Towards a More Generalized Approach in Open Relation Extraction

Anonymous ACL submission

Abstract

Open Relation Extraction (OpenRE) seeks to identify and extract novel relational facts between named entities from unlabeled data without pre-defined relation schemas. Traditional OpenRE methods typically assume that the unlabeled data consists solely of novel relations or is pre-divided into known and novel instances. However, in real-world scenarios, novel relations are arbitrarily distributed. In this paper, we propose a generalized OpenRE setting that considers unlabeled data as a mixture of both known and novel instances. To address this, we propose MixORE, a two-phase framework that integrates relation classification and clustering to jointly learn known and novel relations. Experiments on three benchmark datasets demonstrate that MixORE consistently outperforms competitive baselines in known relation classification and novel relation clustering. Our findings contribute to the advancement of generalized OpenRE research and real-world applications.

1 Introduction

Open Relation Extraction (OpenRE) is a fundamental task in Information Extraction (IE) that aims to identify and extract relational facts between named entities from unlabeled data. Unlike traditional Relation Extraction (RE), which relies on a predefined set of relations and requires end-users to specify their information needs and provide costly annotations, OpenRE operates in a more flexible "open-world" setting. It proactively discovers novel relations, generalizes them into meaningful categories, and identifies additional instances, making it a more adaptable approach for large-scale IE.

In recent years, OpenRE has attracted increasing attention from researchers. Wang et al. (2022) and Li et al. (2022) introduce prompt-based learning methods and advanced clustering techniques, achieving impressive results on unlabeled data. However, existing OpenRE methods typically assume either that the unlabeled data consists entirely of novel relations or that there is prior information indicating whether an instance belongs to a known or novel relation. These assumptions do not accurately reflect the complexities of real-world scenarios.

Hogan et al. (2023) further dispose of the simplifying assumptions and make new assumptions that the unlabeled data includes known and novel instances and that novel relations are typically rare and belong to the long-tail distribution. However, the "long-tail" assumption may not always hold, particularly in scenarios where novel relations emerge as newly-recognized concepts in the real world that have not yet been labeled. Additionally, novel relations may arise when human annotators label only some relations within a large dataset, leaving many potential relations unidentified. Furthermore, we observe that a noticeable performance gap still exists between known and novel instances (Hogan et al., 2023), highlighting the potential for further OpenRE research.

In this paper, we relax the "long-tail" assumption and instead assume the unlabeled data contains both known and novel instances, with no restrictions on the nature of these relations. We propose MixORE model to effectively classify known instances and identify novel relations within unlabeled data. MixORE has two phases: novel relation detection and open-world semi-supervised joint learning (OW-SS joint learning).

In the first phase, our goal is to identify potential novel relations within unlabeled data. We represent each known relation with a one-hot vector in latent space and train a Semantic Autoencoder (SAE) (Kodirov et al., 2017) on labeled data. The trained SAE then maps both labeled and unlabeled instances into the shared latent space, where known instances will cluster around their respective one-hot vectors. In contrast, novel instances, which are

less likely to align with any known relations, tend to appear as outliers in this mapping process. Furthermore, instances in the same novel relation often exhibit a clustering pattern. Therefore, we leverage each unlabeled instance's similarity to the known relation one-hot vectors as a criterion for outlier detection. Subsequently, we apply the Gaussian Mixture Model (GMM) (Pedregosa et al., 2011) to cluster these outliers into novel relation groups and extract instances closest to each cluster centroid as high-quality weak labels for further training.

In the second phase, OW-SS joint learning, we utilize weak labels and adopt a continual learning strategy to align our approach with the evolving nature of OpenRE in real-world applications. MixORE is designed based on the insight that classifying known relations requires learning compact and well-separated feature representations, whereas detecting novel relations benefits from capturing diverse and transferable features. To achieve this, we incorporate contrastive learning by leveraging both labeled instances and data distribution to form positive pairs and propose the OW-SS loss function, which jointly optimizes relation classification and clustering.

In summary, our main contributions are:

- We comprehensively review the assumptions made in previous OpenRE studies and introduce a generalized OpenRE setting.
- We propose a two-phase framework MixORE that learns discriminative features for known relations while continuously incorporating novel information from unlabeled data, making it more adaptable to OpenRE in real-world scenarios.
- Experimental results demonstrate that our approach achieves remarkable performance on both known and novel relations across the FewRel, TACRED, and Re-TACRED datasets.

2 Related Work

Relation Extraction (RE) is an essential Natural Language Processing (NLP) task and has been extensively studied with approaches relying on supervised learning techniques trained on manually annotated datasets (Miller et al., 1998; Zelenko et al., 2002; Peng et al., 2017; Zhong and Chen, 2021; Wadhwa et al., 2023). While RE models achieve high performance, their dependency on large-scale labeled data presents a major limitation. Moreover, they operate under a "closed-world" assumption, where relations are pre-defined, limiting their ability to handle emerging or novel relations. To address these challenges, OpenRE is proposed to proactively identify novel relations from unlabeled data in an "open-world" setting, making it more suitable for real-world, large-scale information extraction.

Existing OpenRE methods mostly operate under two settings. The first setting is unsupervised relation extraction (URE), where models identify relations between named entities from unlabeled data without relying on manual annotations. Liu et al. (2022) propose a hierarchical exemplar contrastive learning framework that refines relation representations by leveraging both instance-level and exemplar-level signals for optimization. Wang et al. (2023) strengthen the discriminative power of contrastive learning with both within-sentence pairs augmentation and augmentation through cross-sentence pairs extraction to increase the diversity of positive pairs.

The second setting of existing methods, semi-supervised OpenRE, involves training models on labeled data with known relations, while the unlabeled data consists entirely of novel relations or is pre-divided into known and novel sets. Wang et al. (2022) develop a novel prompt-based framework that enables the model to generate efficient representations for instances in the open domain and learn clustering novel relational instances. Li et al. (2022) design a co-training framework that combines the advantage of type abstraction and the conventional token-based representation.

There are some recent studies trying to address the open-world semi-supervised learning (Open-world SSL) setting, where unlabeled data contains a mixture of both known and novel classes. Cao et al. (2022) propose ORCA for computer vision tasks. This method introduces uncertainty adaptive margin loss objective to either classify unlabeled image instances into one of the known classes or discover novel classes and assign instances to them. Hogan et al. (2023) later introduce KNoRD for open-world relation extraction. With prompt-based training, KNoRD effectively classifies explicitly and implicitly expressed relations from known and novel relations within unlabeled data. However, the authors assume novel relations are typically rare and belong to the long-tail distribution. In this study, we relax this assumption and instead assume

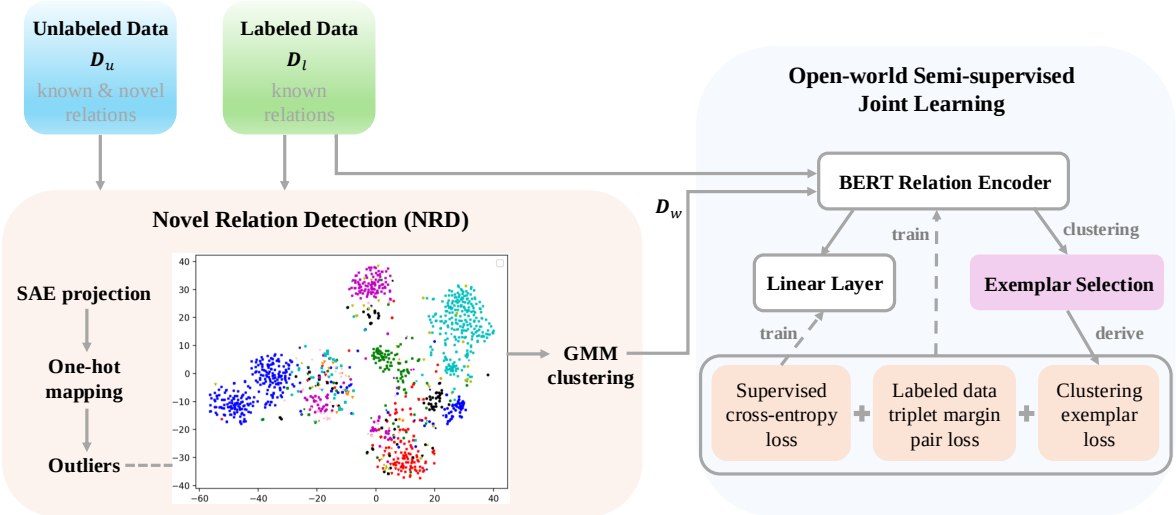

Figure 1: Overview of MixORE Framework.

the unlabeled data comprises both known and novel instances, where the known and novel relations can be arbitrary.

3 Task Formulation

We formalize the OpenRE task as follows. Let $\boldsymbol{x} = [x_1, ..., x_n]$ denote a sentence, where x_i represents the i-th $(1 \leq i \leq n)$ token. In the sentence, a named entity pair $(\boldsymbol{e}_h, \boldsymbol{e}_t)$ is recognized in advance, where $\boldsymbol{e}_h$ represents the head entity and $\boldsymbol{e}_t$ represents the tail entity. Let $\mathcal{D}_l = [(\boldsymbol{x}^1, \boldsymbol{y}^1), ..., (\boldsymbol{x}^M, \boldsymbol{y}^M)]$ be the labeled data consists of M instances with the corresponding sentences, the target entity pairs, and relation labels. Let $\mathcal{D}_u = [(\boldsymbol{x}^1), ..., (\boldsymbol{x}^N)]$ be the unlabeled data consists of N instances with only corresponding sentences and target entity pairs. We denote the set of relations in the labeled data as $\mathcal{C}_{known}$ and the set of relations in the unlabeled test data as $\mathcal{C}_u$. Following Cao et al. (2022), we assume category/class shift $\mathcal{C}_{known} \subseteq \mathcal{C}_u$ (i.e., the relations encountered at test time may not have been explicitly labeled or seen during training). We define the set of novel relations $\mathcal{C}_{novel} = \mathcal{C}_u - \mathcal{C}_{known}$.

The goal of OpenRE is to assign known instances in $\mathcal{D}_u$ to their respective known relations $\mathcal{C}_{known}$, while also identifying $|\mathcal{C}_{novel}|$ novel relation clusters, where $|\mathcal{C}_{novel}|$ represents the number of novel relations in the corpus.

4 Methodology

In this section, we introduce the proposed MixORE, a two-phase framework that integrates relation classification and clustering to jointly learn known and novel relations. Our methodology incorporates novel relation detection for obtaining weak labels and open-world semi-supervised joint learning (OW-SS joint learning) to progressively refine the model. Figure 1 provides an overview of the framework.

4.1 Relation Encoder

Given a sentence along with its named entities and entity types, the relation encoder generates a vector representation that captures the relationship between the entities. To highlight the entities of interest, we adopt entity marker tokens, a widely used technique in relation extraction models (Soares et al., 2019; Xiao et al., 2020; Liu et al., 2022; Wang et al., 2023).

Specifically, for a given sentence $\boldsymbol{x} = [x_1, ..., \boldsymbol{e}_h, ..., \boldsymbol{e}_t, ..., x_n]$, we insert <$e1$:type> and </$e1$:type> to denote the beginning and end of the head entity $\boldsymbol{e}_h$, and similarly, <$e2$:type> and </$e2$:type> for the tail entity $\boldsymbol{e}_t$, where "type" is replaced with the actual entity type. We use BERT$_{base}$ model (Devlin et al., 2019) to obtain the contextualized sentence representation $\boldsymbol{h}$. To effectively capture relational context and enhance focus on the target entity pair, we derive the following fixed-length relation representation:

$$\boldsymbol{h}_r = [h_{<e1:type>} | h_{<e2:type>}] \quad (1)$$

to express the relation between the marked entities in $\boldsymbol{x}$, where $|$ denotes the concatenation.

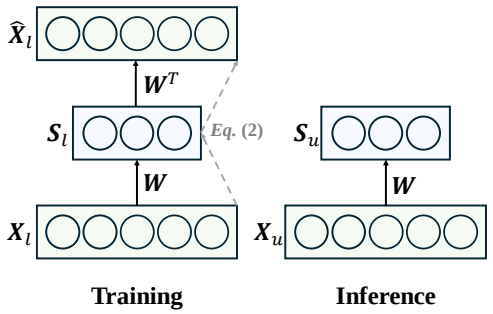

Figure 2: Semantic Autoencoder (SAE)

4.2 Novel Relation Detection

In the first phase of MixORE, our objective is to identify potential novel relations within unlabeled data. These novel relations, once detected, can be leveraged as weak labels to enhance the training process, particularly in real-world scenarios where labeled data is scarce or unavailable.

Our novel relation detection approach is founded on the assumption that known instances naturally cluster around their respective relation centroids, forming well-defined groups. In contrast, novel instances, which do not correspond to any known relations, are likely to appear as outliers. However, in practice, the lack of labeled data for novel relations results in ambiguous feature representations, making it challenging to differentiate between known and novel relations. Additionally, clustering algorithms such as K-Means and Gaussian Mixture Models (GMM) often struggle with high-dimensional feature spaces, further complicating the task of accurately grouping novel instances.

To effectively learn a low-dimensional projection function that generalizes well to both known and novel relations, we employ the encoder-decoder paradigm. In this approach, the encoder maps a feature vector into an intermediate low-dimensional space, while the decoder imposes an additional constraint by ensuring that the projected representation can accurately reconstruct the original feature vector. Specifically, we adopt the Semantic Autoencoder (SAE) (Kodirov et al., 2017), a simple and extremely efficient architecture, as illustrated in Figure 2.

The labeled data $\boldsymbol{\mathcal{D}}_l$ is first processed by the relation encoder (defined in Sec. 4.1), generating the input data matrix $\boldsymbol{X}_l = [(\boldsymbol{h}_r^1), ..., (\boldsymbol{h}_r^M)]$. It is projected into a latent space of $|\boldsymbol{\mathcal{C}}_{known}|$ dimensional with a projection matrix $\boldsymbol{W}$. The latent space is constrained to serve as a semantic representation space. To enforce independence among relations, we incorporate one-hot vectors to encode known relations and obtain the latent representation $\boldsymbol{S}_l$. To further simplify the model, we use tied weights, that is, the transposed projection matrix $\boldsymbol{W}^T$ projects the latent representation $\boldsymbol{S}_l$ back to the feature space, and becomes $\hat{\boldsymbol{X}}_l$. The learning objective is as follows:

$$\min_{\mathbf{W}} \|\mathbf{X}_l - \mathbf{W}^\top \mathbf{S}_l\|_F^2 + \lambda \|\mathbf{W}\mathbf{X}_l - \mathbf{S}_l\|_F^2, \quad (2)$$

where λ is a weighting coefficient that balances the contributions of the first and second terms, corresponding to the losses of the decoder and encoder, respectively. Following Kodirov et al. (2017), we efficiently derive the optimal solution for $\boldsymbol{W}$ with Bartels-Stewart algorithm (Bartels and Stewart, 1972).

During inference, we input all unlabeled data $\boldsymbol{\mathcal{D}}_u$ into the relation encoder, and subsequently pass the resulting relation representations $\boldsymbol{X}_u$ through the encoder of the SAE to obtain the latent representation $\boldsymbol{S}_u$. For each vector v in $\boldsymbol{S}_u$, we calculate its cosine similarity with each known relation one-hot vector and record the highest similarity score as its mapping score. This process assigns each instance in $\boldsymbol{\mathcal{D}}_u$ to the most probable known relation. We designate the 5% of unlabeled instances with the lowest mapping scores as outliers.

Instances belonging to the same novel relation also tend to cluster together. We subsequently employ the Gaussian Mixture Model (GMM) (Pedregosa et al., 2011) to cluster these outliers into $|\boldsymbol{\mathcal{C}}_{novel}|$ novel relation clusters. GMM assumes that the data points are generated from a mixture of several Gaussian distributions, each representing a cluster. The model defines the probability density function (PDF) of the data as:

$$p(v|\Theta) = \sum_{i=1}^{|\boldsymbol{\mathcal{C}}_{novel}|} \pi_i \mathcal{N}(v|\mu_i, \Sigma_i), \quad (3)$$

where $p(v|\Theta)$ is the likelihood of observing data point v, π_i is the mixture weight of the i-th Gaussian component, and $\mathcal{N}(v|\mu_i, \Sigma_i)$ represents the multivariate Gaussian distribution with mean μ_i and covariance matrix Σ_i.

To extract high-quality weak labels for subsequent training, we select instances closest to each cluster centroid. Specifically, we retain instances with a GMM posterior probability greater than 0.95, ensuring that only those with high confidence in

their cluster assignments are used as weak labels. The resulting set of weakly-labeled instances is denoted as $\mathcal{D}_w$.

4.3 Open-world Semi-supervised Joint Learning

The ultimate goal of the Open-world SSL setting is to adaptively expand the model's understanding of novel relations while preserving high performance on known relations. To effectively handle both known and novel relations, we employ a continual learning (Wang et al., 2024) strategy. In the OW-SS joint learning phase, the proposed MixORE model is first warmed up by training on $\mathcal{D}_l$, which consists of known relations only. Following the rehearsal-based strategy in Continual Relation Extraction (Cui et al., 2021; Wu et al., 2024), MixORE is continually trained on both labeled known instances $\mathcal{D}_l$ and the weakly-labeled novel instances $\mathcal{D}_w$.

First, for relation classification, we employ the following cross-entropy loss function:

$$\mathcal{L}_c = -\frac{1}{D_c}\sum_{i=1}^{D_c}\sum_{r=1}^{|\mathcal{C}_u|} y_r^i log(\hat{y_r^i}), \tag{4}$$

where y_r^i is 1 if sample i belongs to relation r otherwise 0, $\hat{y_r^i}$ is the predicted probability of sample i belongs to relation r, $|\mathcal{C}_u| = |\mathcal{C}_{known}| + |\mathcal{C}_{novel}|$ is the number of relations in the unlabeled data, and D_c represents the number of labeled instances in the current epoch.

Zhang et al. (2022) demonstrate that, in computer vision tasks, discriminative features are preferred for classifying known classes, whereas rich and diverse features are essential for identifying novel classes. Such findings should also apply to OpenRE tasks, as classifying known relations requires learning compact and well-separated feature representations, while detecting novel relations benefits from capturing diverse and transferable features that generalize beyond the labeled data.

Contrastive Learning, a strategy widely adopted by state-of-the-art RE models (Liu et al., 2022; Wang et al., 2023; Wu et al., 2024), enhances relation representations by pulling semantically similar relation sentences (positive pairs) closer while pushing apart sentences with different relations (negative pairs). We integrate contrastive learning using two strategies to form positive pairs: sampling from labeled instances and leveraging the data distribution. This approach enables us to jointly capture classification signals and the underlying data distribution, leading to more robust relation representations.

We begin by utilizing labeled data to construct positive pairs. Since weak labels can be noisy, to minimize the risk of introducing false positive pairs from $\mathcal{D}_w$, we restrict the generation of positive pairs to $\mathcal{D}_l$. Specifically, we sample instances from $\mathcal{D}_l$ such that two instances sharing the same relation form a positive pair and ensure that each relation has an equal number of positive pairs sampled, except in cases where there are insufficient instances to enumerate. Let $\mathcal{P} = [(\boldsymbol{a_r}^1, \boldsymbol{p_r}^1), ..., (\boldsymbol{a_r}^{D_m}, \boldsymbol{p_r}^{D_m})]$ denote the set of relation representations of the sampled D_m positive pairs. In this work, we fix the number of sampled positive pairs to $D_m = 5D_c$. As noted by Wang et al. (2023), relation semantics between two sentences should not be treated as a strict "same/different" distinction but rather as a similarity spectrum. To handle this, the triplet margin loss function for the labeled data positive pairs is defined as:

$$\begin{aligned}\mathcal{L}_{lm} = \frac{1}{D_m}\sum_{i=1}^{D_m}\max\{&dist(\boldsymbol{a}_r^i, \boldsymbol{p}_r^i) \\ &- dist(\boldsymbol{a}_r^i, \boldsymbol{n}_r^i) + \gamma, 0\},\end{aligned} \tag{5}$$

where $dist(\cdot)$ denotes the cosine distance function, $\boldsymbol{n}_r^i$ represents a randomly sampled negative example for $\boldsymbol{a}_r^i$, and γ, known as the margin, is a hyperparameter.

To further incorporate data distribution, we encourage relation representations to align more closely with their respective cluster centroids while pushing them away from other clusters. Each instance and its corresponding virtual centroid are treated as a positive pair, reinforcing the cluster structure in the representation space. Following Liu et al. (2022), we select relational exemplars at multiple granularities by computing cluster centroids for different values of k using K-Means algorithm. These exemplars dynamically adjust in response to parameter updates in the relation encoder during each training epoch. Since an instance either belongs to a cluster or not, we use the following clustering exemplar loss function:

$$\mathcal{L}_e = -\sum_{i=1}^{D_c}\frac{1}{L}\sum_{l=1}^{L} log\frac{exp(\boldsymbol{h}_r^i \cdot \boldsymbol{e}_j^l/\tau)}{\sum_{q=1}^{c_l} exp(\boldsymbol{h}_r^i \cdot \boldsymbol{e}_q^l/\tau)}, \tag{6}$$

where $j \in [1, c_l]$ represents the j-th cluster at granularity layer l, $\boldsymbol{e}_j^l$ is relation representation of the exemplar of instance i at layer l, and τ is a is a temperature hyperparameter (Wu et al., 2018).

Our overall OW-SS loss function is defined as the addition of classification loss $\mathcal{L}_c$, labeled data triplet margin loss $\mathcal{L}_{lm}$, and clustering exemplar loss $\mathcal{L}_e$:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_{lm} + \mathcal{L}_e, \tag{7}$$

which jointly optimizes relation classification and clustering. The model architecture consists of a BERT encoder followed by a fully connected linear layer. The BERT encoder is fine-tuned using Eq. 7, while the linear layer parameters are updated based on Eq. 4.

4.4 Inference

During inference, each instance is encoded using the trained model to obtain its relation representation and predicted label. If the predicted label corresponds to a known relation, it is directly accepted as the final result. However, since novel relations are trained with weak labels, their predicted labels may not be accurate. Therefore, we leverage relation representations for novel relations instead. Specifically, employ Faiss K-Means clustering algorithm (Johnson et al., 2021), an efficient implementation of K-Means optimized for large-scale and high-dimensional data, to cluster these relation representations and assign relations based on the clustering results.

4.5 Data Augmentation

Several studies have demonstrated that data augmentation can significantly enhance the performance of RE models (Liu et al., 2021, 2022; Wang et al., 2023). In this work, we apply the data augmentation technique proposed by Wang et al. (2023), which leverages within-sentence pairs augmentation and augmentation through cross-sentence pairs extraction to increase the diversity of positive pairs.

5 Experiments

5.1 Datasets

Following Hogan et al. (2023), we adopt FewRel (Han et al., 2018), TACRED (Zhang et al., 2017), and Re-TACRED (Stoica et al., 2021) datasets to train and evaluate our model. To simulate the OpenRE task in real-world scenarios, we assign $|\mathcal{C}_{novel}| = 6$ relations as novel relations for each dataset, and the remaining relations are considered as known relations. For each known relation, we allocate half of its instances to the labeled dataset $\mathcal{D}_l$. The unlabeled dataset $\mathcal{D}_u$ consists of the remaining half known instances along with all instances from novel relations.

FewRel dataset includes additional relation hierarchies. To challenge the generalizability of OpenRE models, we assign each instance its top-level relation as the ground-truth label. We identify six single relations without a parent and designate them as novel relations. For TACRED and Re-TACRED datasets, novel relations are randomly selected from all relations. For more details about each dataset's split, see Table 1 and Appendix A.1.

Dataset	$\lvert\mathcal{C}_{known}\rvert$	$\lvert\mathcal{D}_l\rvert$	$\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{D}_u\rvert$
FewRel	35	22050	41	26250
TACRED	35	10074	41	11692
Re-TACRED	33	15586	39	18082

Table 1: Statistics of labeled and unlabeled datasets.

5.2 Baselines

We compare the proposed model MixORE with the following state-of-the-art OpenRE methods: (1) ORCA (Cao et al., 2022), (2) MatchPrompt (Wang et al., 2022), (3) TABs (Li et al., 2022), (4) HiURE (Liu et al., 2022), (5) AugURE (Wang et al., 2023), and (6) KNoRD (Hogan et al., 2023). Except for KNoRD, these baselines are not inherently designed for the generalized OpenRE setting and therefore require extensions. The extended baselines and related modifications are discussed as follows.

ORCA: As a computer vision model designed for a similar generalized open-world setting, ORCA does not require structural modifications. It is adapted to the relation extraction task by replacing ResNet with DeBERTa (He et al., 2021) and generating relation representations.

MatchPrompt′ and TABs′: The OpenRE methods MatchPrompt and TABs are inherently limited in their ability to differentiate between known and novel instances within unlabeled data. To address this, we treat all relations as novel and allow these models to effectively cluster the unlabeled data. We then apply the Hungarian Algorithm (Kuhn, 2010) to align some clusters with known relations, enabling performance evaluation on both known and novel relations.

Dataset	Method	P	R	F1	B^3			V-measure			ARI
					Prec.	Rec.	F1	Hom.	Comp.	F1	
FewRel	ORCA	0.6095	0.6328	0.6210	0.6347	0.4823	0.5481	0.6335	0.4848	0.5492	0.4318
	MatchPrompt′	0.7575	0.6271	0.6862	0.3031	0.8196	0.4426	0.4036	0.7599	0.5272	0.2394
	TABs′	0.7296	0.6955	0.7121	0.9193	0.7125	0.8028	0.9088	0.7071	0.7953	0.7746
	HiURE*	0.4441	0.4260	0.4349	0.9660	0.8147	0.8838	0.9615	0.8042	0.8758	0.8735
	AugURE*	0.5005	0.4770	0.4884	**0.9720**	0.7914	0.8723	**0.9647**	0.7941	0.8711	0.8568
	KNoRD	0.7701	0.7775	0.7738	0.8230	0.6587	0.7318	0.8286	0.6519	0.7297	0.6945
	MixORE	**0.8606**	**0.8067**	**0.8328**	0.9585	**0.8426**	**0.8968**	0.9490	**0.8206**	**0.8802**	**0.8817**
TACRED	ORCA	0.6845	0.7534	0.7173	0.7501	0.4751	0.5817	0.7381	0.4696	0.5740	0.4622
	MatchPrompt′	0.7145	0.5989	0.6516	**0.9357**	0.6046	0.7345	**0.9288**	0.6468	0.7626	0.7159
	TABs′	0.7650	0.8175	0.7904	0.8908	0.5462	0.6772	0.8937	0.6214	0.7331	0.6647
	HiURE*	0.4976	0.4699	0.4831	0.8908	0.7289	0.8003	0.9010	0.7520	0.8194	0.7953
	AugURE*	0.4989	0.4751	0.4867	0.8966	0.7743	0.8309	0.9071	0.7718	0.8340	0.8001
	KNoRD	0.8404	0.8638	0.8519	0.8860	0.6778	0.7680	0.8967	0.7033	0.7883	0.7193
	MixORE	**0.8624**	**0.9052**	**0.8833**	0.8973	**0.8429**	**0.8682**	0.9081	**0.8182**	**0.8599**	**0.8473**
Re-TACRED	ORCA	0.6578	0.7520	0.7018	0.6782	0.7810	0.7260	0.6388	0.6783	0.6579	0.5552
	MatchPrompt′	0.7160	0.5564	0.6262	0.9875	0.5416	0.6995	0.9805	0.6301	0.7672	0.6223
	TABs′	0.5976	0.6056	0.6015	0.9715	0.5054	0.6649	0.9653	0.6136	0.7503	0.5582
	HiURE*	0.4341	0.4041	0.4185	0.9721	0.7174	0.8253	0.9694	0.7250	0.8294	0.8494
	AugURE*	0.4551	0.4313	0.4429	**0.9942**	0.7575	0.8596	**0.9908**	0.7639	**0.8625**	0.8767
	KNoRD	0.8493	0.8853	0.8669	0.9698	0.4763	0.6389	0.9583	0.5903	0.7306	0.5081
	MixORE	**0.8972**	**0.9349**	**0.9156**	0.9779	**0.7918**	**0.8750**	0.9718	**0.7733**	0.8613	**0.8925**

Table 2: Performance of all methods on FewRel, TACRED, and Re-TACRED datasets. Precision (P), Recall (R), and F1 score are reported on ground-truth known instances. B^3, V-measure, and ARI evaluate the clustering performance on ground-truth novel instances. The details of baseline methods can be found in Sec. 5.2.

HiURE* and AugURE*: The original HiURE and AugURE models both operate in an unsupervised manner. For fair comparisons, we incorporate a supervised cross-entropy loss in addition to their overall loss function to help fine-tune their relation encoders. Similarly, we leverage the Hungarian Algorithm to assign clusters to known relations. Additionally, we exclude the use of ChatGPT in the AugURE model.

5.3 Evaluation Metrics

We evaluate the model performance on the unlabeled dataset $\mathcal{D}_u$. For instances belonging to ground-truth known relations, we measure the performance using precision, recall, and F1 score. For ground-truth novel relation instances, we evaluate clustering performance using B^3 (Bagga and Baldwin, 1998), V-measure (Rosenberg and Hirschberg, 2007), and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). For all of these metrics, higher values indicate better performance.

- B^3 precision and recall measure the quality and coverage of relation clustering, respectively. B^3 F1 score is computed to provide a balanced evaluation of clustering performance.
- V-measure is another widely used metric for evaluating clustering quality. Unlike B^3, which treats each instance individually, V-measure evaluates both intra-cluster homogeneity and inter-cluster completeness, offering a more comprehensive assessment of clustering performance by considering the overall structure of the clusters.
- Adjusted Rand Index (ARI) measures the level of agreement between the clusters produced by the model and the ground truth clusters. It ranges from $[-1, 1]$, where a value close to 1 indicates strong agreement, 0 represents random clustering, and negative values suggest disagreement.

5.4 Main Results

We evaluate MixORE against state-of-the-art baseline models on the FewRel, TACRED, and Re-TACRED datasets. Additional implementation details are provided in Appendix A.2. For all models, the average performance of two random runs is reported. The main results are shown in Table 2.

On ground-truth known relations, MixORE consistently outperforms the baseline models across all datasets, achieving the highest precision, recall, and F1 score. Notably, MixORE surpasses the previous best OpenRE model, KNoRD, by 5.90%, 3.14%, and 4.87% in F1 score on FewRel, TA-

Method	P	R	F1	B^3			V-measure			ARI
				Prec.	Rec.	F1	Hom.	Comp.	F1	
MixORE	**0.8606**	0.8067	**0.8328**	0.9585	**0.8426**	**0.8968**	0.9490	**0.8206**	**0.8802**	**0.8817**
− NRD (*pred_known*)	0.7374	**0.8440**	0.7871	-	-	-	-	-	-	-
− NRD (*pred_novel*)	-	-	-	**0.9709**	0.8065	0.8807	**0.9612**	0.8016	0.8741	0.8651
− Continual Learning	0.8484	0.8056	0.8264	0.8573	0.7830	0.8134	0.8648	0.7746	0.8154	0.7516
− Clustering Loss $\mathcal{L}_e$	0.8440	0.8063	0.8246	0.9373	0.7972	0.8615	0.9256	0.7771	0.8448	0.8382

Table 3: Ablation study on FewRel dataset.

CRED, and Re-TACRED, respectively. This highlights the effectiveness of MixORE in improving the classification performance of known relations.

In novel relation clustering, MixORE demonstrates competitive performance, consistently ranking among the top-performing models. Although other baselines occasionally achieve higher scores on certain metrics, MixORE exhibits the strongest overall performance, especially on the TACRED dataset, where it attains the highest B^3 F1 score, V-measure F1 score, and ARI. Compared to the second-best model, AugURE*, MixORE achieves improvements of 3.73%, 2.59%, and 4.72% in these metrics on the TACRED dataset, respectively.

These results suggest that MixORE effectively captures meaningful relation representations while maintaining a balance between known relation classification and novel relation clustering.

5.5 Ablation Study

To evaluate the contribution of different components, we conduct an ablation study by systematically excluding specific components. The results on FewRel dataset are presented in Table 3.

To assess the impact of the novel relation detection (NRD) module, we remove all the weakly-labeled novel instances from the training set (referred to as "− NRD"). Without NRD, the model cannot distinguish between known and novel relations in the unlabeled data, so we present the results as two separate settings: (1) *pred_known*, where the model assumes all relations are known and performs classification on the unlabeled data, and (2) *pred_novel*, where the model treats all relations as novel and performs clustering using K-Means algorithm. Subsequently, setting (1) and setting (2) are evaluated against ground-truth known and novel instances, respectively. The results reveal that excluding NRD leads to a notable -4.57% drop in the F1 score of known relation classification and a slight decline in novel relation clustering performance. This indicates that the weak labels play an essential role in enhancing the discriminative power on the relation classification task.

We also evaluate the performance of MixORE without the continual learning paradigm, where the model is initially provided with both labeled data and the weakly-labeled novel instances (referred to as "− Continual Learning"). As a result, we observe a minor decrease in known relation classification performance and a significant drop (-8.34%, -6.48%, and -13.01% in B^3 F1 score, V-measure F1 score, and ARI, respectively) in the clustering performance of novel relations. These results demonstrate that continual learning allows MixORE to use previously acquired knowledge to more effectively learn novel relations, making it well-suited for dynamic and evolving tasks.

To study the advantage of incorporating data distribution, we exclude the clustering exemplar loss function $\mathcal{L}_e$ from MixORE's parameter updates (referred to as "− Clustering Loss $\mathcal{L}_e$"). The results show a small decrease in the classification performance of known relations. For novel relation clustering, we see a performance change of -3.53%, -3.54%, and -4.35% in B^3 F1 score, V-measure F1 score, and ARI, respectively. This suggests that considering data distribution is beneficial for both known relation classification and novel relation clustering tasks.

6 Conclusion

This paper explores the generalized OpenRE task and introduces MixORE, a two-phase framework that jointly optimizes relation classification and clustering. MixORE effectively learns discriminative features for known relations while progressively integrating novel information from unlabeled data. Experiments on three benchmark datasets show the superiority of MixORE over competitive baselines, highlighting its effectiveness in balancing known relation classification and novel relation discovery. Our work advances the OpenRE task by introducing a more adaptable approach and offering valuable insights for both future research and real-world applications.

Limitations

While our proposed framework demonstrates strong performance in generalized OpenRE, it has certain limitations that call for further exploration.

One limitation of our approach is that it cannot automatically determine the number of novel relations present in the unlabeled data. Instead, it relies on a pre-defined number of clusters, which may not always align with the true distribution of novel relations. Future work could explore adaptive clustering techniques to dynamically estimate the number of novel relations, enhancing the flexibility and applicability of our framework.

Another limitation stems from our implicit assumption that relations are independent of each other. In reality, relations may have hierarchical dependencies, such as being child or parent relations of other relations. Our current method does not explicitly model these dependencies, which may lead to suboptimal performance. Future research could incorporate relational hierarchies into the learning process, enabling a more comprehensive understanding of relation dependencies and improving the model's ability to handle complex relation structures.

Ethics Statement

We comply with the ACL Code of Ethics.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Annual Meeting of the Association for Computational Linguistics*.

Richard H. Bartels and G. W. Stewart. 1972. Solution of the matrix equation ax+xb=c [F4] (algorithm 432). *Commun. ACM*, 15(9):820–826.

Kaidi Cao, Maria Brbic, and Jure Leskovec. 2022. Open-world semi-supervised learning. In *ICLR*. OpenReview.net.

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. 2021. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In *ACL/IJCNLP (1)*, pages 232–243. 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.

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 *EMNLP*, pages 4803–4809. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *ICLR*. OpenReview.net.

William Hogan, Jiacheng Li, and Jingbo Shang. 2023. Open-world semi-supervised generalized relation discovery aligned in a real-world setting. In *EMNLP*, pages 14227–14242. Association for Computational Linguistics.

Lawrence J. Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Elyor Kodirov, Tao Xiang, and Shaogang Gong. 2017. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 4447–4456. IEEE Computer Society.

Harold W. Kuhn. 2010. The hungarian method for the assignment problem. In *50 Years of Integer Programming*, pages 29–47. Springer.

Sha Li, Heng Ji, and Jiawei Han. 2022. Open relation and event type discovery with type abstraction. In *EMNLP*, pages 6864–6877. Association for Computational Linguistics.

Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. 2021. Element intervention for open relation extraction. In *ACL/IJCNLP (1)*, pages 4683–4693. Association for Computational Linguistics.

Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip S. Yu. 2022. Hiure: Hierarchical exemplar contrastive learning for unsupervised relation extraction. In *NAACL-HLT*, pages 5970–5980. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Scott Miller, Michael Crystal, Heidi Fox, Lance A. Ramshaw, Richard M. Schwartz, Rebecca Stone, and Ralph M. Weischedel. 1998. BBN: description of the SIFT system as used for MUC-7. In *Seventh*

Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia, USA, MUC 1998, April 29 - May 1, 1998. ACL.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. Comput. Linguistics*, 5:101–115.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 410–420. ACL.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. 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 2895–2905. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *AAAI Conference on Artificial Intelligence*.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. In *ACL (1)*, pages 15566–15589. Association for Computational Linguistics.

Jiaxin Wang, Lingling Zhang, Jun Liu, Xi Liang, Yujie Zhong, and Yaqiang Wu. 2022. Matchprompt: Prompt-based open relation extraction with semantic consistency guided clustering. In *EMNLP*, pages 7875–7888. Association for Computational Linguistics.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5362–5383.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In *EMNLP*, pages 12136–12147. Association for Computational Linguistics.

Ting Wu, Jingyi Liu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Enhancing contrastive learning with noise-guided attack: Towards continual relation extraction in the wild. In *ACL (1)*, pages 2227–2239. Association for Computational Linguistics.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3683–3688. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*, pages 71–78.

Xinwei Zhang, Jianwen Jiang, Yutong Feng, Zhi-Fan Wu, Xibin Zhao, Hai Wan, Mingqian Tang, Rong Jin, and Yue Gao. 2022. Grow and merge: A unified framework for continuous categories discovery. In *NeurIPS*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *EMNLP*, pages 35–45. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 50–61. Association for Computational Linguistics.

A Appendix

A.1 Novel Relations in Each Dataset

The six single relations without a parent we used as FewRel novel relations are as follows:

"publisher"
"nominated for"
"instrument"
"notable work"
"competition class"
"position played on team/speciality"

The randomly selected novel relations from TACRED are as follows:

"per:city_of_birth"
"org:stateorprovince_of_headquarters"
"org:member_of"
"per:date_of_death"
"per:city_of_death"
"per:children"

The randomly selected novel relations from Re-TACRED are as follows:

"per:siblings"
"org:founded_by"
"org:city_of_branch"
"per:countries_of_residence"
"per:date_of_birth"
"per:city_of_death"

A.2 Implementation Details

In the first phase, we set the weighting coefficient to $\lambda = 100$. During the second phase, we optimize the loss using AdamW (Loshchilov and Hutter, 2019). The encoder is warmed up for 2 epochs and continually trained for 5 epochs, all with a learning rate of $1e-5$. We set the margin for the triplet margin loss on labeled data to $\gamma = 0.75$. For the clustering exemplar loss function, we use a temperature parameter of $\tau = 0.02$ and include $J = 10$ negative examples. We implement the granularity layer with $L = 4$, setting $c_l \in [16, 32, 41, 64]$ for FewRel and TACRED, and $c_l \in [16, 32, 39, 64]$ for Re-TACRED. All experiments are conducted on an NVIDIA Tesla V100 GPU.

This work and its associated artifacts are licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) License, allowing unrestricted use, distribution, and reproduction, provided the original work is properly cited using standard academic practices.