Relation of the Relations: A New Paradigm of the Relation Extraction Problem

Zhijing Jin[*]
AWS Shanghai AI Lab
zhijing.jin@connect.hku.hk

Yongyi Yang[*]
Fudan University
17300240038@fudan.edu.cn

Xipeng Qiu
Fudan University
xpqiu@fudan.edu.cn

Zheng Zhang
AWS Shanghai AI Lab & NYU Shanghai
zhaz@amazon.com

Abstract

In natural language, often multiple entities appear in the same text. However, most previous works in Relation Extraction (RE) limit the scope to identifying the relation between two entities at a time. Such an approach induces a quadratic computation time, and also overlooks the interdependency between multiple relations, namely the relation of relations (RoR). Due to the significance of RoR in existing datasets, we propose a new paradigm of RE that considers as a whole the predictions of all relations in the same context. Accordingly, we develop a data-driven approach that does not require hand-crafted rules but learns by itself the RoR, using Graph Neural Networks and a relation matrix transformer. Experiments show that our model outperforms the state-of-the-art approaches by +1.12% on the ACE05 dataset and +2.55% on SemEval 2018 Task 7.2, which is a substantial improvement on the two competitive benchmarks.[1]

Figure 1: An example of the RE task.

1 Introduction

Relation Extraction (RE) is the task to identify the relation of given entities, based on the text that they appear in. As a fundamental task of knowledge extraction from text, RE has become an active area of research in the past several decades (Miller et al., 1998; Zelenko et al., 2003; Bunescu and Mooney, 2005; Zhou et al., 2005; Hendrickx et al., 2010).

In natural language, most text includes multiple entities. For example, 99.76% data of the widely used ACE 05 dataset (Walker et al., 2006) have more than two entities, and there are 9.21 relations in each text on average. However, most previous research has been confined to the simplified setting of only classifying the relation between every *two* entities at a time (Zeng et al., 2014; Luan et al., 2018;
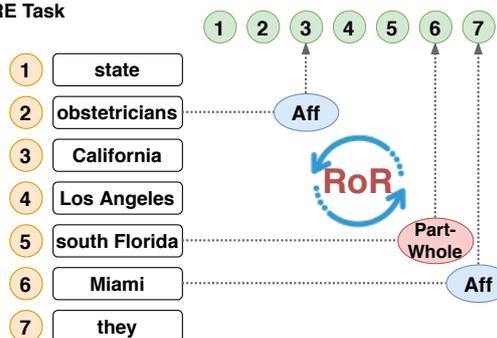
Li and Ji, 2014; Gormley et al., 2015; Miwa and Bansal, 2016). For the sentence in Figure 1 with 7 entities, most previous approaches would perform 49 independent relation classification tasks (if including self-reflexive relations). It is not feasible to reduce this number of classification tasks because existing methods require explicit annotation of the entities in the input. For example, to predict the relation between the entity pair (obstetricians, California), the input needs to be transformed into "...⟨e1⟩ obstetricians ⟨\e1⟩ in ⟨e2⟩ California ⟨\e2⟩ will pay $60,000 in Los Angeles ...".

The problems exposed by such a previous paradigm is that it is not only inefficient, but also overlooks the interdependency among the multiple relations in one context. For example, in the 49 relations in Figure 1, if we already know the relationship (Miami, is part of, south Florida), where

---

"is part of" is a relation defined on two objects, then it is very unlikely for *Miami* to be in any other person-social relationship such as "is the father of...". Remember that, on average, there are 9.21 relations in each text in the ACE 05 dataset, for example, and each relation can provide information to other relations in the same text. We denote the frequently-appearing interdependency of the many relations in the same text as the "relation of relations" (RoR) phenomenon.

To capture RoR, we propose a new paradigm of RE by treating the predictions of all relations in the same text as a whole. Note that our work is distinct from (Wang et al., 2019), which still treats the relation of each entity pair as independent classification tasks, but saves the computation power at the cost of accuracy by encoding all entities in one pass. Instead, our newly proposed paradigm is not about tradeoffs between computation cost and accuracy, but to increase the performance by capturing RoR.

In this paper, we first highlight the importance of RoR by identifying several types and their frequent occurrences in RE datasets in Section 3. We then propose a data-driven approach without hand-crafted rules, using Graph Neural Networks (GNNs) to model each relation as a node to learn the pair-wise dependency of every two relations, and then a matrix transformer to learn the correlations involving multiple relations or numerical correlations of the count of relations. We evaluate the model on two benchmark datasets, ACE 2005 (ACE05) (Walker et al., 2006) and SemEval 2018 Task 7.2 (SemEval2018) (Gábor et al., 2018). Our system outperforms the previous state-of-the-art (SOTA) models by +1.12% on ACE05 and +2.55% on SemEval2018. The contributions of our paper are as follows:

- We propose a new paradigm of RE, according to the frequent RoR phenomenon. It provides a new perspective for future research in RE.

- We develop a model to learn the interdependency of all relations in the same text, based on a GNN and matrix transformer.

- We validate the effectiveness of our model, which outperforms SOTA models by a clear margin on two benchmark datasets.

- We open-source our model and evaluation codes.

2 New formulation of RE

Most previous work formulates RE as multiple independent classification problems limited to two entities and the text:

> Given two entity mentions e_1 and e_2, a text sequence $t = \{w_1, w_2, \ldots, w_N\}$ involving e_1 and e_2, and a finite set relation types $\mathcal{R}$, the task is to predict the relation type between the two entities.

Under this setting, RE can be solved by the well-researched sentence classification task.

However, based on the motivations in Section 1, we propose a new paradigm of RE:

> Given a text sequence $t = \{w_1, w_2, \ldots, w_N\}$ and all the entities $e_1, \ldots, e_M$ mentioned in t, the model needs to predict the relationship r_{ij} between each two entities (e_i, e_j), where $i, j \in \{1, \ldots, M\}$.

We use a matrix $\boldsymbol{R} = (r_{ij}) \in \mathbb{R}^{M \times M}$ to represent all the relations of interest, as shown in Figure 2.
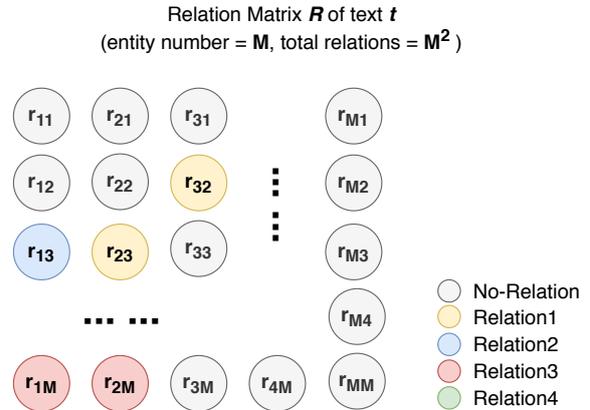

Figure 2: New formulation of RE that considers all M entities in the same text.

3 Statistical analysis of RoR

We conduct our case study using the benchmark dataset ACE05 (Walker et al., 2006). We will introduce two forms of RoR: (1) biRoR, which is only between two relations, and (2) multiRoR, which involves three or more relations.

Note that the purpose of the following analyses is to demonstrate the importance of RoR. Our model introduced in the subsequent section will *not* hand-craft such detailed rules but will learn RoR in a data-driven way.

Relation	Arg0	Arg1
Person-Social (Per-Soc) (e.g., family)	PER	PER
Part-Whole (e.g., city-and-country)	FAC, LOC, GPE, ORG	FAC, LOC, GPE, ORG
Physical (Phys) (e.g., near)	PER, FAC, LOC, GPE	PER, FAC, LOC, GPE
Organization-Affiliation (Org-Aff) (e.g., employment)	PER, ORG, GPE	ORG, GPE
Agent-Artifact (Art) (e.g., owner-and-object)	PER, ORG, GPE	FAC
General-Affiliation (Gen-Aff) (e.g., citizen)	PER	PER, LOC, GPE, ORG

Table 1: Valid entity types for each relation. The entity type abbreviations refer to Person (PER), Facility (FAC), Location (LOC), Geo-PoliticalEntity (GPE), and Organization (ORG).

3.1 Data overview

ACE05 (Walker et al., 2006) is the most widely used dataset for RE. Its text is extracted from a variety of sources, including news programs, newspapers, newswire reports, and audio transcripts. The 6 relation types are shown in Table 1. There are 7 entity types valid for the relations: Facility (FAC), Geo-PoliticalEntity (GPE), Location (LOC), Organization (ORG), Person (PER), Vehicle (VEH), Weapon (WEA).

3.2 BiRoR: Interdependency of two relations

3.2.1 Entity type-constrained biRoR

We first introduce the simplest form of RoR, based on the entity types. Specifically, given the triple (e_1, RelType_a, e_2) of two entities and their relation, we can infer whether e_1 is unlikely to co-occur with a different relation type RelType_b.

To elucidate such constraint, we will make an example using the seven entity types in ACE05. As detailed in Table 1, only certain types are allowed to be the arguments of the relations. Therefore, we can deduct 12 rules of *incompatibility*. For example, the same entity cannot be both the arg0 of Per-Soc and the arg0 of Part-Whole, because an entity must be a person (PER) in order to satisfy the Per-Soc relationship, but Part-Whole *cannot* involve PER. The full list of incompatibility rules are listed in Appendix A.1.

3.2.2 Semantic-constrained biRoR

Another type of biRoR is constrained by the semantics of the relation. The intuition is that what a relation means can imply whether it can be shared or must be disjoint with another relation. For example, the Art relationship can describe a person (arg0) owning a facility (arg1), where the arg1 *must* be a facility. If such relation already exists, the same facility cannot be involved in the Part-Whole relation with a city, because cities (e.g., Boston)

cannot be a part of a facility, semantically. This kind of incompatibility is not a hard constraint by the entity type, but it is implied by the semantics of relations.

Semantics can also imply whether a relation should be symmetric. For example, Per-Soc is always symmetric because family and friends are commutative relations, whereas Org-Aff is always an asymmetric relation (Walker et al., 2005). Hence, if a relation $r_{ij} = $ Per-Soc, then $r_{ji} = $ Per-Soc. And if $r_{ij} = $ Org-Aff, then $r_{ji} \neq $ Org-Aff.

3.2.3 Empirical biRoR

To form a more direct understanding of biRoR, we calculate the correlation of every two relations in Figure 3. We can see that the incompatibility rules in Section 3.2.1 is proven by the red negative color, the symmetric property of Per-Soc and Phys in Section 3.2.2 is proven by the darker blue. There are also other correlations such as frequent
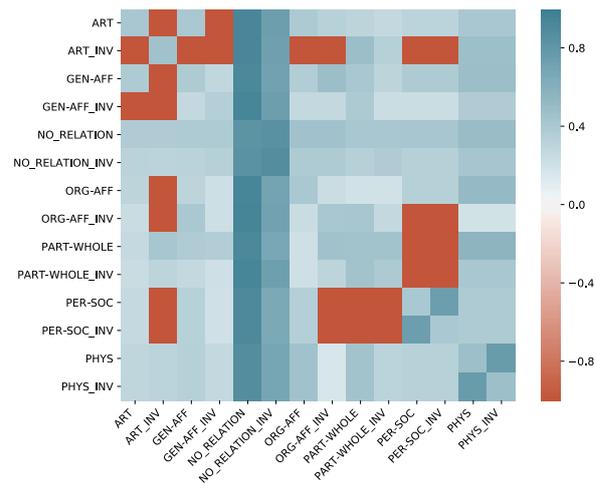

Figure 3: Correlation of every two relations. Each cell $(\text{RelType}_i, \text{RelType}_j)$ represents the conditional probability that an entity will have RelType_j given the existing RelType_i. For better visual effect, the zero probability is converted to -1 (dark red), signaling that these two relations have 0% co-occurrence.

co-occurrences of Part-Whole and Phys relations.

3.3 MultiRoR: Correlation of 3+ relations

3.3.1 Entity type-constrained multiRoR

Other than biRoR, which only involves two relations, there are more complicated rules acting on multiple relations, namely multiRoR. The entity type-constrained multiRoR extends Section 3.2.1 from the incompatibility between two relations to among 3+ relations. For example, if we already know two relations of an entity, which is both the arg1 of the Org-Aff relation and the Phys relation, then this entity cannot be the arg1 of any Art relations (because this entity must be GPE). As the number of relations centered on one entity increases in Figure 4, the percentage of invalid combinations of multiple relations among all possible combinations will soon be over 50%, and even reached 83% when there are 7 relations.
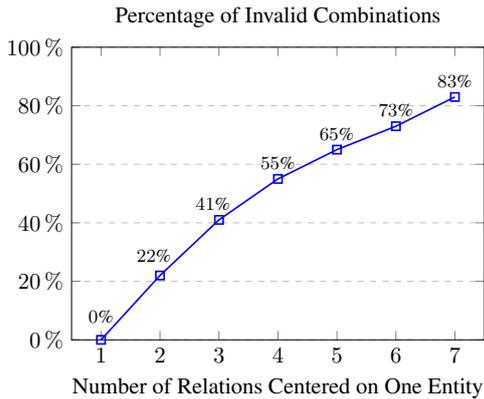
Figure 4: Percentage of invalid combinations over all possible set of relations of an entity, as the total number of non-empty relations centered on the entity increases.

3.3.2 Numerically correlated multiRoR

We can also discover numerical correlations of multiple relations. From all relation matrices R in the dataset, we find that the number of occurrences of a specific type of relations can correlate with the number of another relation type. Note that it is counted as multiRoR, because the correlation is defined not between two single relations, but the total count of two relation types (each of which can include multiple occurrences). From the correlation plot in Figure 5, Per-Soc and Gen-Aff show a strong positive linear dependency, whereas Art and Org-Aff are negatively related by numbers.
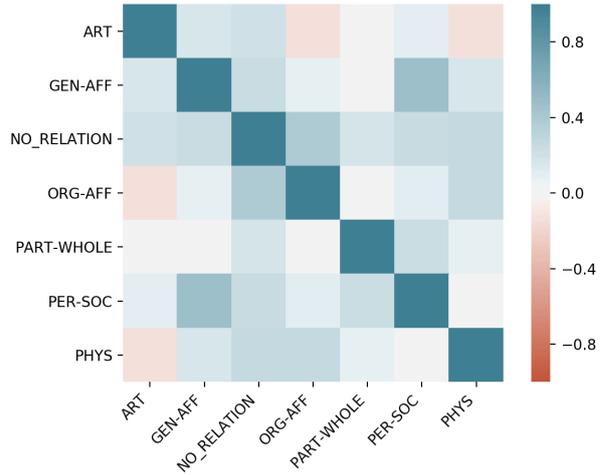
Figure 5: Numerical correlation of every two relation types. Each cell $(\mathrm{RelType}_i, \mathrm{RelType}_j)$ represents the linear correlation between the number of occurrences of $\mathrm{RelType}_i$ and that of $\mathrm{RelType}_j$.

4 Method

Based on the rich RoR phenomenon analyzed in Section 3, we aim to design a model that can mine these properties from data. A naive solution is to hand-craft many rules to impose every type of RoR, but it is not scalable when there are datasets of different features, or when there are some RoR that are difficult to be manually identified. Hence, we aim to design a model that has the capacity to learn RoR with no hand-crafting.

Our overall training strategy is in Figure 6. In the following, we will elaborate on three key components: (1) initial embeddings of entities and relations, (2) the GNN-based biRoR learner, and (3) the matrix transformer that learns multiRoR.

4.1 Initialization of entities and relations

As a preparation step, we first obtain the embedding of each entity. We pass the text through a pretrained BERT model (Devlin et al., 2019), and obtain each entity representation by average pooling over its tokens' hidden states in the last layer of BERT. Note that our framework can easily adapt to other ways to retrieve pretrained embeddings, such as (Yang et al., 2019; Liu et al., 2019).

Then, we obtain the initial embedding of each relation by applying an feedforward layer on the concatenated embeddings of the two involved entities.

4.2 BiRoR Learner

We use a GNN to learn the interdependency between every two relations, namely biRoR. For each
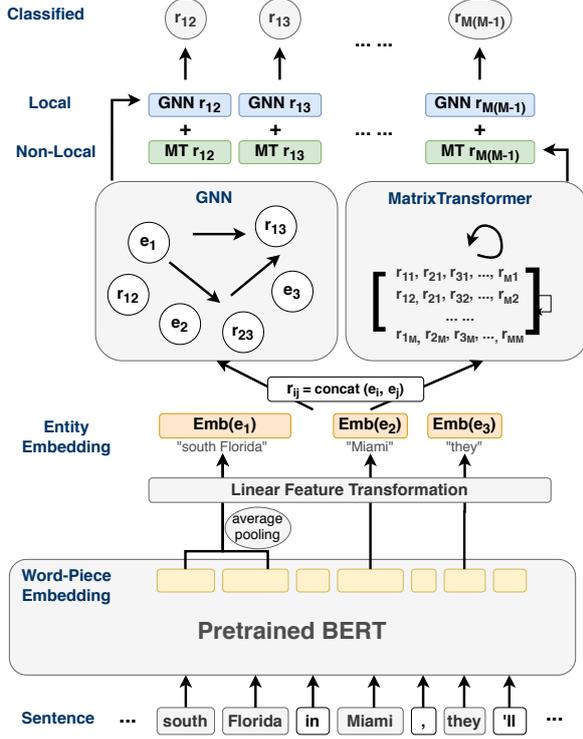
Figure 6: The model architecture of RoR.

text sequence t with M entities, we formulate a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where all entities and relations form the nodes of the graph. Accordingly, we link every relation node to its two entity nodes.

We use a GNN to learn node-to-node interactions, and especially among the relation nodes. The first layer of GNN is initialized with the node embeddings (including all entity embeddings and relation embeddings) obtained from Section 4.1. In each layer ℓ, it aggregates all hidden states passed from the neighbors to update its representation in the next layer $\ell + 1$. More specifically, we have

$$ h_u^{\ell+1} = \text{FFN}\left(W^O \sum_{v \in \mathcal{N}(u)} \alpha_{u,v} h_v^\ell \right), \quad (1) $$

where h_i^ℓ is the hidden state of the i-th vertex in the ℓ-th layer, $\text{FFN}(\cdot)$ is a feed-forward network, W^O is the weight matrix, $\mathcal{N}(u)$ is the set of neighbor nodes to the vertex u, and $\alpha_{u,v}$ is the attention weight that u has to v. This attention weight $\alpha_{u,v}$ is obtained by

$$ \alpha_{u,v} = \frac{\exp\left[W^Q h_v^\ell \left(W^K h_u^\ell \right)^T \right]}{\sum_{v' \in \mathcal{N}(u)} \exp\left[W^Q h_{v'}^\ell \left(W^K h_u^\ell \right)^T \right]}, $$

where W^K and W^Q are the key and query weight matrices when calculating the attention.

4.3 MultiRoR Learner

The GNN introduced in Section 4.2 has a strong ability to model node-to-node interaction, which corresponds to the biRoR, but it is not as strong when seizing the more complicated multiRoR as analyzed in Section 3.3. For example, if an entity has both the relation Org-Aff and Phys, then it cannot have the relation Art. But the GNN structure does not necessarily capture such complicated multiRoR that involve nested conditions.

Therefore, we need another module to directly model the relation matrix $\boldsymbol{R} \in \mathbb{R}^{M \times M}$, which takes into consideration the dynamics among all relations as a whole, in order to capture multiRoR.

To this end, we propose a simple but effective module, a relation matrix transformer. As each relation r_{ij} in the relation matrix $\boldsymbol{R}$ needs to attend to all other relations, we build our relation matrix transformer by customizing the Transformer encoder architecture, which allows extensive mutual attention among all elements (Vaswani et al., 2017). Specifically, we customize the position encoding in the Transformer into two parts: row encoding and column encoding, each of which is a learnable mapping from the position index to a d-dimensional vector space.

Our relation matrix transformer adds the position encoding, namely the sum of the row and column embedding, to the initial representations of relations obtained by procedures in Section 4.1. The input is thus a tensor $\boldsymbol{T} \in \mathbb{R}^{M \times M \times d}$, where d is the dimension of features in the embedding. The matrix transformer then learns the dynamics among all relations as a whole, and outputs new features of all relations by a transformed matrix $\boldsymbol{T}_{\text{Transformed}} \in \mathbb{R}^{M \times M \times d'}$ which captures the multiRoR.

Finally, we add together the relation embeddings learned by the GNN and the matrix transformer, and feed them into the final classification layer to obtain the type of each relation.

5 Experiments

5.1 Datasets

We use two benchmark RE datasets to evaluate the performance of our model.[2]

---

[2]Because our focus is RE, we do not use the more recent RE dataset, DocRED, as it also contains entity linking annotations which can be a multi-tasking objective, and turns the task into a different setting.

ACE05 The ACE 2005 Multilingual Training Corpus (ACE05) (Walker et al., 2006) is a standard RE dataset. It has six relation categories, detailed in Section 3. We process and split the dataset following the practice in (Gormley et al., 2015; Plank and Moschitti, 2013).[3] There are six subdomains in the dataset: Broadcast Conversations (BC), Broadcast News (BN), Conversational Telephone Speech (CTS), Newswire (NW), Usenet Newsgroups (UN), and Weblogs (WL). The statistics of the resulted dataset is in Table 2. Macro F1-scores is the primary evaluation metric of the data, and for more elaborate analysis, we also use macro precision and recall. As "no-relation" is regarded as the negative label, all non-negative relations on wrong entities as treated as false positives.

	Domain	# Relations
Train	NW+BN	34,669
Valid	BC_dev	5,717
Test	BC_test	6,692
	CTS	11,683
	WL	11,022

Table 2: Statistics of ACE05 dataset.

SemEval2018 We use SemEval 2018 Task 7.2 (Gábor et al., 2018) as the second RE dataset to evaluate our model.[4] The corpus is collected from abstracts and introductions of scientific papers, and there are six types of semantic relations in total. Note that there are three subtasks of it: Subtask 1.1 and Subtask 1.2 are relation *classification* on clean and noisy data, respectively; Subtask 2 is the standard relation *extraction*, where the same training set as Subtask 1 is used, but the evaluation is to identify all relations including "no-relation." Following the main systems (Rotsztejn et al., 2018; Nooralahzadeh et al., 2018), we combine the training data of both Subtask 1.1 and 1.2 as the training data. The dataset consists of 136,965 train, and 27,316 test examples. We count the relation of every entity pair as a sample, and if no specific relation is annotated for some entity pairs, we use "no-relation" as the label. The standard evaluation metric is the same as ACE05.

---

[3]https://github.com/mgormley/ace-data-prep
[4]We choose Task 7.2 instead of the Task 7.1 that Wang et al. (2019) use because Task 7.1 only tests the *classification* of positive relations, but Task 7.2 is the standard relation *extraction* which tests on both positive and negative relations.

5.2 Baselines for ACE05

For ACE05, we compare our model with the following systems.

BERT$_{EntEmb}$ We use the baseline from (Wang et al., 2019) (called "BERT$_{sp}$ with entity-indicator on input-layer" in the original paper), which is essentially a BERT with entity embedding on the input layer.

OnePass$_{MRE}$ and OnePass$_{SRE}$ We also compare our model with the state-of-the-art systems on ACE05, OnePass by (Wang et al., 2019). As they found that SRE with OnePass resulted in high scores, we will list the performance of both OnePass$_{MRE}$ and OnePass$_{SRE}$.

Other models We also compare our model with other previous RE models such as (Gormley et al., 2015; Nguyen and Grishman, 2015; Fu et al., 2017; Shi et al., 2018).

5.3 Baselines for SemEval 2018 Task 7.2

For SemEval2018, we compare our results with the top 3 systems on the leaderboard. For a fair comparison, we compare our models against the non-ensemble and ensemble models separately.

Ensemble models The top 1 system (Rotsztejn et al., 2018) is an ensemble of CNNs and RNNs. Its training uses an ensemble size of 20, data augmentation, multi-task learning, and many other meticulous designs. Notably, when implementing our models, we only use a small ensemble size of 5, with no further tricks.

Non-ensemble models The 2nd model (Luan et al., 2018) uses LSTMs to encode both word sequences and dependency tree structures, and perform relation extraction between concepts on top of them. And the 3rd model (Nooralahzadeh et al., 2018) uses a CNN model over the shortest dependency path between two entities.

5.4 Our Models

We investigate the following four settings of our model.

RoR$_{base}$ For the base model in our architecture, we use a customized version of the BERT$_{EntEmb}$ baseline. Instead of the extra embedding for entities in BERT$_{EntEmb}$, we use the sentence index to indicate entities, so that the pretrained properties of BERT can be maintained as much as possible.

	Dev	Test			
	BC_dev	BC_test	CTS	WL	Overall
HybridFCM (Gormley et al., 2015)	–	63.48	56.12	55.17	58.26
DAN (Fu et al., 2017)	–	65.16	55.55	57.19	59.30
GSN (Shi et al., 2018)	–	66.38	57.92	56.84	60.38
BERT$_{EntEmb}$ (Wang et al., 2019)	65.32	66.86	57.65	53.56	59.36
OnePass$_{MRE}$ (Wang et al., 2019)	67.46	69.25	61.7	58.48	63.14
OnePass$_{SRE}$ (Wang et al., 2019)	68.90	**69.76**	63.71	57.20	63.14
RoR$_{base}$	69.95	66.51	61.84	57.67	62.01
RoR$_{bi-only}$	68.50	**68.86**	64.18	59.57	64.20
RoR$_{multi-only}$	**71.64**	66.77	59.90	57.62	61.43
RoR$_{full}$	70.59	68.63	**64.49**	**59.67**	**64.26**

Table 3: Macro F1 on ACE05. Following (Wang et al., 2019), we reported the performances on the development set, and all subsets of the test set. "Overall" is the average of all three subsets in the test set.

We use this model as the base model that we build our biRoR and multiRoR learning modules on.

RoR$_{bi-only}$ and RoR$_{multi-only}$ Based on RoR$_{base}$, we implement RoR$_{bi-only}$, which only the biRoR module based on GNN, and RoR$_{multi-only}$, which uses only the multiRoR module based on a relation matrix transformer.

RoR$_{full}$ Finally, we implement the full model RoR$_{full}$ which adopts both the GNN and the relation matrix transformer to learn all types of RoR.

5.5 Implementation Details

We follow (Wang et al., 2019)'s practice to use BERT-base (uncased) as the pretrained model. The GNN and matrix transformer has 4 layers, 8-head attention, and the hidden size of attention layers is 512. The feedforward layers of GNN have 1024 hidden units, and that of matrix transformer has 4096 hidden units. The batch size is 8. The learning rate is set to 5e-5 on ACE05 and 1e-4 on SemEval2018, and the warmup is 0.1. We use the Adam optimizer and the cosine scheduler of the python package Transformers.[5] We train the model for 30 epochs on both datasets.

6 Results and Analysis

6.1 Main Results

ACE05 We first analyze the experiments on the larger dataset, ACE05. From the main results in Table 3, we can see that in the last column (overall performance), our full model RoR$_{full}$ achieves the strongest performance, outperforming all previous

models. Specifically, our model improves over the strongest setting of OnePass by 1.12%, which is a large margin on the ACE05 dataset. Our full model Moreover, on the subsets of the test set, such as CTS and WL, which is different from the BC domain seen in the development set, our model also demonstrate consistent improvement over the previous best model.

Model	Macro
Non-ensemble models	
(Luan et al., 2018)	39.1
(Nooralahzadeh et al., 2018)	33.6
RoR$_{base}$	38.83
RoR$_{bi-only}$	41.99
RoR$_{multi-only}$	42.87
RoR$_{full}$	**45.46**
Ensemble models	
(Rotsztejn et al., 2018)	49.3
E-RoR$_{base}$	46.47
E-RoR$_{bi-only}$	51.5
E-RoR$_{multi-only}$	**51.85**
E-RoR$_{full}$	51.56

Table 4: Results on SemEval 2018 Task 7.2

SemEval2018 In Table 4, we compare our system with the top 3 systems (Rotsztejn et al., 2018; Luan et al., 2018; Nooralahzadeh et al., 2018) on the SemEval 2018 Task 7.2 Leaderboard. For the non-ensemble setting, our model is +6.26% higher in performance than the most competitive models on the leaderboard. For the more difficult ensemble model leaderboard, we use a simpler setting (i.e., smaller ensemble size, and no data augmenta-

---

[5]https://github.com/huggingface/transformers

tion), but surpasses the top model by +2.55%. The improvement shows that our model with no complicated designs can demonstrate strong performance.

Summary of main experiments Experiments on both datasets validate the effectiveness of our new perspective into RE. Because we model all the relation extraction tasks of the same text together, our method absorbs more knowledge from the same data than other models. Therefore, without data augmentation for SemEval2018, and without domain adaptation on ACE05 (Fu et al., 2017; Shi et al., 2018), our model still surpasses all systems that utilizes extra data.

6.2 Does the model learn multi-relation instances well?

We want to further analyze our model's ability to learn complicated multiRoR relations, especially for some entities with a high number of non-empty relations. For the model outputs on the ACE05 test set, we look into the subset of relations centered on entities with ≥ 2 relations. On this subset, the macro F1 of the previous model OnePass is 61.19%, which drops 2 percent under its overall reported performance. However, our method achieves 63.55%, keeping almost the same performance as the reported 64.26% on the whole test set. On the more challenging subset involving entities with ≥ 3 relations, our model performs 56.63%, almost 4 percent over OnePass's F1 of 52.82%. This shows that the advantage of our model is stronger than others as the relation prediction gets more challenging.

6.3 Does the model learn the conditional probability of relations well?

We also compare our model's ability to learn biRoR versus OnePass. Remember the ground-truth conditional probability of every two relations mentioned before in Figure 3 of Section 3.2.3. For each row of the heatmap, it is normalized to probability 1, so each value in the row forms the probability distribution of any other relation conditioned on the observation of an existing relation. On the overall test set of ACE05, we analyze the output of our model and OnePass to obtain the corresponding heatmap, with each row representing a probability distribution. We calculate the average Jenson-Shannon (JS) distance of all the distributions. We find that OnePass is 0.0805 away from the gold distribution, and we are 0.0326 away from the gold

distribution, -60% nearer than the OnePass, which is the strongest previous model.

7 Related Work

Relation Extraction is one of the most important tasks in NLP. Conventional classification approaches have made use of contextual, lexical and syntactic features combined with richer linguistic and background knowledge such as WordNet and FrameNet (Hendrickx et al., 2010; Rink and Harabagiu, 2010), as well as kernel-based methods (Zelenko et al., 2003; Bunescu and Mooney, 2005; Zhou et al., 2005).

The recent advancement of deep neural networks result in a revolution in the methodology of RE. Many CNN-based (Zeng et al., 2014; dos Santos et al., 2015; Nguyen and Grishman, 2015), and RNN-based (Socher et al., 2012; Zhang and Wang, 2015; Miwa and Bansal, 2016; Zhou et al., 2016) models achieve high performance in many datasets. The popularity of the field also gives birth to many shared tasks (Hendrickx et al., 2010) which turned into the cradle of many competitive, well-designed systems (Luan et al., 2018; Rotsztejn et al., 2018; Jin et al., 2018). Recently, as the model innovation in single relation extraction gradually slow down, most work shifted to the direction of distant supervision (Mintz et al., 2009; Zeng et al., 2015; Lin et al., 2016).

These data-augmenting methods through distant supervision are orthogonal to the innovation in supervised models, and the focus of this paper is to innovate the supervised models to learn the same supervised data with more thorough exploitation. To the best of our knowledge, we are the first paper in MRE to aim at learning the correlation of relations in the training data.

8 Conclusion

In this paper, we proposed a new paradigm of RE, which is capable of modeling the interdependency among multiple relations in the same text. Our model uses a GNN and a matrix transformer to capture the RoR in data. Experiments validated that our model has a substantial improvement on the two benchmark datasets where most models compete for improvement on the decimal point. We also conducted analyses to reflect on our model's performance on learning multi-relation data and the proximity of distributions of correlation of the gold and predicted relations.

Acknowledgments

References

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 724–731. The 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.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 425–429. Asian Federation of Natural Language Processing.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 679–688.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1774–1784.

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, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 33–38.

Di Jin, Franck Dernoncourt, Elena Sergeeva, Matthew B. A. McDermott, and Geeticka Chauhan. 2018. MIT-MEDG at semeval-2018 task 7: Semantic relation classification via convolution neural network. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 798–804.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 402–412.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. 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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. The UWNLP system at semeval-2018 task 7: Neural relation extraction model with selectively incorporated concept embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 788–792.

Scott Miller, Michael Crystal, Heidi Fox, Lance A. Ramshaw, Richard M. Schwartz, Rebecca Stone, and Ralph M. Weischedel. 1998. Algorithms that learn to extract information BBN: TIPSTER phase III. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, MD, USA, October 13-15, 1998*. Morgan Kaufmann.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

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*.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, pages 39–48.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Sirius-ltg-uio at semeval-2018 task 7: Convolutional neural networks with shortest dependency paths for semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 805–810.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1498–1507.

Bryan Rink and Sanda M. Harabagiu. 2010. UTD: classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 256–259.

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. 2018. Eth-ds3lab at semeval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 689–696.

Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 626–634.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1018–1023. Association for Computational Linguistics.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 1201–1211.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. The ace 2005 (ace05) evaluation plan – evaluation of the detection and recognition of ace entities, values, temporal expressions, relations, and events.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting multiple-relations in one-pass with pre-trained transformers. 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 1371–1377.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *CoRR*, abs/1508.01006.

Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 427–434. The Association for Computer Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

A   Appendices

A.1   Full List of Incompatibility Rules

The 12 incompatibility rules of entity type-constrained biRoR introduced in Section 3.2.1 are

1. Part-Whole (arg0) and Per-Soc (arg0)

2. Part-Whole (arg0) and Gen-Aff (arg0)

3. Part-Whole (arg1) and Per-Soc (arg0)

4. Part-Whole (arg1) and Gen-Aff (arg0)

5. Per-Soc (arg0) and Org-Aff (arg1)

6. Per-Soc (arg0) and Art (arg1)

7. Org-Aff (arg0) and Art (arg1)

8. Org-Aff (arg1) and Art (arg1)

9. Org-Aff (arg1) and Gen-Aff (arg0)

10. Art (arg0) and Art (arg1)

11. Art (arg1) and Gen-Aff (arg0)

12. Art (arg1) and Gen-Aff (arg1)