. Most work treat them as independent(Han and Liu, 2022; Gao et al., 2024) while some work (Zeng et al., 2018; Luo et al., 2024) employ different interaction mechanism to leverage their correlation. For instance, TP-Linker (Wang et al., 2020; Zeng et al., 2020) proposes a table filling paradigm to integrate ER and RE together, and SINET (Luo et al., 2024) designs a cross-attention mechanism between ER and RE. In this stream of works, after performing ER and RE, their outputs are then somehow merged to produce triplets. While this interaction reduces error propagation to some extent, the coordination remains insufficient due to the lack of direct integration with the RTE process. Except for the correlation between ER and RE, their correlations between the main task RTE are much more beneficial and important, while neglected by previous works. As a result, the full potential of multi-task synergy remains underutilized.

Recalling the example previously mentioned, it is commonly realized that the ER information of `The Three-Body Problem` and `Cixin Liu`, as well as the RE information of `author`, is crucial to the prediction information of the triplet (The Three-Body Problem, author, Cixin Liu). In this paper, we argue that there also exists a potential inverse flow that the RTE information of `(The Three-Body Problem, author, Cixin Liu)` can help derive the extraction of relation `author` and the recognition of entity `The Three-Body Problem` and `Cixin Liu`. The feedback loop among RTE, ER, and RE can form a virtuous cycle. Their mutual promotion is beneficial and urgent to be achieved.

To this end, we propose a new paradigm, as shown in Fig. 1 (c). Inspired from Xing and Tsang (2023a,b), our paradigm performs comprehensive mutual learning between RTE, ER and RE, which are deeply integrated, as indicated by the dense interconnections. This holistic paradigm not only mitigates error propagation but also ensures that all tasks reinforce each other, leading to a significant improvement in the overall performance of the RTE process. In this way, our paradigm is potential to set a new standard for effective and collaborative RTE.

To implement our paradigm and achieve the virtuous cycle among RTE, ER and RE, we propose RTE-GMoE, a novel and model-agnostic framework centered on graph-based mixture-of-experts mutual learning. We propose a task-specific semantics space projection module to process the basic semantic representation obtained from the encoder into three distinct task-specific spaces: entity-specific representation, triplet-specific representation, and relation-specific representation. These three streams of representations are then routed via anchored routing to the respective experts: ER expert, RTE expert, and RE expert, enabling parallel execution within the graph-based MoE framework with mutual learning. Each expert's outputs are subsequently processed by the corresponding decoder through adaptive gating and then produce the final prediction. During training, the three experts are dedicated to ER, RE, and RTE tasks, respectively. In the testing phase, the framework solely executes on the RTE stream. Since our framework is model-agnostic and scalable, the ER, RE and RTE experts can be implemented by any concrete model architecture.

In summary, our contributions are three-folds:

1. We propose a unified model-agnostic RTE framework that deeply integrates ER, RE, and RTE, forming a co-enhancement virtuous cycle.

2. We propose a novel graph-based MoE mutual learning mechanism to dynamically and comprehensively enhance the knowledge and semantics interaction among RTE and its two auxiliary tasks: ER and RE.

3. Extensive experiments on four backbone models and seven benchmark datasets demonstrate the effectiveness and robustness of our method.

2 Related work

2.1 Relation Triplet Extraction

Over the years, a large number of RTE methods have been proposed and this field evolved significantly. Generally, existing RTE methods can be categorized into following two paradigms:

Pipeline models treat entity recognition (ER) and relation extraction (RE) as separate tasks, where outputs from one stage serve as inputs for the next (Miwa and Bansal, 2016; Wu et al., 2019; Zhong and Chen, 2021; Wang et al., 2024a; Chia et al., 2022; Ning et al., 2023; Hennen et al., 2024). Although advancements have been widely achieved by these models, they often suffer from error propagation.

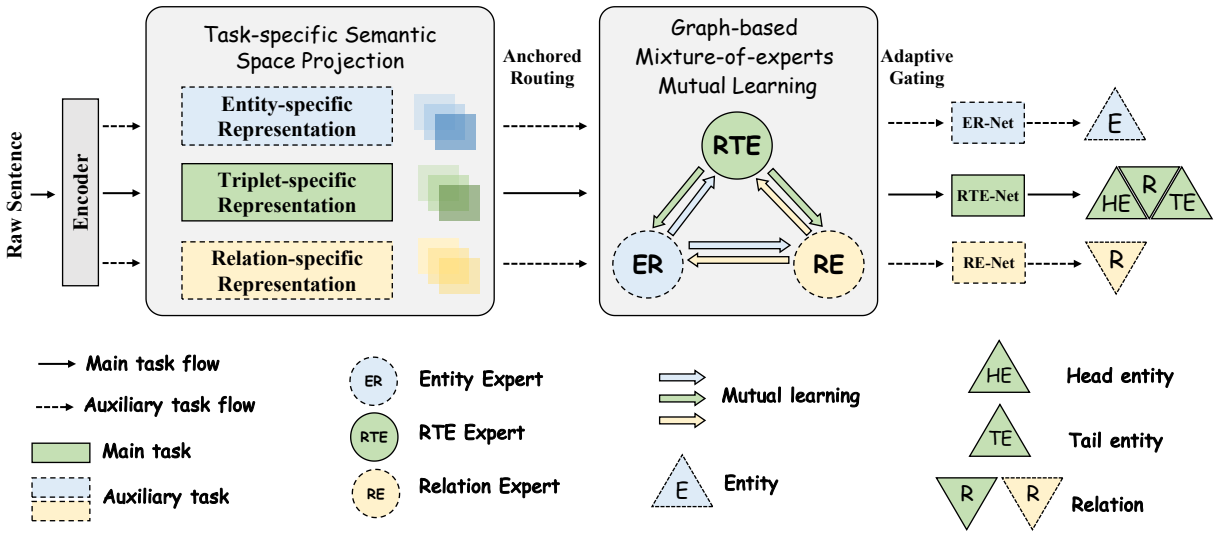

Figure 2: Illustration of our proposed RTE-GMoE framework.

Joint or parallel models integrate ER and RE into a single model to capture interdependencies (Katiyar and Cardie, 2016; Bekoulis et al., 2018; Nguyen and Verspoor, 2019; Zhao et al., 2020; Li et al., 2023a; Zhao et al., 2023; Cheng et al., 2025; Luo et al., 2024). (Fu et al., 2019) utilizes two phases' joint entity and relation extractions with graph convolutional networks. (Nayak and Ng, 2020) proposes two novel encoder-decoder approaches, including a pointer network-based framework, to jointly extract entities and relations. (Yan et al., 2023) introduces HGERE, combining span pruning and hypergraph neural networks to improve entity and relation extraction. (Wang et al., 2024b) proposes a relational triplets extraction method, which uses CRT to accomplish entity recognition and Tucker decomposition to relation extraction in combination.

Existing methods ignore the comprehensive and beneficial interactions among RTE, ER and RE, leaving room for improvement. In this paper, we propose RTE-GMoE as a model-agnostic framework to address this gap by fostering deep integration and interactions among ER, RE, and RTE.

2.2 Mixture of Experts

Mixture of Experts (MoE) (Jacobs et al., 1991) is a dynamic architecture where multiple specialized 'expert' modules are selectively activated for specific tasks. A gating mechanism determines the contribution of each expert, optimizing task-specific learning while managing computational resources efficiently. Sparsely Activated Models (Shazeer et al., 2017) introduces sparse activations to enhance scalability without compromising representational power. Switch Transformers (Fedus et al., 2022) simplifies MoE by activating a single expert per input, reducing computational costs while maintaining performance. In perspective of large language model, GShard (Lepikhin et al., 2021) implements scalable MoE for LLMs, dynamically routing inputs through specialized experts. (Puigcerver et al., 2024) proposes Soft MoE, a fully differentiable sparse Transformer that mitigates traditional MoE challenges, offering scalability and efficiency by performing soft token assignments to experts.

In this paper, we borrow the idea of MoE to establish a graph-based MoE architecture for deeply coupling the knowledge of RTE, ER and RE.

3 Methodology

Problem Definition Given one sentence S including n words $\{s_1,...,s_n\}$, the task of RTE is to extract all relational triplets in the form of (head-entity, relation, tail-entity). In some scenarios, the entity type e for each entity should also be predicted. The entity is a span included in S, while the entity type e and relation type r are from predefined label sets: $\{e_1,...,e_n\}$ and $\{r_1,...,r_n\}$.

3.1 Semantics Encoding

The semantics encoder takes S as input and output the contextualized hidden states $H = \{h_{cls}, h_1, h_2, \ldots, h_n\}$, where $H \in \mathbb{R}^{n \times d}$, d denotes the dimension of hidden state and h_{cls} denotes the hidden state of the special token [CLS].

It can be implemented by any language models, such as LSTM (Hochreiter and Schmidhuber, 1997), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2021), etc..

3.2 Task-specific Semantics Space Projection

In our RTE-GMoE framework, this module serves to generate task-specific representations tailored for ER, RE and RTE. This is achieved by projecting the hidden states obtained from the encoder into task-specific sematnics spaces. Specifically, we apply different linear transformations for different tasks.

Given original hidden states H, we obtain the entity-specific representations R_{ent} via: $R_{\text{ent}} = HW_{\text{ent}} + b_{\text{ent}}$, where $W_{\text{ent}} \in \mathbb{R}^{d\times d}$ and $b_{\text{ent}} \in \mathbb{R}^d$ are trainable parameters. This streams of representations emphasizes ER-specific features in the sentence.

Similarly, we obtain the relation-specific representations R_{relation} via: $R_{\text{rel}} = HW_{\text{rel}} + b_{\text{rel}}$, where $W_{\text{rel}} \in \mathbb{R}^{d\times d}$ and $b_{\text{rel}} \in \mathbb{R}^d$ are trainable parameters. R_{relation} focuses on capturing RE-specific semantics.

As for triplet-specific representation R_{triplet}, we directly assign H to it: $R_{\text{tri}} = H$. It conveys the RTE-specific semantics.

3.3 Graph-based MoE with Mutual Learning

We initialize the structure of graph-based MoE as a fully-connected graph consisting of three nodes representing ER expert, RE expert and RTE expert, corresponding to ER, RE and RTE task, respectively. The initial node representations are R_{ent}, R_{rel} and R_{tri}. Then we use a Graph-Attention-based MoE Mutual Transformation (GAMMT) mechanism to achieve the interactions among the three experts:

$$\widetilde{R_{\text{ent}}}, \widetilde{R_{\text{rel}}}, \widetilde{R_{\text{tri}}} = \text{GAMMT}(\mathcal{G}, [R_{\text{ent}}, R_{\text{rel}}, R_{\text{tri}}], AK, \theta) \tag{1}$$

where $\mathcal{G}$ denotes the MoE graph structure, AK denotes the attention kernel and θ denotes the training parameters of GAMMT.

AK in GAMMT can be implemented by any graph neural networks or graph-based interaction mechanism. In this work, we simply adopt a multi-head graph-attention mechanism for AK, which

Algorithm 1 Training Procedure of RTE-GMoE

Input: Training samples $\mathcal{S}_{train}$, Backbone $\mathcal{M}$, EpochNum, BatchNum
Output: Optimized RTE-GMoE Model $\mathcal{M}_{gmoe}$

```
Define RTE-Net, ER-Net and RE-Net considering M, forming our model M_gmoe.
for e in 1:EpochNum do
    for i in 1:|S_train| do
        L ← 0
        for b in 1:BatchNum do
            H ← Encode(S_i), R_rte ← H
            R_ent ← TSSP_ent(H), R_rel ← TSSP_rel(H)
            R̃_ent, R̃_rel, R̃_tri = GAMMT(R_ent, R_rel, R_tri)
            logit_rte, l_rte ← RTE-Net
            logit_ent, l_ent ← ER-Net
            logit_rel, l_rel ← RE-Net
            L ← L + l_rte + α l_ent + β l_rel
            M_gmoe ← Optimizer(M_gmoe, L)
        end for
    end for
end for
return M_gmoe
```

can be formulated as:

$$\begin{aligned}\tilde{h}_i' &= \|_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i}\gamma_{ij}^k W_h^k \tilde{h}_j\right)\\ \gamma_{ij} &= \frac{\exp\left(\text{ReLU}\left(a^\top\left[W_h\tilde{h}_i \| W_h\tilde{h}_j\right]\right)\right)}{\sum_{j'\in\mathcal{N}_i}\exp\left(\text{ReLU}\left(a^\top\left[W_h\tilde{h}_i \| W_h\tilde{h}_{j'}\right]\right)\right)}\end{aligned} \tag{2}$$

where $W_h \in \mathbb{R}^{d\times d}$ and $a \in \mathbb{R}^{2d}$ are trainable parameter matrices; $\mathcal{N}_i$ denotes the neighbors of node i; γ_{ij} is the normalized attention cores and σ notes the nonlinearity activation function; K is the head number.

Through the above process, the information and knowledge of RTE expert, ER expert and RE expert are comprehensively shared and exchanged. For each expert, it dynamically receives the beneficial knowledge from the other experts in an adaptive gating manner.

3.4 Task-specific Networks

For ER, RE and RTE tasks, task-specific networks are adopted to produce the corresponding predictions with taking $\widetilde{R_{\text{ent}}}$, $\widetilde{R_{\text{rel}}}$, $\widetilde{R_{\text{tri}}}$ as basic semantics representations, respectively. Theoretically, the task-specific networks can be implemented by any concrete models for ER, RE and RTE, respectively. For simplification, in this work, taking an RTE backbone as the RTE-Net, we just adopt the same

architecture of the ER and RE modules in RTE-Net as the ER-Net and RE-Net, respectively. Next, we take SINET (Luo et al., 2024) as the backbone to introduce the task-specific networks. For more details, please refer to the original paper.

ER-Net takes three kinds of features as input: entity-specific sentence representation, the span representation obtained by MaxPooling , and the width embedding got from a trainable matrix. The corresponding entity class distribution for each span is obtained by:

$$y^e = \text{SoftMax}(W_{er}\, e_i + b^{er}) \tag{3}$$

where e_i is the final representation of the $i-$th entity candidate; W_{er} and b^{er} are trainable parameters.

RE-Net takes entity-pair candidate's representation, two entities' representations and the relation-specific sentence representation as input. Then the features are merged into r_c, which is fed into a multi-label relation classier:

$$y^r = \text{Sigmoid}(W_{re}\, r_c + b^{re}) \tag{4}$$

where W_{re} and b^{re} are trainable parameters.

RTE-Net first conducts the cross-task attention mechanism between $\widetilde{R_{\text{ent}}}, \widetilde{R_{\text{rel}}}$, with the consideration of sentence semantics h_{cls}. In this process, the interactions between ER and RE are enhanced. Then the modules of ER-Net and RE-Net are performed to obtain the final predicted relation triplets.

3.5 Training and Inference

The training procedure of RTE-GMoE is illustrated in Algorithm 1. Given the training samples $\mathcal{S}_{train}$ and the backbone $\mathcal{M}$, our goal is to obtain the optimized RTE-GMoE-based model $\mathcal{M}_{gmoe}$. Before training, we first implement the RTE-Net, ER-Net and RE-Net in $\mathcal{M}_{gmoe}$ based on $\mathcal{M}$. For each batch in training procedure, we first obtain the task specific representations via Task-specific Semantics Space Projection (TSSP). Then the graph-based MoE mutual learning mechanism is performed to obtain the task expert representations $\widetilde{R_{\text{ent}}}, \widetilde{R_{\text{rel}}}, \widetilde{R_{\text{tri}}}$, which are used to produce the logit and loss in respective subnetwork. The training object is the weighted sum of ER loss l_{ent} and RE loss l_{rel} as well as the RTE loss l_{rte}. The coefficient α and β balance the impact of ER expert and RE expert in our framework. Finally, the model is updated by the optimizer with the calculated loss.

In the inference stage, only the RTE modules are activated, while the ER- and RE-related modules are only leveraged in the training stage to enhance RTE.

4 Experiments

4.1 Settings

Benchmarks We conduct evaluation experiments on seven widely-used benchmarks: ACE04, ACE05, SciERC, NYT, WebNLG, NYT* and WebNLG*, whose detailed statistics are shown in Table 1.

Dataset	Train	Dev	Test
ACE04	8683(5-fold)		
ACE05	10051	2424	2050
SciERC	1861	275	551
NYT	56195	5000	5000
NYT*	56195	4999	5000
WebNLG*	5019	500	703
WebNLG	5019	500	703

Table 1: Dataset statistics

Evaluation Metrics The evaluation metrics for ACE04, ACE05 and SciERC include Ent, Rel, and Rel+, all in micro-F1 score. On ACE04, the 5-fold cross-validation approach is adopted for performance evaluation. Ent is considered correct when both the entity types and boundaries match those in the ground truth. For relation extraction, Rel is regarded as correct when both the relation types and entity spans align with the ground truth. Therefore, the score of Rel can directly reflect the performance on RTE. The Rel+ evaluation is a more lenient version of Rel, where only the entity types need to be correct, but the entity spans can differ slightly, allowing for more flexibility in the evaluation. As for NYT and WebNLG, we follow the convention to adopt exact match evaluation, where a prediction is considered correct only if it exactly matches the ground truth. Following Dual-Dec and OD-RTE, we also adopt NYT* and WebNLG*, corresponding to partial match evaluation, where a prediction is considered correct if it shares a certain degree of overlap with the ground truth.

Backbones and Baselines We adopt four recently proposed state-of-the-art (SOTA) baselines as the backbones to augment them with our proposed RTE-GMoE framework:

Model	SciERC			ACE04			ACE05		
	Ent	Rel	Rel+	Ent	Rel	Rel+	Ent	Rel	Rel+
SINET (Luo et al., 2024)	69.92	51.99	40.42	**88.26**	62.72	59.47	86.37	64.56	61.44
+ *RTE-GMoE (Ours)*	**71.33**	**54.84**	**43.27**	88.24	**63.13**	**59.62**	**86.52**	**65.07**	**62.02**
Δ	+1.41	+2.85	+2.85	-0.02	+0.41	+0.15	+0.15	+0.51	+0.58

Table 2: Performance comparison based on SINET backbone.

Model	NYT			WebNLG			NYT*			WebNLG*			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg.F1
Dual-Dec (Cheng et al., 2025)	90.05	91.39	90.72	88.54	**88.49**	88.52	89.68	**92.03**	90.84	90.84	92.22	91.53	90.40
+ *RTE-GMoE (Ours)*	**90.66**	**91.76**	**91.21**	**89.57**	87.62	**88.58**	**90.69**	91.71	**91.20**	**91.53**	**92.28**	**91.91**	**90.73**
Δ	+0.61	+0.37	+0.49	+1.03	-0.87	+0.06	+1.01	-0.32	+0.36	+0.69	+0.06	+0.38	+0.33
TLRel (Wang et al., 2024b)	81.28	**89.59**	85.24	**91.10**	90.22	90.66	-	-	-	-	-	-	87.95
+ *RTE-GMoE (Ours)*	**83.70**	88.53	**86.04**	90.48	**91.30**	**90.89**	-	-	-	-	-	-	**88.47**
Δ	+2.42	+1.06	+0.80	-0.62	+1.08	+0.23	-	-	-	-	-	-	+0.52
OD-RTE (Ning et al., 2023)	90.54	92.83	91.67	93.04	**93.21**	93.13	89.74	**92.47**	91.09	92.77	95.00	93.87	92.44
+ *RTE-GMoE (Ours)*	**91.07**	**92.88**	**91.96**	**94.04**	92.40	**93.22**	**91.68**	91.85	**91.76**	**93.55**	**95.45**	**94.49**	**92.86**
Δ	+0.53	+0.05	+0.29	+1.00	-0.81	+0.09	+1.94	-0.62	+0.67	+0.78	+0.45	+0.62	+0.42

Table 3: Performance comparison based on Dual-Dec, TLRel and OD-RTE backbones.

- SINET (Luo et al., 2024) leverages cross-task attention mechanisms to enable synergetic learning between entity recognition and relation extraction tasks. It integrates shared semantic features across tasks to enhance the joint extraction of relational triples.

- Dual-Dec (Cheng et al., 2025) introduces a cascade framework with two decoders: a text-specific relation decoder to detect relations and a relation-corresponded entity decoder to extract associated entities. This design tackles overlapping triples by sequentially decoding relations and entities.

- TLReL (Wang et al., 2024b): incorporates Tucker decomposition to capture correlations among relations via a tensor learning approach. By modeling relation extraction as a tensor learning problem, it achieves robust joint learning for entities and relations.

- OD-RTE (Ning et al., 2023): Inspired by object detection in computer vision, this one-stage model represents relational triple extraction as a bounding-box detection task. It uses vertices-based tagging and auxiliary region detection to improve efficiency and handle complex overlapping triples effectively.

For each above model, it is taken to implement the RTE expert in our RTE-GMoE framework.

We also compare with SOTA in-context-learning (ICL) approaches and supervised fine-tuning (SFT) methods: CodeLlama-34B+C-ICL (Mo et al., 2024), Text-davinci-003+CodeKGC (Bi et al., 2024), ChatGPT+I^{2}CL$_{top-k}$ (Li et al., 2024), Orca-mini-3-7b (SFT) and Vicuna-33b(SFT) (Zhang et al., 2024).

Implementation Details We adopt the uncased base version of BERT (Devlin et al., 2019) as the encoder. For fair comparison, we reproduce the results of all backbones by strictly adhering to the original parameter settings, including the number of epochs, learning rates, batch sizes, and other hyper-parameters as outlined in their papers. We only tune two coefficient α and β balancing the impact of ER expert and RE expert, while keeping all other hyper-parameters consistent with the backbones. α and β are tuned from the range of [0.01, 0.1 , 1]. We conduct all experiments on an NVIDIA A40 GPU.

4.2 Main Results

Experiment results based on the four SOTA backbones are shown in Table 2 and 3. From Table 2, we can observe that RTE-GMoE achieves consistent improvements over SINET backbone across SciERC, ACE04, and ACE05. Specifically, on SciERC, RTE-GMoE improves Rel and Rel+ scores by 2.85% and 2.85%, respectively. Table 3 show that augmented with our framework, OD-RTE, Dual-Dec and TLRel consistently achieve better F1 scores across all datasets. Augmenting the three backbones, RTE-GMoE respectively brings average improvements of 0.42%, 0.33% and 0.52% in

Model	SciERC (SINET)			NYT (TLRel)			WebNLG* (OD-RTE)		
	Ent	Rel	Rel+	Prec.	Rec.	F1	Prec.	Rec.	F1
RTE-GMoE	**71.33**	**54.84**	**43.27**	**83.70**	**88.53**	**86.04**	**93.55**	**95.45**	**94.49**
w/o mutual learning	71.10	53.83	43.18	81.51	90.33	85.69	91.96	95.45	93.67
w/o ER expert	70.31	53.51	42.32	82.68	89.00	85.73	92.99	94.75	93.86
w/o RE expert	70.55	52.55	40.49	80.41	90.42	85.12	92.73	95.19	93.95

Table 4: Ablation Study of RTE-GMoE.

Dataset	Method	F1
SciERC	CodeLlama-34B + C-ICL	17.33
	Text-davinci-003 + CodeKGC	24.00
	RTE-GMOE + SINET	**54.84**
NYT	ChatGPT + I^2CL_{top-k}	35.66
	CodeLlama-34B + C-ICL	60.92
	RTE-GMOE + OD-RTE	**91.96**
WebNLG	Orca-mini-3-7B(SFT)	71.90
	Vicuna-33B(SFT)	72.70
	RTE-GMOE + OD-RTE	**93.22**

Table 5: Comparison with LLM-based SOTAs.

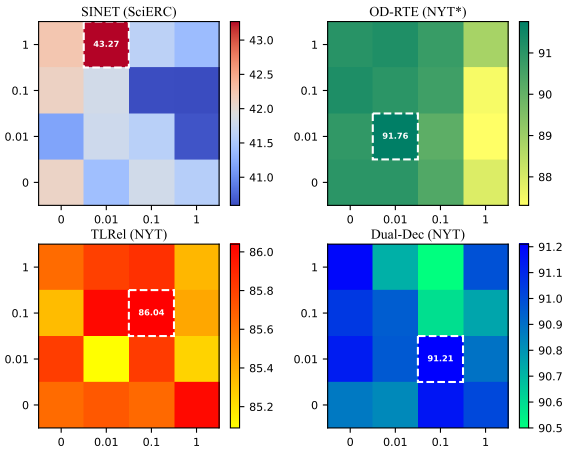

Figure 3: F1 heat-map of four backbones with different values of α (vertical axes) and β (horizontal axes)

F1 score on all datasets. Specifically, RTE-GMoE improves OD-RTE's F1 score by 0.67% and 0.62% on NYT* and WebNLG* datasets, respectively. On Dual-Dec, RTE-GMoE brings 0.49% F1 improvement on NYT dataset. As for TLRel, our RTE-GMoE framework improves it by 0.8% on NYT dataset in F1 score. These results highlight the robustness of our model for both entity and relation extraction tasks.

Besides, we compare our models with LLM-based SOTA models, as shown in Table 5. We can find that our models significantly outperform all of the ICL and SFT methods.

Overall, RTE-GMoE demonstrates consistent and promising performance improvements, achieving higher precision, recall, and F1 scores across various benchmarks. These results validate the effectiveness of our proposed method for all of ER, RE and RTE tasks, proving the ability of our framework in tacking challenging scenarios involving multiple datasets and tasks.

4.3 Ablation study

To verify the effectiveness of each component of our framework, we conduct a group of ablation experiments. The results are shown in Table 4.

Effect of Mutual Learning The graph-based MoE mutual learning mechanism is crucial for enhancing the interaction among ER, RE and RTE. Removing this mechanism leads to a performance drop across all datasets, particularly in RTE. For example, on WebNLG*, the F1 score decreases by 0.82%. This proves the necessity of mutual learning in promoting the comprehensive knowledge interactions among RTE, RE and ER to improve extraction performance.

Effect of ER Expert Excluding the ER expert results in a notable decline in performance. On SciERC, the NER F1 drops 1%, and the triplet F1 decreases from 43.27% to 42.32%. Similar trends are observed on NYT and WebNLG*, with the F1 scores dropping by 0.31% and 0.63%, respectively. This demonstrates the essential role of ER knowledge in promoting RTE.

Effect of RE Expert RE Expert is another vital component in our framework. Removing it causes the most significant drop in triplet extraction performance across all datasets. For instance, on SciERC, the triplet F1 drops to 40.49%, a decrease of 2.78% compared to the full model. Likewise, on NYT and WebNLG*, the F1 scores drop to 85.12% and 93.95%, respectively. These findings underline the importance of RE knowledge in capturing the relational context for precise RTE.

Another interesting observation from the results of SINET is that removing ER expert also harms Rel, and inversely, removing RE expert also leads to performance decrease in Ent. This can be attributed to the fact that our framework achieves the

Unseen Labels	Model	Single Triplet		Multi Triplet					
		Fewrel	Wiki-ZSL	Fewrel			Wiki-ZSL		
		Acc.	Acc.	Prec.	Rec.	F1	Prec.	Rec.	F1
m=5	RSED (Lan et al., 2024)	18.40	22.67	38.14	36.84	37.84	43.91	34.79	38.93
	TAG (Xu et al., 2024)	23.12	28.94	39.36	37.51	38.24	37.56	40.24	38.81
	Re-Cent (Li et al., 2025)	44.97	41.01	46.14	43.53	43.90	52.77	54.7	53.22
	Re-Cent + RTE-GMoE (Ours)	**45.36**	**42.14**	51.17	47.02	**48.21**	49.34	61.29	**54.18**
m=10	RSED (Lan et al., 2024)	22.30	24.91	27.09	39.09	32.00	30.89	29.90	30.39
	TAG (Xu et al., 2024)	17.24	28.16	31.37	32.53	31.88	31.04	33.49	32.18
	Re-Cent (Li et al., 2025)	35.64	29.96	42.85	36.82	38.83	35.62	52.52	42.26
	Re-Cent + RTE-GMoE (Ours)	**37.03**	**30.67**	42.34	38.22	**39.39**	34.93	56.36	**43.07**
m=15	RSED (Lan et al., 2024)	21.64	25.14	25.37	33.80	28.98	27.00	23.55	25.16
	TAG (Xu et al., 2024)	16.41	22.53	26.52	31.34	29.18	25.35	25.88	25.59
	Re-Cent (Li et al., 2025)	30.22	26.91	35.23	30.12	31.56	28.36	52.64	36.48
	Re-Cent + RTE-GMoE (Ours)	**31.49**	**26.95**	35.28	30.95	**32.61**	30.21	54.59	**38.56**

Table 6: Performance comparison on zero-shot RTE.

deeply coupled the ER, RE and RTE, thus their knowledge and semantics information is comprehensively shared and co-promoted. Therefore, moving anyone of them would harm the others' performance.

4.4 Effect of Coefficient α and β

To further study the impact of the ER expert and RE expert, we conduct a group of experiments to analysis the effect of α and β, which determine the degrees of ER and RE in the final loss. The results are shown in Figure 3.

For SINET, its F1 peaks at 43.27% when $\alpha = 1$ and $\beta = 0.01$ on SciERC dataset. As for OD-RTE on NYT* dataset, the F1 score achieves a maximum of 91.76% when $\alpha = 0.01$ and $\beta = 0.01$, while slight variations in α still yield competitive performance.On NYT dataset, the F1 of TLRel reaches its highest score of 86.04% when $\alpha = 0.1$ and $\beta = 0.1$. Its performance is relatively sensitive to α, as reducing α leads to a steady decline in F1, confirming the importance of the ER expert's contribution. On NYT dataset, the optimal F1 of Dual-Dec is 91.21% with $\alpha = 0.01$ and $\beta = 0.1$.

As α or β decreases to 0, the F1 scores significantly drops. While the coefficients increase to 1, the F1 scores do not always achieves the optimal ones. This verifies that the necessity of coefficients α and β in balancing the impact of ER and RE to reach the best performance.

4.5 Experiments on Zero-shot RTE

To verfity the zero-shot generalization ability of our method, we conduct experiments on zero-shot RTE with taking Re-Cent (Li et al., 2025) as the backbone and Fewrel (Han et al., 2018) as well as Wiki-ZSL (Chen and Li, 2021) datasets as the testbeds. Two recent SOTAs RSED (Lan et al., 2024) and TAG (Xu et al., 2024) are taken as baselines. Experiments results are shown in Table 6. We can observe that equipped with our RTE-GMoE framework, Re-Cent can further gain significantly improvments. Specifically, on Fewrel dataset in the m=5 setting, our method achieves 4.3% F1 improvement on Multi Triplet. This can be attributed to the fact that our method focus on the knowledge mixture of RTE and its subtasks, and this ability can be effectively generalized to unseen labels.

4.6 Qualitative Analysis

In Figure 4, we demonstrate some cases to compare the predictions of backbones and the ones augmented with our framework. These cases demonstrate how our models address typical errors in RTE, including entity type errors, relation errors, and entity prediction errors.

In case 1, SINET mis-identifies the entity type of `reconstruction process` as a `Task`, resulting in an Entity Type Error. In contrast, our SINET+RTE-GMoE model correctly identifies the entity type as `Method` and maintains accurate predictions for the relation `Used-for`. This improvement highlights our model's ability to refine entity type predictions through graph-based MoE mutual learning.

In case 2, SINET fails to predict the relation between `detector` and `linear and kernel SVM`, resulting in a Missing Relation Error. Augmented with our framework, SINET+RTE-GMoE correctly predicts the relation `Compare`, demonstrating its capability to capture subtle relation-specific clues.

In case 3, the prediction of Dual-Dec incorrectly associates `Anita Lerman` with `Staten Island`, causing an Entity Error in location identification. Our DualDec+RTE-GMoE model corrects this by

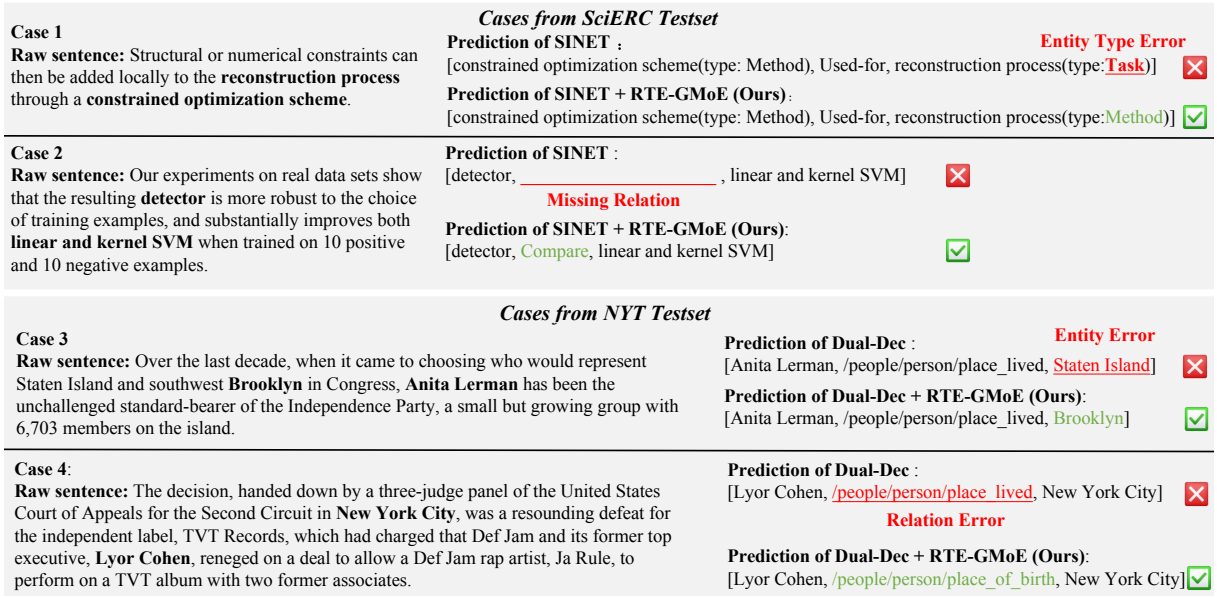

Figure 4: Qualitative analysis on SciERC and NYT datasets.

accurately linking `Anita Lerman` with `Brooklyn`. This showcases the effectiveness of our framework in resolving entity disambiguation issues.

In case 4, Dual-Dec predicts an incorrect relation `/people/person/place_lived` between `Lyor Cohen` and `New York City`, resulting in a Relation Error. Our DualDec+RTE-GMoE predicts right relation triplet by identifying the correct relation as `/people/person/place_of_birth`. This example emphasizes the advantage of graph-based MoE mutual learning in our model to disambiguate complex relational contexts and nuances.

5 Conclusion

This paper presents RTE-GMoE, model-agnostic framework that effectively addresses the challenges in RTE. By incorporating graph-based MoE architecture, mutual learning mechanism, and specialized MoE loss function, RTE-GMoE achieves co-enhancement of entity recognition, relation extraction and relation triplet extraction. The synergistic collaboration between specialized experts and dynamic knowledge sharing through graph-based multi-expert mutual learning significantly improves the model's performance. Extensive experimental evaluations on various backbones and benchmarks show that RTE-GMoE brings consistent and significant improvements, outperforming state-of-the-art models. Further analysis demonstrate the robustness, flexibility, and scalability of RTE-GMoE, positioning it as a promising framework for advancing the field of relation triplet extraction.

Acknowledgment

This work was supported by the National Key Research and Development Project of China (No. 2022YFC3502303), Fundamental Research Funds for the Central Universities (No. FRF-TP-25-035).

Limitation

While our method demonstrates significant improvements in RTE, it still relies on domain-specific correctly labeled samples. Its generalization to unseen or highly complex relational patterns remains a potential limitation which should be tackled via further exploration. Addressing this limitations could enhance the robustness and applicability of the framework.

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.*, 114:34–45.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. Codekgc: Code language model for generative knowledge graph construction. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 23(3):45.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online,*

June 6-11, 2021, pages 3470–3479. Association for Computational Linguistics.

Jian Cheng, Tian Zhang, Shuang Zhang, Huimin Ren, Guo Yu, Xiliang Zhang, Shangce Gao, and Lianbo Ma. 2025. A cascade dual-decoder model for joint entity and relation extraction. *IEEE Trans. Emerg. Top. Comput. Intell.*, 9(2):1130–1142.

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, Dublin, Ireland, May 22-27, 2022*, pages 45–57. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1409–1418. Association for Computational Linguistics.

Chen Gao, Xuan Zhang, Zhi Jin, Weiyi Shang, Yubing Ma, Linyu Li, Zishuo Ding, and Yuqin Liang. 2024. Few-shot relational triple extraction with hierarchical prototype optimization. *Pattern Recognit.*, 156:110779.

Xi Han and Qi-Ming Liu. 2022. Joint extraction of entities and relations by entity role recognition. *Cognitive Robotics*, 2:234–241.

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 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 169–174. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4803–4809. Association for Computational Linguistics.

Moritz Hennen, Florian Babl, and Michaela Geierhos. 2024. ITER: iterative transformer-based entity recognition and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11209–11223. Association for Computational Linguistics.

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

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87.

Arzoo Katiyar and Claire Cardie. 2016. Investigating lstms for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Yuquan Lan, Dongxu Li, Yunqi Zhang, Hui Zhao, and Gang Zhao. 2024. RSED: zero-shot relation triplet extraction via relation selection and entity boundary detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11256–11260. IEEE.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Guozheng Li, Wenjun Ke, Peng Wang, Zijie Xu, Ke Ji, Jiajun Liu, Ziyu Shang, and Qiqing Luo. 2024. Unlocking instructive in-context learning with tabular prompting for relational triple extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 17131–17143. ELRA and ICCL.

Qibin Li, Nianmin Yao, Nai Zhou, Jian Zhao, and Yanan Zhang. 2023a. A joint entity and relation extraction model based on efficient sampling and explicit interaction. *ACM Trans. Intell. Syst. Technol.*, 14(5):77:1–77:18.

Shiyang Li, Yifan Gao, Haoming Jiang, Qingyu Yin, Zheng Li, Xifeng Yan, Chao Zhang, and Bing Yin.

2023b. Graph reasoning for question answering with triplet retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3366–3375. Association for Computational Linguistics.

Zehan Li, Fu Zhang, Kailun Lyu, Jingwei Cheng, and Tianyue Peng. 2025. Re-cent: A relation-centric framework for joint zero-shot relation triplet extraction. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 7344–7354. Association for Computational Linguistics.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized BERT pre-training approach with post-training. In *Chinese Computational Linguistics - 20th China National Conference, CCL 2021, Hohhot, China, August 13-15, 2021, Proceedings*, volume 12869 of *Lecture Notes in Computer Science*, pages 471–484. Springer.

Da Luo, Run Lin, Qiao Liu, Yuxiang Cai, Xueyi Liu, Yanglei Gan, and Rui Hou. 2024. Synergetic interaction network with cross-task attention for joint relational triple extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15457–15468. ELRA and ICCL.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ying Mo, Jiahao Liu, Jian Yang, Qifan Wang, Shun Zhang, Jingang Wang, and Zhoujun Li. 2024. C-ICL: contrastive in-context learning for information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10099–10114. Association for Computational Linguistics.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8528–8535. AAAI Press.

Dat Quoc Nguyen and Karin Verspoor. 2019. End-to-end neural relation extraction using deep biaffine attention. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, volume 11437 of *Lecture Notes in Computer Science*, pages 729–738. Springer.

Jinzhong Ning, Zhihao Yang, Yuanyuan Sun, Zhizheng Wang, and Hongfei Lin. 2023. OD-RTE: A one-stage object detection framework for relational triple extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11120–11135. Association for Computational Linguistics.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. 2024. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jiaxin Wang, Lingling Zhang, Jun Liu, Kunming Ma, Wenjun Wu, Xiang Zhao, Yaqiang Wu, and Yi Huang. 2024a. TGIN: translation-based graph inference network for few-shot relational triplet extraction. *IEEE Trans. Neural Networks Learn. Syst.*, 35(7):9147–9161.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1572–1582. International Committee on Computational Linguistics.

Zhen Wang, Hongyi Nie, Wei Zheng, Yaqing Wang, and Xuelong Li. 2024b. A novel tensor learning model for joint relational triplet extraction. *IEEE Trans. Cybern.*, 54(4):2483–2494.

Shanchan Wu, Kai Fan, and Qiong Zhang. 2019. Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7273–7280. AAAI Press.

Bowen Xing and Ivor W Tsang. 2023a. Co-evolving graph reasoning network for emotion-cause pair extraction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 305–322. Springer.

Bowen Xing and Ivor W Tsang. 2023b. Co-guiding for multi-intent spoken language understanding. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 46(5):2965–2980.

Ting Xu, Haiqin Yang, Fei Zhao, Zhen Wu, and Xinyu Dai. 2024. A two-agent game for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7510–7527. Association for Computational Linguistics.

Zhaohui Yan, Songlin Yang, Wei Liu, and Kewei Tu. 2023. Joint entity and relation extraction with span pruning and hypergraph neural networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7512–7526. 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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9507–9514. AAAI Press.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 506–514. Association for Computational Linguistics.

Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Zijun Min, Qingguo Hu, and Xiaodong Shi. 2022. Towards better document-level relation extraction via iterative inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8306–8317. Association for Computational Linguistics.

Yujia Zhang, Tyler Sadler, Mohammad Reza Taesiri, Wenjie Xu, and Marek Z Reformat. 2024. Fine-tuning language models for triple extraction with data augmentation. In *The First Workshop on Knowledge Graphs and Large Language Models*, page 116.

Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. 2020. Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3948–3954. ijcai.org.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11):293:1–293:39.

Xiaoyan Zhao, Min Yang, Qiang Qu, Ruifeng Xu, and Jieke Li. 2023. Exploring privileged features for relation extraction with contrastive student-teacher learning. *IEEE Trans. Knowl. Data Eng.*, 35(8):7953–7965.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 50–61. Association for Computational Linguistics.