Cluster-aware Pseudo-Labeling for Supervised Open Relation Extraction

Bin Duan*, **Shusen Wang***, **Xingxian Liu***, **Yajing Xu**
Pattern Recognition & Intelligent System Laboratory,
Beijing University of Posts and Telecommunications, Beijing, China
{bobtuan,shusenw,liuxingxian,xyj}@bupt.edu.cn

Abstract

Supervised open relation extraction aims to discover novel relations by leveraging supervised data of pre-defined relations. However, most existing methods do not achieve effective knowledge transfer from pre-defined relations to novel relations, they have difficulties generating high-quality pseudo-labels for unsupervised data of novel relations and usually suffer from the error propagation issue. In this paper, we propose a **C**luster-**a**ware **P**seudo-**L**abeling (**CaPL**) method to improve the pseudo-labels quality and transfer more knowledge for discovering novel relations. Specifically, the model is first pre-trained with the pre-defined relations to learn the relation representations. To improve the pseudo-labels quality, the distances between each instance and all cluster centers are used to generate cluster-aware soft pseudo-labels for novel relations. To mitigate the catastrophic forgetting issue, we design the consistency regularization loss to make better use of the pseudo-labels and jointly train the model with both unsupervised and supervised data. Experimental results on two public datasets demonstrate that our proposed method achieves new state-of-the-arts performance[1].

1 Introduction

Open relation extraction (OpenRE) has been proposed to extract the novel relations that are not defined or observed beforehand. Previous methods can be classified into two types: unsupervised and supervised. Unsupervised OpenRE (Yao et al., 2011, 2012; Marcheggiani and Titov, 2016; Elsahar et al., 2017; Tran et al., 2020; Hu et al., 2020) regards the OpenRE as a totally unsupervised task which first extracts the feature and then clusters them. However, these methods don't take full advantage of the large amounts of existing relational facts in knowledge bases. Hence, the supervised OpenRE is proposed which leverages the supervised data of pre-defined relations to guide the learning of the unsupervised data of novel relations. In this paper, we focus on the latter setting, supervised OpenRE.

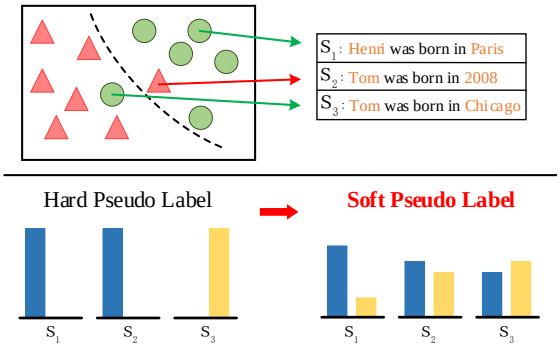

Figure 1: Relation instance S_1 and S_3 belong to the same relation type while S_2 with similar context is from another relation type. However, the hard pseudo-labels produced by common clustering methods contain much noise which causes the error propagation issue. We generate soft pseudo-labels that contain the information about true cluster to improve the pseudo-labels quality.

Since the classes between pre-defined relations and novel relations are disjoint, the main challenge of supervised OpenRE is how to make the best use of the prior knowledge in pre-defined relations to extract novel relations. Wu et al. (2019) proposes relational siamese networks to transfer the knowledge from pre-defined relations to novel relations. However, many studies have shown that high-dimensional embeddings learn much about the complex linguistic information (Peters et al., 2018; Jawahar et al., 2019; Clark et al., 2019; Goldberg, 2019), which makes it hard to produce ideal clusters. Zhao et al. (2021) proposes a relation-oriented clustering method that explicitly aligns the derived clusters with relational semantic classes.

However, we find that the pseudo-labels produced by previous method are not robust to transfer

*The first three authors contribute equally. Yajing Xu is the corresponding author.

[1]Code is available at https://github.com/BobTuan/CaPL

Proceedings of the 29th International Conference on Computational Linguistics, pages 1834–1841
October 12–17, 2022.

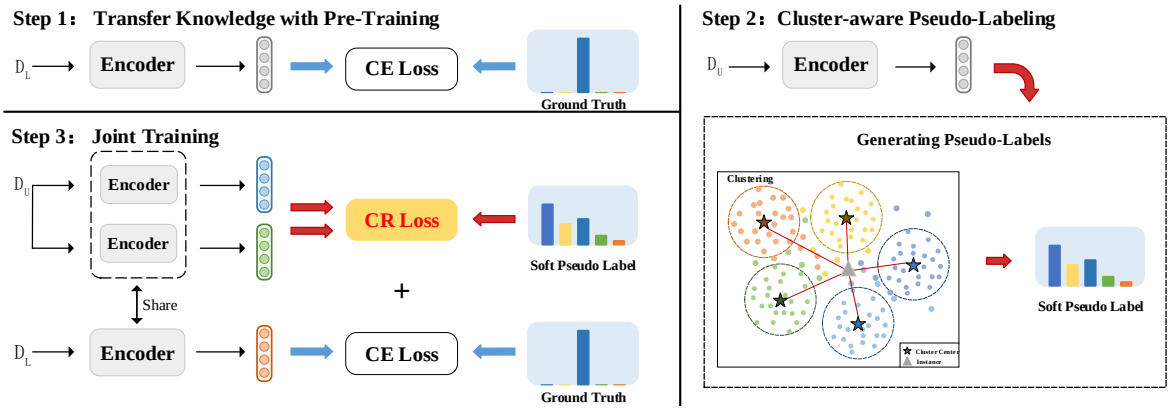

Figure 2: Overview of our CaPL method. Firstly, we pre-train the model with supervised data of pre-defined relations. Then, we generate cluster-aware soft pseudo-labels for unsupervised data of novel relations. Finally, we jointly train the model with both supervised data and unsupervised data. After step 1, step 2 and step 3 are performed iteratively to gradually improve model performance on novel relations.

the knowledge from pre-defined relations to novel relations which means that using unreliable pseudo-labels could cause the error propagation issue. As is shown in Figure 1, relation instance S_1 and S_3 belong to the same relation type *BORN_IN_PLACE* while S_2 with similar context is from another relation type *BORN_IN_DATE*. Owing to the spurious correlations (Liu et al., 2021), existing method that selects the nearest cluster center as the pseudo-labels may produce noise pseudo-labels. Hence if we further exploit the hard pseudo-labels, the model would be wrongly optimized. We argue that utilizing the information about all cluster centers to generate the soft pseudo-labels would be helpful to reduce the error propagation issue. The intuition is that if we exploit the soft pseudo-labels, we can utilize the information about the true cluster to guide the learning for discovering novel relations.

In this paper, we propose a **C**luster-**a**ware **P**seudo-**L**abeling (**CaPL**) method to improve the pseudo-labels quality and transfer more knowledge for discovering novel relations. Firstly, we pre-train the model under the supervision of cross-entropy loss to leverage the prior knowledge in pre-defined relations. Then, to effectively transfer the knowledge, rather than directly using the hard pseudo-labels produced by common clustering algorithms, we use the distances between each instance and all cluster centers to generate cluster-aware soft pseudo-labels for novel relations. Finally, we design consistency regularization loss to make the best use of the knowledge stored in the cluster-aware pseudo-labels and jointly train the model with both supervised and unsupervised data to mitigate the catastrophic forgetting issue.

To summarize, the major contributions of our work are as follows: (1) We propose a simple but effective framework based on the CaPL for supervised OpenRE whicn can transfer more knowledge for discovering novel relations. (2) We design the consistency regularization loss to keep the cluster predictions and pseudo-labels of unsupervised data to be consistent for making better use of the pseudo-labels. (3) Experimental results and analyses on two public datasets demonstrate the effectiveness of our proposed method.

2 Method

In the supervised OpenRE settings, training data is split into two sets: a supervised dataset of pre-defined relations $D_l = \left\{\left(x_i^l, y_i^l\right), i = 1, \ldots, N\right\}$ and an unsupervised dataset of novel relations $D_u = \{x_i^u, i = 1, \ldots, M\}$, where x_i^l in D_l and x_i^u in D_u is an relation instance and y_i^l is a categorical label. Our goal is to cluster the D_u to discover C^u novel relations where we assume C^u to be known a priori. The set of C^l labeled classes is assumed to be disjoint from the set of C^u unlabeled classes.

In this work, we propose a simple but effective framework based on the CaPL to improve the pseudo-labels quality for discovering novel relations. Figure 2 shows the overall architecture of our proposed method. We will introduce these step details in the following subsections.

2.1 Leverage Knowledge with Pre-Training

To leverage the prior knowledge in pre-defined relations, we use the supervised data of pre-defined relations to pre-train the model. The goal of pre-training with pre-defined relations is to make the

model adapt to the relational feature space while be less biased towards pre-defined relations.

Specifically, we learn the relational feature representations under the supervision of cross-entropy loss due to its simplicity and efficacy.

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\phi\left(\boldsymbol{s}_i\right)^{y_i}\right)}{\sum_{j=1}^{K}\exp\left(\phi\left(\boldsymbol{s}_i\right)^{j}\right)}, \quad (1)$$

where $\boldsymbol{s}_i$ is the i^{th} relation feature representation of pre-defined relations, $\phi\left(\cdot\right)$ is a linear classifier and $\phi\left(\cdot\right)^j$ are the output logits of the j^{th} class.

2.2 Cluster-aware Pseudo-Labeling

Since there is a mass of unsupervised data of novel relations, it's important to effectively leverage these samples to discover novel relations. Pseudo-Labeling is a well-established technique for transfer learning in general (Cui et al., 2021). After leveraging the prior knowledge in pre-defined relations with pre-training, we propose a simple but effective method CaPL for transferring the knowledge to discover novel classes.

To transfer more knowledge, rather than directly generating the hard pseudo-labels with common clustering algorithm like Zhao et al. (2021), we generate more robust cluster-aware soft pseudo-labels. More specifically, we first obtain the relation instance representations $H = \{h_1, \ldots, h_N\}$, and then perform k-means algorithm in the relational feature space to obtain K cluster centers, denoted as μ_k, $k \in \{1, \ldots, K\}$. Different from that the standard k-means algorithm regards the indicator of the nearest cluster center as the hard pseudo-labels, we adopt a soft assignment to K cluster centers for each instance. Inspired by Hu et al. (2020), we use the Student's t-distribution to compute the assigning probability between relation instance h_j and each cluster center μ_k:

$$p_{jk} = \frac{\left(1 + \|h_j - \mu_k\|_2^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K}\left(1 + \|h_j - \mu_{k'}\|_2^2/\alpha\right)^{-\frac{\alpha+1}{2}}}, \quad (2)$$

where α denotes the degree of freedom of the Student's t-distribution and p_{jk} can be regarded as the probability of assigning the sample j to the cluster center k. Without explicit mention, we set α = 1 for all experiments in this paper. In addition, we can also use other common clustering algorithms to generate the cluster centers.

2.3 Joint Training

Conventional cross-entropy loss cannot work with the cluster-aware pseudo-labels. To make better use of the knowledge in the pseudo-labels, we design the consistency regularization loss. The idea of consistency is that the cluster prediction and pseudo-labels on a relation instance h_j and its transformed counterpart $h_j^{'}$ should be the same. In our case, we use dropout twice to get h_j's transformed counterpart $h_j^{'}$ like Gao et al. (2021) and then map these relation representations into the cluster predictions q_j and $q_j^{'}$ with the same equation 2. Finally, we use the KL-divergence to keep the consistency between cluster predictions and pseudo-labels:

$$\ell_j = \mathrm{KL}\left[p_j \| q_j\right] + \mathrm{KL}\left[p_j \| q_j^{'}\right] + \mathrm{KL}\left[q_j \| q_j^{'}\right] \quad (3)$$

$$\mathrm{KL}\left[p_j \| q_j\right] = \sum_{k=1}^{K} p_{jk}\log\frac{p_{jk}}{q_{jk}} \quad (4)$$

$$\mathcal{L}_{cr} = \sum_{j=1}^{N}\frac{\ell_j}{N}, \quad (5)$$

To let the supervised data of pre-defined relations better guide the learning of discovering novel relations and mitigate the catastrophic forgetting issue, we jointly train the model with both supervised and unsupervised data. The overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \omega(r)\mathcal{L}_{cr}, \quad (6)$$

where $\omega(r)$ is a ramp-up function slowly increasing from 0 to 1 along with the training. Following Zhao et al. (2021), we use the sigmoid-shaped function $\omega(r) = \lambda e^{-5\left(1-\frac{r}{T}\right)^2}$, where r is current epoch and T is ramp-up length and a positive scalar factor λ.

3 Experiments

3.1 Datasets

To assess the performance of our method, we conduct experiments on two relation extraction datasets: FewRel (Han et al., 2018) and TACRED (Zhang et al., 2017). **FewRel** is a human-annotated dataset which contains 80 types of relations, each with 700 relation instances. **TACRED** is also a human-annotated dataset with 41 relation types. Following the setup of Zhao et al. (2021), we split the FewRel dataset into 64 pre-defined relations and 16 novel relations and randomly select 1,600 instances in novel relations as the test set. For TACRED, we also remove the instances labeled as

Dataset	Method	B^3			V-measure			ARI	Avg.
		Prec.	Rec.	F1.	Homo.	Comp.	F1.		
FewRel	VAE(Marcheggiani and Titov, 2016)	0.309	0.446	0.365	0.448	0.500	0.473	0.291	0.376
	RW-HAC(Elsahar et al., 2017)	0.256	0.492	0.337	0.391	0.485	0.433	0.250	0.340
	Etype+(Tran et al., 2020)	0.238	0.485	0.319	0.364	0.463	0.408	0.249	0.325
	SelfORE(Hu et al., 2020)	0.672	0.685	0.678	0.779	0.788	0.783	0.647	0.703
	RSN(Wu et al., 2019)	0.486	0.742	0.589	0.644	0.787	0.708	0.453	0.583
	RSN-BERT	0.585	0.899	0.709	0.696	0.889	0.781	0.532	0.674
	RoCORE(Zhao et al., 2021)	0.752	0.846	0.796	0.838	0.883	0.860	0.709	0.788
	CaPL	0.815	0.822	**0.819**	0.875	0.873	**0.889**	**0.794**	**0.834**
	CaPL w/o Pre-training	0.735	0.802	0.767	0.834	0.865	0.850	0.693	0.770
	CaPL w/o Consistency Regularization	0.752	0.785	0.768	0.840	0.855	0.847	0.738	0.784
	CaPL w/o Joint Training	0.768	0.820	0.793	0.845	0.875	0.860	0.718	0.790
TACRED	VAE(Marcheggiani and Titov, 2016)	0.247	0.564	0.343	0.208	0.362	0.264	0.159	0.255
	RW-HAC(Elsahar et al., 2017)	0.426	0.633	0.509	0.469	0.597	0.526	0.281	0.439
	Etype+(Tran et al., 2020)	0.302	0.803	0.439	0.260	0.607	0.364	0.143	0.315
	SelfORE(Hu et al., 2020)	0.576	0.510	0.541	0.630	0.608	0.619	0.447	0.536
	RSN(Wu et al., 2019)	0.628	0.634	0.631	0.624	0.663	0.643	0.459	0.578
	RSN-BERT	0.795	0.878	0.834	0.849	0.870	0.859	0.756	0.816
	RoCORE(Zhao et al., 2021)	0.871	0.849	0.860	0.895	0.881	0.888	0.812	0.853
	CaPL	0.858	0.888	**0.873**	0.891	0.906	**0.898**	**0.829**	**0.867**
	CaPL w/o Pre-training	0.834	0.847	0.840	0.868	0.870	0.869	0.789	0.833
	CaPL w/o Consistency Regularization	0.856	0.795	0.824	0.883	0.843	0.862	0.743	0.810
	CaPL w/o Joint Training	0.835	0.827	0.831	0.870	0.855	0.862	0.788	0.827

Table 1: Experimental results produced by baselines and proposed model on FewRel and TACRED in terms of B^3, V-measure, ARI and average performance. The horizontal line divides unsupervised and supervised methods.

no_relation. We seperately select 30 pre-defined relations and 10 novel relations. In addition, we randomly select 15% of the instances from the novel relations as the test set.

3.2 Baselines

For comparison, we consider both unsupervised and supervised OpenRE baselines for comparison:

- **Unsupervised.** We first compare with unsupervised OpenRE methods. **VAE** (Marcheggiani and Titov, 2016) proposes a VAE-based model learned by the self-supervised signals. **RW-HAC** (Elsahar et al., 2017) first extracts entity types and re-weights the word embeddings and then clusters them. **Etype+** (Tran et al., 2020) solely uses entity types as the input. **SelfORE** (Hu et al., 2020) proposes a self-supervised framework which learns the embeddings with pseudo-labels.

- **Supervised.** We also compare our method with supervised OpenRE methods. **RSN** (Wu et al., 2019) proposes the relation similarity metrics to transfer the knowledge to discover novel relations. **RSN-BERT** replaces the static word embeddings with the pre-trained BERT embeddings for a fair comparison. **RoCORE** (Zhao et al., 2021) proposes a relation-oriented method to explicitly align the derived clusters with relational semantic classes.

3.3 Implement Details

We use the pre-trained model (bert-base-uncased[2], with 12-layer transformer) as our network backbone. To avoid overfitting and improve the training efficiency, as suggested in Zhao et al. (2021), we freeze all the parameters of BERT and only fine-tune the parameters of the layer 8. The training batch size is 128, the learning rate is 1e-4, and we use Adam (Kingma and Ba, 2014) as optimizer. All experiments are conducted by using a GeForce RTX 3090Ti with 24 GB memory.

3.4 Main Results

The main results are shown in Table 1. The proposed method CaPL achieves SOTA performance in all datasets and evaluation metrics. It demonstrates the effectiveness that our method leverages the pre-defined relations to extract novel relations. In addition, we find that most supervised methods perform better than unsupervised methods. It indicates that transferring knowledge from pre-defined relations is helpful to discover novel relations.

3.5 Ablation Analysis

To study the effect of different components in CaPL, we conduct ablation experiments. From Table 1, we find that the performance of CaPL will severely degrade without these modules, which

[2]https://huggingface.co/bert-base-uncased

Task	Method	Prec.	Rec.	F1.	Avg.
FewRel	RoCORE	0.752	0.846	0.796	0.788
	CaPL-hard	0.753	0.811	0.781	0.781
	CaPL-soft	0.815	0.822	**0.819**	**0.834**
TACRED	RoCORE	0.871	0.849	0.860	0.853
	CaPL-hard	0.832	0.850	0.841	0.827
	CaPL-soft	0.858	0.888	**0.873**	**0.867**

Table 2: Experimental results with different pseudo-labels under the same pre-training setting on FewRel and TACRED. CaPL-hard adopts the same hard pseudo-labels with RoCORE while our method CaPL-soft adopts the cluster-aware soft pseudo-labels. This table only lists the results of metric B^3. For results of other metrics, please refer to the Appendix D.

Task	Method	Prec.	Rec.	F1.	Avg.
F → T	RSN	0.349	0.590	0.439	0.387
	RSN-BERT	0.337	0.866	0.486	0.400
	RoCORE	0.621	0.602	0.611	0.642
	CaPL	0.813	0.601	**0.691**	**0.847**
T → F	RSN	0.225	0.529	0.316	0.359
	RSN-BERT	0.261	0.861	0.400	0.438
	RoCORE	0.687	0.766	0.724	0.796
	CaPL	0.722	0.757	**0.739**	**0.802**

Table 3: Results on two cross-domain tasks. F means FewRel, which is from encyclopedia domain. T means TACRED, which is from news and web domain. This table only lists the results of metric B^3. For results of other metrics, please refer to the Appendix D.

demonstrates that all modules are important to the final model performance. It is worth noting that without consistency regularization the performance is seriously hurt which indicates that the loss we designed makes better use of the pseudo-labels. Further study about consistency regularization can be found in Appendix B.

3.6 Effect of Pseudo-labels Quality

In this section, we analyse the effect of pseudo-labels quality for transferring knowledge to discover novel relations. Specifically, we adopt the same method as RoCORE to generate and utilize the hard pseudo-labels and combine it into our framework which is named as CaPL-hard. From Table 2, we can see that under the same pre-training setting, our method that generates and utilizes cluster-aware soft pseudo-labels significantly outperforms the CaPL-hard method which indicates that our method generates the high-quality pseudo-labels and makes the best use of them for discovering novel relations.

3.7 Cross Domain Analysis

To further study the knowledge transfer ability, we adopt more strict cross-domain settings to evaluate

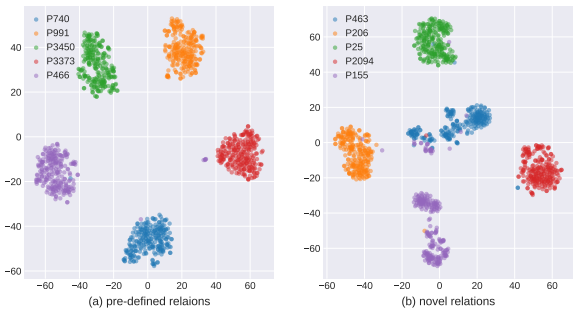

Figure 3: Visualization of the learned relation representations by CaPL on both pre-defined and novel relations.

the model in which the pre-defined and novel relations come from different domains. Specifically, we conduct two cross-domain tasks: FewRel to TACRED and TACRED to FewRel. In Table 3, we observe that in the more realistic cross-domain settings, our model shows better generalization performance on novel relations which indicates that our method effectively transfers the knowledge to discover novel relations.

3.8 Visualization Analysis

To explore the effectiveness on the refinement of relation representations in semantic space, we visualize the representations of both pre-defined and novel relations. We randomly choose 5 relations and sample 250 instances for each relation separately in pre-defined and novel relations. As is shown in Figure 3, the relation representations from both pre-defined and novel relations are mostly separated in our proposed method which means that our method not only fully leverages the prior knowledge for discovering novel relations but also mitigates the catastrophic forgetting issue for pre-defined relations.

4 Conclusion

In this paper, we propose an effective framework based on Cluster-aware Pseudo-Labeling (CaPL) to transfer more knowledge for discovering novel relations. Our main contribution is to improve the knowledge transfer ability of the model. The proposed method makes better use of the prior knowledge in pre-defined relations and generalizes to novel relations with the high-quality pseudo-labels. Experiments and analyses confirm the effectiveness of CaPL for supervised OpenRE.

Acknowledgements

We thank all anonymous reviewers for their valuable suggestions. This work was supported by the National Natural Science Foundation of China (NSFC No.62076031).

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 232–243, Online. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest. 2017. Unsupervised open relation extraction. In *European Semantic Web Conference*, pages 12–16. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

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

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: Self-supervised relational feature learning for open relation extraction. *CoRR*, abs/2004.02438.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. 2021. Element intervention for open relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4683–4693, Online. Association for Computational Linguistics.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

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

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2020. Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7498–7505, Online. Association for Computational Linguistics.

Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2019. Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 219–228, Hong Kong, China. Association for Computational Linguistics.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2012. Unsupervised relation discovery with sense disambiguation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 712–720, Jeju Island, Korea. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9707–9718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Evaluation Metrics

For evaluation metrics, we adopt B^3 (Bagga and Baldwin, 1998), V-measure (Rosenberg and Hirschberg, 2007), and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), the same as Wu et al. (2019); Zhao et al. (2021). We take the average for comprehensive evaluation, since that any of the three metrics can measure the clustering performance from different angles.

B Effect of Consistency Regularization Loss

Table 4 shows the effect of consistency regularization loss for exploiting the pseudo-labels. We individually add different components of the consistency regularization loss in equation 3, including the consistency between cluster predictions and cluster predictions (CC) and the consistency between cluster predictions and pseudo-labels (CP). We can observe that both CC and CP consistency are helpful to extract novel relations while CP consistency performs better than CC consistency. We argue it's the knowledge in the pseudo-labels that makes it.

Dataset	Method	Prec.	Rec.	F1.	Avg.
FewRel	Baseline	0.752	0.785	0.768	0.784
	+ CC	0.747	0.822	0.783	0.786
	+ single CP	0.753	0.832	0.790	0.790
	+ double CP	0.794	0.802	0.798	0.812
	+ double CP + CC	0.815	0.822	**0.819**	**0.834**
TACRED	Baseline	0.856	0.795	0.824	0.810
	+ CC	0.845	0.823	0.835	0.830
	+ single CP	0.842	0.856	0.849	0.841
	+ double CP	0.844	0.871	0.857	0.849
	+ double CP + CC	0.858	0.888	**0.873**	**0.867**

Table 4: Evaluation of the effectiveness of the proposed consistency regularization loss. **Baseline**: CaPL without consistency regularization loss, **CC**: consistency between cluster predictions and cluster predictions, **CP**: consistency between cluster predictions and pseudo-labels. This table only lists the results of metric B^3. Refer to Table 5 for detailed results.

C More details about Encoder

In this section, we introduce how we encode the relation instance using the pre-trained models. Given a relation instance with n tokens as $w = [w_1, w_2, \ldots, w_n]$, where head entity e_h and tail entity e_t are marked with the start and end position of the entity. In addition, we adopt the pre-trained language model BERT (Devlin et al., 2019) to encode each token w_t to the corresponding representation $h_t \in R^d$ where d denotes the dimension of representation vectors. Then, we obtain the hidden state vectors of two entities h_{ent} by averaging their respective token's hidden state vectors:

$$h_{ent} = mean\text{-}pooling([h_s, \ldots, h_e]) \quad (7)$$

where $h_{ent} \in R^d$. s and e represent start and end position of the corresponding entity respectively. Finally, we use the concatenation of representations of two entity as the representation of the relation instance $h \in R^{2 \cdot d}$:

$$h = [h_{head}, h_{tail}] \quad (8)$$

D Detailed Results of Other Experiments

In this section, detailed results of consistency regularization loss, different pseudo-labels analysis and cross domain analysis are listed in Table 5, Table 6 and Table 7 respectively.

Dataset	Method	B^3			V-measure			ARI	Avg.
		Prec.	Rec.	F1.	Homo.	Comp.	F1.		
FewRel	Baseline	0.752	0.785	0.768	0.840	0.855	0.847	0.738	0.784
	+ CC	0.747	0.822	0.783	0.837	0.867	0.852	0.722	0.786
	+ single CP	0.753	0.832	0.790	0.839	0.872	0.855	0.726	0.790
	+ double CP	0.794	0.802	0.798	0.861	0.869	0.865	0.773	0.812
	+ double CP + CC	0.815	0.822	**0.819**	0.875	0.873	**0.889**	**0.794**	**0.834**
TACRED	Baseline	0.856	0.795	0.824	0.883	0.843	0.862	0.743	0.810
	+ CC	0.845	0.823	0.835	0.879	0.855	0.867	0.789	0.830
	+ single CP	0.842	0.856	0.849	0.870	0.874	0.872	0.801	0.841
	+ double CP	0.844	0.871	0.857	0.870	0.884	0.877	0.812	0.849
	+ double CP + CC	0.858	0.888	**0.873**	0.891	0.906	**0.898**	**0.829**	**0.867**

Table 5: The detailed results of the proposed consistency regularization loss.

Dataset	Method	B^3			V-measure			ARI	Avg.
		Prec.	Rec.	F1.	Homo.	Comp.	F1.		
FewRel	RoCORE	0.752	0.846	0.796	0.838	0.883	0.860	0.709	0.788
	CaPL-hard	0.753	0.811	0.781	0.843	0.873	0.858	0.705	0.781
	CaPL-soft	0.815	0.822	**0.819**	0.875	0.873	**0.889**	**0.794**	**0.834**
TACRED	RoCORE	0.871	0.849	0.860	0.895	0.881	0.888	0.812	0.853
	CaPL-hard	0.832	0.850	0.841	0.867	0.878	0.872	0.769	0.827
	CaPL-soft	0.858	0.888	**0.873**	0.891	0.906	**0.898**	**0.829**	**0.867**

Table 6: The detailed results of different pseudo-labels analysis.

Task	Method	B^3			V-measure			ARI	Avg.
		Prec.	Rec.	F1.	Homo.	Comp.	F1.		
F $\rightarrow$ T	RSN	0.349	0.590	0.439	0.387	0.533	0.448	0.279	0.389
	RSN-BERT	0.337	0.866	0.486	0.400	0.777	0.528	0.352	0.455
	RoCORE	0.621	0.602	0.611	0.642	0.666	0.654	0.451	0.572
	CaPL	0.813	0.601	**0.691**	0.847	0.703	**0.769**	**0.650**	**0.703**
T $\rightarrow$ F	RSN	0.225	0.529	0.316	0.359	0.507	0.420	0.243	0.326
	RSN-BERT	0.261	0.861	0.400	0.438	0.822	0.571	0.263	0.411
	RoCORE	0.687	0.766	0.724	0.796	0.836	0.815	0.658	0.732
	CaPL	0.722	0.757	**0.739**	0.802	0.830	**0.816**	**0.664**	**0.740**

Table 7: The detailed results of cross domain analysis.