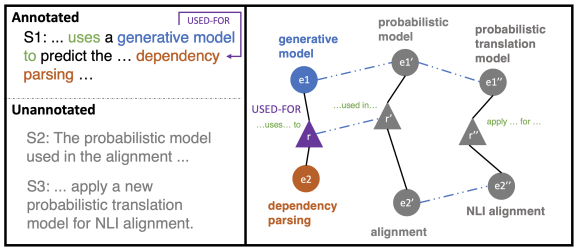
Jointprop: Joint Semi-supervised Learning for Entity and Relation Extraction with Heterogeneous Graph-based Propagation

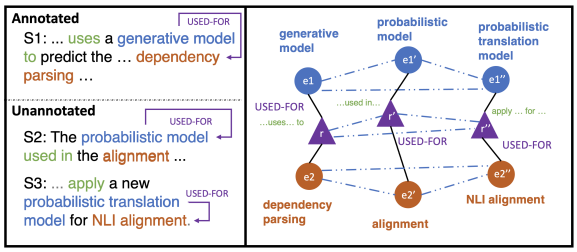
Zheng Yandan[1,2], **Hao Anran**[1] **and Luu Anh Tuan**[1]
[1]School of Computer Science and Engineering
[2] Interdisciplinary Graduate Program-HealthTech
Nanyang Technological University, Singapore
{yandan002, s190003}@e.ntu.edu.sg
anhtuan.luu@ntu.edu.sg

Abstract

Semi-supervised learning has been an important approach to address challenges in extracting entities and relations from limited data. However, current semi-supervised works handle the two tasks (i.e., Named Entity Recognition and Relation Extraction) separately and ignore the cross-correlation of entity and relation instances as well as the existence of similar instances across unlabeled data. To alleviate the issues, we propose *Jointprop*, a Heterogeneous Graph-based Propagation framework for joint semi-supervised entity and relation extraction, which captures the global structure information between individual tasks and exploits interactions within unlabeled data. Specifically, we construct a unified span-based heterogeneous graph from entity and relation candidates and propagate class labels based on confidence scores. We then employ a propagation learning scheme to leverage the affinities between labelled and unlabeled samples. Experiments on benchmark datasets show that our framework outperforms the state-of-the-art semi-supervised approaches on NER and RE tasks. We show that the joint semi-supervised learning of the two tasks benefits from their codependency and validates the importance of utilizing the shared information between unlabeled data.



(a) Before label propagation



(b) After label propagation

Figure 1: An example of label propagation. We represent the sentence as a triplet ($entity_1$, relation, $entity_2$) which consists of an entity pair (circle) and a relation (triangle) in a graph structure. The colored nodes indicate labeled semantic units (entity or relation candidates), while the uncolored nodes represent the unlabeled semantic units. Purple denotes relation label `Used-for`, blue denotes for entity label `Method`, and orange denotes another entity label label `Task`.

1 Introduction

Named Entity Recognition (NER) and Relation Extraction (RE) are two crucial tasks in Information Extraction. Supervised learning schemes have made significant progress in NER and RE research by leveraging rich annotated data (e.g., Lin et al. (2020); Yamada et al. (2020); Baldini Soares et al. (2019). However, high-quality data annotation still involves extensive and expensive labor. Moreover, training NER and RE models in various domains and applications require diverse annotated data. Semi-supervised learning approaches (SSL) employ a small amount of annotated data as a source of supervision for learning powerful models at a lower cost.

SSL in NER and RE have performed very well in recent years by employing bootstrapping, distant supervision or graph-based approach (Batista et al., 2015; Zeng et al., 2015; Delalleau et al., 2005). However, they train a NER model (Yang et al., 2018; Chen et al., 2018; Lakshmi Narayan et al., 2019) or a RE model (Lin et al., 2019; Hu et al., 2021a; Li et al., 2021) separately. Therefore, they neglect the underlying connections between entity recognition and relation extraction under a semi-supervised learning scheme, making it harder for the model to assign accurate annotation to unla-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14541–14555
July 9-14, 2023

beled data. For instance, in Figure 1, the annotated entity "*generative model*" in sentence S1 and the unannotated "*probabilistic model*" in sentence S2 are syntactically similar. Likewise, the context phrases "*uses... to*" and "*used in*" are also similar. If such similarities are ignored, the model may fail to draw a syntactic analogy between "*dependency parsing*" and "*alignment*", and thereby miss labeling the latter as an entity that shares the same type with the former. To the best of our knowledge, there is no universal framework to integrate semi-supervised learning for different tasks in IE, despite evidence of the effectiveness of a joint or multi-task learning approach (Luan et al., 2018a, 2019; Ye et al., 2021; Luan et al., 2018a, 2019; Lin et al., 2020).

In addition, existing semi-supervised approaches devote considerable effort to aligning labeled and unlabeled data but do not exploit similarities between instance pairs that are structurally parallel, which exist across unlabeled data. Consequently, they do not perform classification from the perspective of global consistency. For example, given the sentences S1 to S3 in Figure 1, we expect a model to recognize the entities and relations as (`Method`, `Used-for`, `Task`) in triplet form. However, it is hard to infer the correct pseudo label to the unlabeled entities "*alignment*" or "*NLI alignment*" from the annotated entity "*dependency parsing*". Because they are not semantically or lexically similar. Likewise, the affinity between "*uses to*" and "*apply*" is not obvious; and hence it would be difficult to extract the relation `Used-for` between entities. Nonetheless, the "*alignment*" and "*NLI alignment*" pair are alike, and so are the pair "*probabilistic model*" and "*probabilistic model*". Exploiting the relationships between unlabeled data would integrate the information hidden in the text and make use of the large quantity of unlabeled data for semi-supervised learning.

To address the above limitations, we propose a semi-supervised method based on label propagation over a heterogeneous candidate graph to populate labels for the two tasks (see Figure 3). More specifically, we introduce a joint semi-supervised algorithm for the two tasks, where unannotated and annotated candidates (entities and relations) are treated as nodes in a heterogeneous graph, and labels are propagated across the graph through similarity-scored edges. Our framework *Jointprop* considers the interactions among the unlabeled data by constructing the graph using the union of labeled and unlabeled data into one learning diagram. We evaluate *Jointprop* on multiple benchmark datasets and our proposed framework achieve state-of-the-art results on both semi-supervised NER and RE tasks. To the best of our knowledge, this is the first work that performs semi-supervised learning for entity and relation extraction in a unified framework to leverage unannotated data for both tasks.

Our contributions are summarized as following:

- We propose a joint learning scheme using heterogeneous graph-based label propagation for semi-supervised NER and RE. The model exploits the interrelations between labeled and unlabeled data and the similarity between unlabeled examples from both tasks by propagating the information across a joint heterogeneous graph. To the best of our knowledge, this is the first work that combines semi-supervised NER and RE.

- We propose a unified semi-supervised framework for both entity and relation extraction. The framework generates candidate spans from the unlabeled data, automatically constructs a semantic similarity-based graph for all the candidates, and performs label propagation across the graph.

- We show that our proposed method can reliably generate labels for unlabeled data and achieve good performance under a limited data scenario. Our model outperforms strong baselines in two- and single-task settings and establishes new state-of-the-art F1 on benchmark datasets.

2 Related Work

Joint Entity and Relation Extraction Name Entity Recognition, and Relation Extractions are two essential problems in information extraction (Grishman, 1997). Exploiting their interrelationships, models that combine the identification of entities and relations have attracted attention. Conventional joint extraction systems combine the tasks in a pipelined approach (e.g., Ratinov and Roth (2009); Chan and Roth (2011); Luu et al. (2014, 2015); Tuan et al. (2016)): first identifying entities and employing the detected entity for relation extraction. However, they overlook their inherent correlation. Recent works have proposed coupling various IE

tasks to avoid error propagation issues. For example, joint extract entities and relations (Miwa and Sasaki, 2014; Li and Ji, 2014; Luu et al., 2016) or end-to-end multi-task learning (Luan et al., 2018a, 2019; Wadden et al., 2019; Lin et al., 2020; Zhang et al., 2017). Despite evidence of the efficiency of joint or multi-task learning, there is currently no framework that integrates semi-supervised learning for both tasks in a joint entity and relation extraction system.

Semi-supervised learning The Semi-Supervised learning seeks to enhance limited labeled data by leveraging vast volumes of unlabeled data (Søgaard, 2013) which mitigate data-hungry bottleneck and supervision cost. SSL has a rich history (Scudder, 1965). There have been substantial works in semi-supervised settings in NLP, such as bootstrapping (Gupta and Manning, 2014, 2015; Batista et al., 2015), co-training (Blum and Mitchell, 1998), distant supervision (Zeng et al., 2015; Yang et al., 2018), and graph-based methods (Delalleau et al., 2005; Subramanya and Bilmes, 2011; Subramanya et al., 2010; Luan et al., 2017).

In particular, graph-based SSL algorithms have gained considerable attention (Zhu and Ghahramani, 2002; Seeger, 2001; Delalleau et al., 2005). There are two underlying assumptions for the label propagation (Zhou et al., 2004). First, similar training samples are more likely to belong to the same class. Second, nodes in similar structures are likely to have the same label. Label propagation exploits these assumptions by propagating label information to surrounding nodes based on their proximity. The metric-based method had been applied in a graph-based SSL setting for its ability to infer labels for unseen classes directly during inference. For example, Luan et al. (2017) propagates the label based on estimating the posterior probabilities of unlabeled data. Meanwhile, Liu et al. (2019) sought to exploit the manifold structure of novel class space in a transitive setting.

3 Methodology

Problem Definition The input of the problem is a sentence $X = \{x_1, ..., x_n\}$ consisting of n tokens, from which we derive $S = \{s_1, ..., s_d\}$, the set of all possible within-sentence word sequence spans (up to length L) in the sentence. Let START(i) and END(i) denote the start and end indices of span s_i, $\mathcal{E}$ denote a set of predefined entity types, and $\mathcal{R}$ denote the set of relational types. The full data is defined as $D = (X, Y)$. In *Jointprop*, the goal is to learn from the small portion of labelled data D_l and generalize to the unlabelled portion of data D_u. The labelled data D_l and unlabelled data D_u are originally split from the training set D_{train}, where $D_l \cap D_u = \emptyset$.

The purpose of this work is to predict a possible entity type $y_e(s_i) \in \mathcal{E}$ for each span $s_i \in S$ while predicting a possible relation types $y_r(s_i, s_j) \in R$ for every pair of spans $s_i \in S, s_j \in S$ under SSL settings. The label can also be a 'null' label for a span (i.e. $y_e(s_i) = \epsilon$) or a span pair (i.e. $y_r(s_i, s_j) = \epsilon$). The output of the task are $Y_e = \{(s_i, e) : si \in S, e \in \mathcal{E}\}$ and $Y_r = \{(s_i, s_j, r) : s_i, s_j \in S, r \in \mathcal{R}\}$.

Model Overview Figure 2 illustrates an overview architecture of the proposed *Jointprop* framework. Our framework consists of 1) SPAN FEATURE GENERATION that learns the discriminative contextualized features for labelled data D_l and unlabeled span D_u; 2) HETEROGENEOUS GRAPH CONSTRUCTION which maps both labelled-unlabeled, labelled-labelled and unlabeled-unlabeled relationships for both entities and relations; 3) JOINT LABEL PROPAGATION which disseminates labels over the whole heterogeneous graph is produced by unlabeled nodes, and 4) MODEL DECODE AND FINETUNE MODULE that decodes and select the refined propagated pseudo labels to perform fine-tuning.

3.1 Span feature generation

Our feature extractor is a standard span-based model following prior work (Wadden et al., 2019; Luan et al., 2018a,b). For each input token x_k, we obtain contextualized representations $\mathbf{x}_k$ using a pre-trained language model (e.g., BERT (Devlin et al., 2019)). For the i-th span $s_k \in S$, the span representation $\boldsymbol{h}_e(s_i)$ is as follows:

$$\boldsymbol{h}_e(s_i) = [\mathbf{x}_{START(i)}; \mathbf{x}_{END(i)}; \phi(s_i)] \quad (1)$$

where $\phi(s_i) \in \mathbb{R}^{1 \times d_F}$ denotes the learned embeddings of span width features.

For each pair of spans input $s_i, s_j \in S$, the span pair representation is defined as:

$$\boldsymbol{h}_r(s_i, s_j) = [\boldsymbol{h}_e(s_i); \boldsymbol{h}_e(s_j); \mathcal{F}_{af}] \quad (2)$$

where $\mathcal{F}_{af} = \boldsymbol{h}_e(s_i) \cdot \boldsymbol{h}_e(s_j)$ refers to the entity affinity function of $e(s_i)$ and $e(s_j)$.

Both pairwise span feature $\boldsymbol{h}_r(s_i, s_j)$ and span feature $\boldsymbol{h}_e(s_i)$ will be fed into feedforward neural networks (FFNNs) respectively. The probability distribution of entity is denoted as

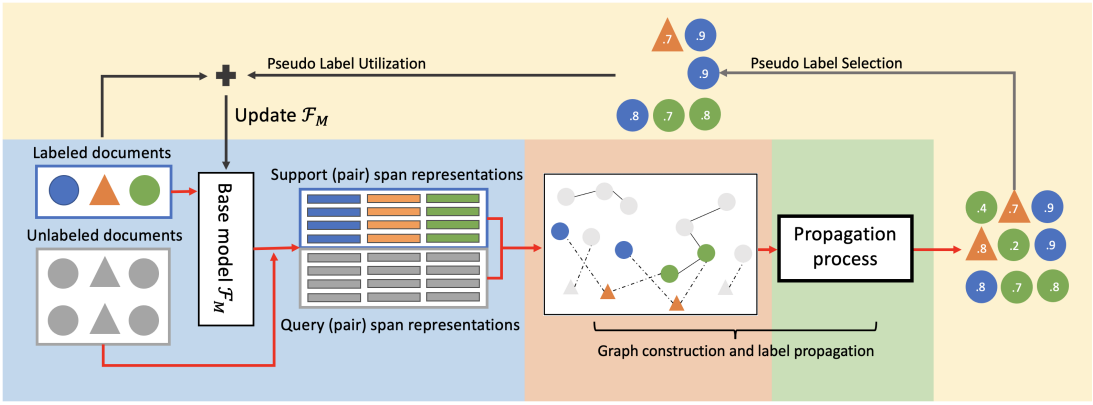


Figure 2: Overview of our proposed framework for semi-supervised joint learning. SPAN FEATURE GENERATION, HETEROGENEOUS GRAPH CONSTRUCTION, JOINT LABEL PROPAGATION are represented in **red arrows**, and MODEL OPTIMIZATION is illustrated in **black arrows**

$P_e(e|s_i), (e \in \mathcal{E} \cup \epsilon)$ and entity pairs is denoted as $P_r(r|s_i, s_j), (r \in \mathcal{R} \cup \epsilon)$.

The classification loss will be defined as:

$$\mathcal{L} = \sum_{t \in T} w_t \mathcal{L}_t \tag{3}$$

where w_t is the predefined weight of a task t and T is the total number of tasks.

We then use labelled data D_l to train the classifier C_l. The C_l generates contextualized span or span pair feature from Equation 1 and Equation 2 which converts unlabeled data D_u into unlabeled (query) entity presentation $\boldsymbol{h}_{u,e}$ or query entity pair representation $\boldsymbol{h}_{u,r}$. For labelled data D_l, we denote the C_l generated labelled (support) entity presentation as $\boldsymbol{h}_{l,e}$ and labelled entity pair representation as $\boldsymbol{h}_{l,r}$.

3.2 Joint Semi-supervised Learning

Heterogeneous Graph Construction We construct the heterogeneous graph to exploit the manifold structure of the class space and exploit the combination of labelled data D_l and unlabeled data D_u. Specifically, we examine the similarity relations among pairs of unlabeled data as well as the similarity relationships between the labelled data in order to take advantage of the smoothest constraints among neighbouring unlabelled data in our semi-supervised joint entity and relation extraction task.

For computational efficiency, we construct a k Nearest Neighbor (kNN) graph instead of a fully-connected graph. Let N be the number of labelled entity representations and let M be the number of unlabelled entity representations. Specifically, we take N entity representations and M unlabelled entity representations as nodes of an entity graph with size $T_e = N + M$. For the relation graph, we take span pair representation as nodes with size $T_r = ((N+M) \times (N+M))$. We construct a sparse affinity matrix, denoted as $\mathbf{A} \in \mathbb{R}^{\mathbf{T} \times \mathbf{T}}$, where $T_e, T_r \in \mathbf{T}$ by computing the Gaussian similarity function between each node:

$$\mathbf{A}_{ab} = exp(-\frac{||(\mathbf{h}_a, \mathbf{h}_b)||_2^2}{2\sigma^2}) \tag{4}$$

where $\mathbf{h}_a$ denotes the a-th entity representation or pairwise entity representation (i.e. $\{h_r(s_i, s_j), h_e(s_i), h_e(s_j)\} \in \mathbf{h}_a$). The σ is the length scale parameter.

Subsequently, we symmetrically normalize the non-negative and symmetric matrix $O = \mathbf{A} + \mathbf{A}^T$ by applying Normalized Graph Laplacian on O:

$$S = H^{(-1/2)} O H^{(-1/2)} \tag{5}$$

where H is a diagonal matrix with its (i, i)-value to be the sum of the i-th row of O.

For pairwise span representation $\boldsymbol{h}_r(s_i, s_j)$ is essentially a function of $\boldsymbol{h}_e(s_i)$ and $\boldsymbol{h}_e(s_ij)$. The entity nodes and the relation nodes are automatically associated via their representation.

Label propagation Based on the embedding space, we propose the use of transductive label propagation to construct a graph from the labelled support set and unlabeled set, and then propogate the labels based on random walks to reason about relationships in labelled and unlabeled sets. Figure 3 illustrates the whole process of heterogeneous graph-based propagation $\mathcal{G}$. The circle node is the entity span representation and the triangle node is the relation representation. We define a label matrix $Z \in \mathbb{R}^{V \times U}$ where U is either the size of entity

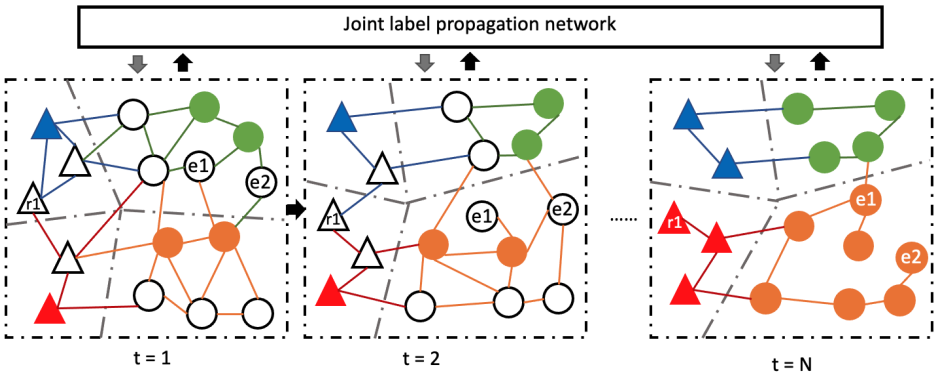


Figure 3: A conceptual demonstration of the label propagation process. Through the heterogeneous graph, our proposed joint semi-supervised learning method propagates labels to entity or relation candidates in the unlabeled data alternatively. As shown in the figure, the pseudo label for entities or relations will be refined every time t until converged.

types or relation types $U = \{\mathcal{E}; \mathcal{R}\}$. For label matrix, Z, the corresponding labelled data are one-hot ground truth labels and the rest are 0. Additionally, we denote Y_t as a representation of the predicted label distributions at iteration t. Initially, we set the rows in $Y_0 = Z$. Starting from Y_0, message passing via label propagation in an iterative manner selects the type of the span or span pairs in the unlabeled set D_u according to the graph structure according to the following operation:

$$Y_{t+1} = cSY_t + (1-c)Z \tag{6}$$

where $c \in (0, 1)$ controls the probability of information being obtained from a node's adjacency nodes or its initial label. Y_t refers to the predicted labels at time t.

Given $Y_0 = Z$, and equation (6) , we have:

$$Y_t = (cS)^{t-1}Z + (1-c)\sum_{i=0}^{t-1}(cS)^i Y \tag{7}$$

As the parameter $c \in (0, 1)$, taking the limit of equation (7) $(t- > \infty)$ we have:

$$\lim_{t\to\infty} Y_t = (1-c)(1-cS)^{-1} = Y_{converge} \tag{8}$$

The label propagation will converge to $Y_{converge}$.

3.3 Model optimization

After we obtain the $Y_{converge}$, we use the $softmax$ function followed by a standard `argmax` operation to determine the pseudo labels $\{\hat{y}\}$ for all the instances in the unlabeled set based on the final label probability matrix $Y_{converge}$. After generating the pseudo labels $\{\hat{y}\}$ for all the labelled data D_l, we filter those of lower quality with a confidence threshold of g and combine the rest (of confidence above the threshold) with the labelled data D_l to retrain the classification model:

$$\{\hat{y}\} = \{\hat{y} \mid \text{confidence}(y) \geq g\}$$
$$(X, Y) = (X, Y)_{D_l} + \{(x, \hat{y}) | x \in D_u\}$$

As shown in the Figure 2, the final step in our proposed joint semi-supervised learning framework is re-training. The retraining model remains the same as the baseline model, as does the joint NER-RE classification function.

4 Experiments

We evaluate the effectiveness of *Jointprop* against models from two lines of work: semi-supervsied NER and semi-supervsied RE. We also provided a detailed analysis to demonstrate the benefits of our framework. For implementation details and dataset descriptions please refer to Appendix A and Appendix B.

Datasets We perform experiments to assess the efficacy of our framework on four public datasets: SciERC (Luan et al., 2018b), ACE05 (Walker et al., 2006), SemEval (Hendrickx et al., 2010) and ConLL (Tjong Kim Sang and De Meulder, 2003).

4.1 Main Results

Tables 1 to 4 provide the framework performance on the joint entity and relation extraction task, the NER task, and the RE task, respectively. Note that *Beforeprop* only trains using the labelled corpus. (i.e., The *Beforeprop* only trains with 5%, 10% and 30% training data.) As no unlabeled data are used in the training, this indicates the lower bound performance and establishes a new baseline.

Settings% labeled Data	Task	5%			10%			20%			30%		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Beforeprop(baseline)	NER	46.78	47.25	47.01	52.44	59.80	55.94	55.80	62.37	58.90	60.42	67.56	63.79
	RE	20.89	15.40	17.73	35.75	16.74	22.80	38.68	23.51	29.25	43.41	29.77	35.32
Jointprop	NER	**52.67**	**48.46**	**51.02**	**60.15**	**61.95**	**61.04**	**62.03**	**64.52**	**63.25**	**66.55**	**65.73**	**66.19**
	RE	**40.82**	**33.78**	**36.97**	**44.42**	**26.34**	**39.98**	**44.55**	**45.28**	**44.91**	**57.94**	**39.32**	**46.85**

Table 1: Performance on SciERC with various amount of labeled data.

Settings% labeled Data	Task	5%			10%			20%			30%		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Beforeprop (baseline)	NER	78.32	76.88	77.59	80.81	81.68	81.24	81.01	85.17	83.04	84.51	86.98	85.72
	RE	46.33	20.85	28.76	49.10	30.76	37.82	46.71	46.31	46.51	57.59	48.78	52.83
Jointprop	NER	**81.91**	**78.39**	**80.11**	**83.38**	**82.76**	**83.07**	**86.82**	**83.76**	**85.27**	**87.69**	**86.40**	**87.04**
	RE	**48.10**	**29.63**	**36.67**	**48.89**	**36.23**	**42.00**	**60.26**	**44.67**	**51.30**	**61.54**	**48.65**	**54.34**

Table 2: Performance on ACE05 with various amounts of labelled data.

Methods / % labeled Data	5%			10%			30%		
	P	R	F1	P	R	F1	P	R	F1
Mean-Teacher	70.33	68.55	69.05	74.01	72.08	73.37	79.09	82.23	80.61
Self-Training	73.10	70.01	71.34	75.54	73.00	74.25	80.92	82.39	81.71
DualRE	73.32	77.01	74.35	75.51	78.81	77.13	81.30	84.55	82.88
MRefG	73.04	78.29	75.48	76.32	79.76	77.96	81.75	84.91	83.24
MetaSRE	75.59	81.40	78.33	78.05	82.29	80.09	82.01	87.95	84.81
GradLRE	75.96	83.72	79.65	78.90	82.94	81.69	82.74	88.49	85.52
Jointprop†	**76.09**	**86.35**	**80.89**	**79.10**	**88.64**	**83.60**	**83.62**	**89.35**	**86.39**
Gold labels	-	-	84.64	-	-	84.40	-	-	87.08

Table 3: Performance on SemEval with various labelled data and 50% unlabeled data. We provide the *Gold labels* serves as the upper bound of the model. († indicates our framework.)

Methods / % labeled Data	5%			10%			30%		
	P	R	F1	P	R	F1	P	R	F1
VSL-GG-Hier	84.13	82.64	83.38	84.90	84.52	84.71	85.37	85.67	85.52
MT + Noise	83.74	81.49	82.60	84.32	82.64	83.47	84.98	84.78	84.88
Semi-LADA	86.93	85.74	86.33	88.61	88.95	88.78	89.98	90.52	90.25
Jointprop†	**89.88**	**85.98**	**87.68**	**88.76**	**90.25**	**88.89**	**91.16**	**90.58**	**90.87**

Table 4: Performance on CoNLL 2003 with various labelled data. († indicates our framework.)

Results on SciERC Table 1 illustrate our main results on semi-supervised joint learning on the SciERC dataset. We observed *Jointprop* improve significantly on both entity recognition and relation extraction. *Jointprop* achieves 3.97% and 15.89% F1 improvements, respectively, comparing to *Beforeprop*. This improvement validates the robustness of *Jointprop* by performing joint learning on NER and RE.

Results on ACE05 Table 2 we summarize the results of comparing to the baseline performance. As can be seen from the table, *Jointprop* improves by around 2% and 5% on F1 for entity recognition and relation extraction task respectively. The results of this study provide further evidence of the consistency of the framework for multitask datasets.

Results on SemEval Table 3 summarizes the experimental results on the SemEval dataset using various labelled data and 50% unlabeled data. *Jointprop* improves on the *Beforeprop* by 5.47% on average. We can observe that *Jointprop* attains 1.24%, 1.91% and 0.81% F1 improvements over the state-of-the-art model GradLRE (Hu et al., 2021b) with 5%, 10% and 30% training data. Moreover, the model's performance consistently improves while narrowing down the gap towards the upper bound as the proportion of labelled data increases. *Jointprop* establishes a new state-of-the-art result, indicating that our framework is relatively robust even when performing a single task: semi-supervised

RE.

Results on CoNLL Experimental results on CoNLL dataset are shown in Table 4. Semi-LADA (Chen et al., 2020) is the current state-of-the-art semi-supervised NER model. In multiple training data settings, *Jointprop* achieves an average improvement of 0.9% over Semi-LADA. Semi-LADA reports a 91.83% F1 score in a fully supervised setting, as the upper bound of the semi-supervised model. *Jointprop* achieves 90.87% in F1 score with 30% of training data. The difference between the upper bound and the model performance narrows to less than 1%. Moreover, *Jointprop* surpasses the current state-of-the-art semi-supervised NER model, showing our model's effectiveness on another single task: semi-supervised NER.

4.2 Analysis

4.2.1 Ablation Studies

This section provides comprehensive ablation studies to show the efficacy of *Jointprop* frameworks. Tables 5 and 7 show the effect of joint label propagation on single-task (NER or RE) prediction accuracy. *w/o REprop* denotes ablating the relation propagation while *w/o NERprop* denotes ablating the entity propagation. As a lower bound to the framework, we provide the *Beforeprop* result, which is the base model without any propagation. As shown in Table 5, although *w/o REprop* achieved an average 0.85% improvement on F1 compared to *Beforeprop*. The *Jointprop* further improve the performance significantly by 4.01%, 4.98%, 3.65% and 2.19% across 5%, 10%, 20% and 30% training data, respectively. From Table 7, we observed that *w/o REprop* attain an average of 2.94% performance gain in F1 compared to *Jointprop*. Though *w/o REprop* shows its effectiveness, *w/o NERprop* has 7.03% further overall across different proportions of training data. In general, we observe that joint label propagation is very helpful to *Jointprop* performance, especially for relation extraction tasks.

We investigate a real and illustrative example in Figure 1. Given sentences S1 to S3. *w/o REprop* is unable to identify the label of "alignment" in S2 and "NLI alignment" in S3. Moreover, *w/o NERprop* tends to miss predict the pseudo label as `no_relation`. More specifically, in annotated S1, the entity "dependency parsing" has no direct link to the entity "alignment" in S2 and entity "NLI alignment" in S3. Consequently, *w/o REprop* makes the wrong prediction. Similar to *w/o REprop*, the relation indicator "uses..to" in annotated S1 is semantically similar to "used in" in S2 but not akin to "apply..for.." in S3, hence *w/o NERprop* miss identify the label of r'''. Whereas *Jointprop* can assign the correct pseudo label to entities and relations in all three sentences for it benefits from the shared information from NER and RE. The results indicate that our framework *Jointprop* could leverage the interactions across the two tasks and derive useful information from a broader context. Therefore achieve significant improvement across NER and RE tasks.

4.2.2 Case study

We perform a case study examining our framework's performance on four sentences (i.e., S1, S2, S3, and S4) in comparison to the benchmark models Semi-LADA and GradLRE. Semi-LADA performs semi-supervised NER task while GradLRE performs semi-supervised RE task. Meanwhile *Jointprop* performs the semi-supervised style joint for NER and RE.

S1 has a simple structure, and all three models correctly classify the label for relation and entity. For S2, the GradLRE misclassifies the "Statistical machine translation" entity as Task. Most of the labelled samples with given entity pair are likely as in (e1: `Method`, e2: `Task`), plus there is a relation indicator "in order to," which misguides the GradLRE into the wrong prediction. Similarly, in S4, Semi-LADA predicts the entity as `Generic`, the dominant class in the training set. *Jointprop* can assign the correct label without being sensitive to the label distribution in the training data.

Moreover, Semi-LADA fails to recognize the entity "correlation of dependency relation paths" in S3, while GradLRE cannot identify the relation `Used-for`. One possible reason is that there were not many similar long sequences in the training data. Consequently, Semi-LADA is insufficient in entity learning, especially for long lines, while the GradLRE fails to establish edges with samples in the training set. *Jointprop* not only builds a connection between labelled and unlabeled data but also within labelled/unlabeled data. The extra connections hence help our model to make the correct prediction.

4.2.3 Qualitative Analysis

Table 8 shows the qualitative results of our proposed method Joint Semi-supervised Learning for

Model / Task	Name Entity Recognition											
	5%			10%			20%			30%		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Beforeprop	46.78	47.25	47.01	52.44	59.80	55.94	55.80	62.37	58.90	60.42	67.56	63.79
w/o REprop	51.82	45.10	48.23	58.92	53.46	56.06	61.55	57.77	59.60	64.71	63.32	64.01
Jointprop	**52.67**	**48.46**	**51.02**	**60.15**	**61.95**	**61.04**	**62.03**	**64.52**	**63.25**	**66.55**	**65.73**	**66.19**

Table 5: Ablation study on pure NER task on SciERC dataset.

Model / Task	Relation Extraction											
	5%			10%			20%			30%		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Beforeprop	20.89	15.40	17.73	35.75	16.74	22.80	38.68	23.51	29.25	43.41	29.77	35.32
w/o NERprop	38.92	13.35	19.88	19.20	44.97	26.91	22.27	62.53	32.74	32.12	44.56	37.33
Jointprop	**40.82**	**33.78**	**36.97**	**44.42**	**26.34**	**39.98**	**44.55**	**45.28**	**44.91**	**57.94**	**39.32**	**46.85**

Table 6: Ablation study on pure RE task on SciERC dataset.

Sentence	**Semi-LADA**	**GradLRE**	***Jointprop***
S1: We propose a Cooperative Model for natural language understanding in a dialogue system.	e1: Method e2: Task R: -	e1: - e2: - R: Used-for	e1: Method e2: Task R: Used-for
S2: We address appropriate user modelling in order to generate cooperative responses to each user in spoken dialogue systems.	e1: Method e2: Task R: -	e1: - e2: - R: Used-for (x)	e1: Method e2: Task R: Part-of
S3: We explore correlation of dependency relation paths to rank candidate answers in answer extraction.	e1: - (x) e2: Task R: -	e1: - e2: - R: no_relation (x)	e1: OST e2: Task R: Used-for
S4: We present a syntax-based constraint for word alignment, known as the cohesion constrain.	e1: Generic (x) e2: Generic (x) R: -	e1: - e2: - R: Hyponym-of	e1: OST e2: OST R: Hyponym-of

Table 7: Case study of *Jointprop*. The red marked span denotes the head ($e1$) entity while the blue marked span represents the tail ($e2$) entity. Semi-LADA performs `OtherScientificTerm` abbreviated as `OST`. (x) indicates the wrong prediction and - means the model does not have certain predictions (i.e., The model does not predict entity type or relation type).

Entity and Relation Extraction with Heterogeneous Graph-based Propagation. We show the performance of the propagated pseudo labels with the ground truths under 10% split training set on ACE05 dataset. As we can see from the performance Table 8, in both NER and RE, the recall of the predictions indicates that most of the positive candidates have been propagated a positive label.Meanwhile, the precision of the predictions for the NER task is also high. However, the precision for the RE task is low, showing that almost half of the null candidates have been assigned a positive label. The propagation of RE tasks is still quite challenging.

%	P	R	F1
NER	86.23	92.78	89.34
RE	52.17	98.82	68.57

Table 8: Qualitative results of our method in 10% split on ACE05 dataset. (Average F1)

In spite of this, our method still generally produces more accurate predictions. Given a sentence in ACE05: 'Although the Russian government...'. Our model prediction for the phrase "Russian government" is "Organization", which is more accurate than the ground truth GPE-Geographic Entities.

5 Conclusion

In this paper, we propose a novel heterogeneous graph-based propagation mechanism for joint semi-supervised learning of entity and relation extraction. For the first time, we explore the interrelation between different tasks in a semi-supervised learning setting. We show that the joint semi-supervised learning of two tasks benefits from their codependency and validates the importance of utilizing the shared information between unlabeled data. Our experiments show that combining the two tasks boost the model performance. We also evaluate two public datasets over competitive baselines and achieve

state-of-the-art performance. We also conduct ablation studies of our proposed framework, which demonstrate the effectiveness of our model. We further present case studies of our model output.

6 Limitations

May extend to other domains In this paper, we present a generic framework and evaluate the effectiveness of our proposed model *Jointprop* on three public datasets. We may further extend the framework to various datasets in different domains. For example, ACE05 (Walker et al., 2006) in social networks, journalism, and broadcasting, as well as GENIA corpus (Ohta et al., 2002) in biomedical research.

May extend to other NLP tasks Our proposed model focus on two tasks, namely NER and RE. We may extend our framework to include more information extraction tasks, such as coreference resolution and event extraction. Moreover, we may contract knowledge graphs from extracted structural information.

Acknowledgment

This research is supported by Nanyang Technological University, under SUG Grant (020724-00001)

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

David S. Batista, Bruno Martins, and Mário J. Silva. 2015. Semi-supervised bootstrapping of relationship extractors with distributional semantics. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 499–504, Lisbon, Portugal. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT' 98, page 92–100, New York, NY, USA. Association for Computing Machinery.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 551–560, Portland, Oregon, USA. Association for Computational Linguistics.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1241–1251, Online. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel. 2018. Variational sequential labelers for semi-supervised learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 215–226, Brussels, Belgium. Association for Computational Linguistics.

Olivier Delalleau, Yoshua Bengio, and Nicolas Le Roux. 2005. Efficient non-parametric function induction in semi-supervised learning. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, volume R5 of Proceedings of Machine Learning Research, pages 96–103. PMLR. Reissued by PMLR on 30 March 2021.

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.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In Information Extraction, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 11–27. Springer Verlag. International Summer School on Information Extraction, SCIE 1997 ; Conference date: 14-07-1997 Through 18-07-1997.

Sonal Gupta and Christopher Manning. 2014. Improved pattern learning for bootstrapped entity extraction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pages 98–108, Ann Arbor, Michigan. Association for Computational Linguistics.

Sonal Gupta and Christopher D. Manning. 2015. Distributed representations of words to guide bootstrapped entity classifiers. In Proceedings of the

2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1215–1220, Denver, Colorado. Association for Computational Linguistics.

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

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. 2021a. Semi-supervised relation extraction via incremental meta self-training. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 487–496, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021b. Gradient imitation reinforcement learning for low resource relation extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

Pooja Lakshmi Narayan, Ajay Nagesh, and Mihai Surdeanu. 2019. Exploration of noise strategies in semi-supervised named entity classification. In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), pages 186–191, Minneapolis, Minnesota. Association for Computational Linguistics.

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 (Volume 1: Long Papers), pages 402–412, Baltimore, Maryland. Association for Computational Linguistics.

Wanli Li, Tieyun Qian, Xu Chen, Kejian Tang, Shaohui Zhan, and Tao Zhan. 2021. Exploit a multi-head reference graph for semi-supervised relation extraction. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–7.

Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. In The World Wide Web Conference, WWW '19, page 1073–1083, New York, NY, USA. Association for Computing Machinery.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7999–8009, Online. Association for Computational Linguistics.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. 2019. Learning to propagate labels: Transductive propagation network for few-shot learning. In International Conference on Learning Representations.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018a. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018b. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In Proc. Conf. Empirical Methods Natural Language Process. (EMNLP).

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2641–2651, Copenhagen, Denmark. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. 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 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Anh Tuan Luu, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy construction using syntactic contextual evidence. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 810–819.

Anh Tuan Luu, Jung-jae Kim, and See Kiong Ng. 2015. Incorporating trustiness and collective synonym/contrastive evidence into taxonomy construction. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1013–1022.

Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 403–413.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1858–1869, Doha, Qatar. Association for Computational Linguistics.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The genia corpus: An annotated research abstract corpus in molecular biology domain. In Proceedings of the Second International Conference on Human Language Technology Research, HLT '02, page 82–86, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

H. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory, 11(3):363–371.

Matthias Seeger. 2001. Learning with labeled and unlabeled data.

Amarnag Subramanya and Jeff Bilmes. 2011. Semi-supervised learning with measure propagation. Journal of Machine Learning Research, 12(102):3311–3370.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, page 167–176, USA. Association for Computational Linguistics.

Anders Søgaard. 2013. Semi-supervised learning and domain adaptation in natural language processing. Synthesis Lectures on Human Language Technologies, 6:1–103.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147.

Luu Anh Tuan, Siu Cheung Hui, and See Kiong Ng. 2016. Utilizing temporal information for taxonomy construction. Transactions of the Association for Computational Linguistics, 4:551–564.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

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

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442–6454, Online. Association for Computational Linguistics.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2159–2169, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Deming Ye, Yankai Lin, and Maosong Sun. 2021. Pack together: Entity and relation extraction with levitated marker. CoRR, abs/2109.06067.

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, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-end neural relation extraction with global optimization. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1730–1740, Copenhagen, Denmark. Association for Computational Linguistics.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In Advances in Neural Information Processing Systems, volume 16. MIT Press.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation.

A Experimental Settings

Framework We show our overall framework *Jointprop* in Figure 2. Following (Liu et al., 2019), the hyper-parameter c in Equation 6 is set to 0.99. According to our empirical findings, the best values for the settings of k and σ in graph construction in Section 3.2 are varied in datasets. We select σ as two and the k as 50. Meanwhile, We adopt the affinity function $\mathcal{F}_{af}$ with all the generated spans between relation spans. Moreover, we perform average pooling for them. The optimal hyperparameters and settings are selected based on the model's performance.

Training We employ the BERT-cased as an encoder for SemEval and ConLL datasets and adopt the SciBERT-SCIVOCAB-cased (Beltagy et al., 2019) encoder for the SciERC dataset as suggested in (Luan et al., 2019). The rest will be treated as an unlabeled set. To maximize the loss, we use BERTAdam with a 1e-3 learning rate. The maximum span width is set at 8.

B Datasets and baselines

B.1 Dataset implementation

For semi-supervised joint task, we consider SciERC (Luan et al., 2018b) and ACE05(Walker et al., 2006) datasets and follow the pre-processing steps in (Wadden et al., 2019). For a single task, we conduct experiments against models from two types of work: semi-supervised NER, and semi-supervised RE. For the semi-supervised NER task, we consider ConLL 2003 (ConLL) (Tjong Kim Sang and De Meulder, 2003) and adopt the pre-processing in (Chen et al., 2020). For semi-supervised RE we evaluate our approach on SemEval 2010 Task 8 (SemEval) (Hendrickx et al., 2010) dataset and adopt the pre-processing in (Hu et al., 2021b). Note that the entity mentioned in the sentences in the **SemEval** has been identified and marked in advance.

Table 9 shows the statistics of each dataset.

Dataset	Sentences			Types	
	Train	Dev	Test	# E	# R
SciERC	1861	275	551	6	7
ACE05	10051	2424	2050	7	6
ConLL	14,987	3466	3684	4	-
SemEval	7199	800	1864	-	19

Table 9: **Statistics for the SciERC, ACE05, ConLL and SemEval datasets**. #E: Number of entity classes. #R: Number of relation classes.

Data split for semi-supervised settings We follow split settings in (Chen et al., 2020), (Hendrickx et al., 2010) respectively for ConLL and SemEval and generate different proportions (5%, 10% and 30%) of training data to investigate how training set size impacts performance and to retain the original development set and test set for evaluation purposes. Noted that we sample 50% of the training set as the unlabeled set as (Hendrickx et al., 2010) for fair comparisons. For ACE05 and SciERC datasets, we split the training data based on documents and generate 5%, 10%l 20% and 30% of training data. In particular, we endeavour to ensure that each proportion of data contains as many types of entity types and relation types possible.

B.2 Evaluation Metrics

We consider the same criteria to apply as previous works (Hu et al., 2021b,a; Li et al., 2021; Lin et al., 2019) where precision and recall serve as supplementary metrics, while F1 score serves as the primary evaluation metrics. Note that the evaluation excludes the accurate prediction for `no_relation`.

B.3 Compared baselines

Semi-supervised joint learning For joint learning, because there is no prior study on semi-supervise joint learning, we use DYGIE++ (Wadden et al., 2019) (i.e. *Beforeprop*) as our baseline model to train.

Semi-supervised NER In order to show that our *Jointprop* framework works with unlabeled data, we compared it to three recent state-of-the-art semi-supervised NER models that were already in use:

- **VSL-GG-Hier** (Chen et al., 2018) introduced a hierarchical latent variables models into semi-supervised NER learning.
- **MT + Noise** (Lakshmi Narayan et al., 2019) explored different noise strategies including word-dropout, synonym-replace, Gaussian noise and network-dropout in a mean-teacher framework.
- **Semi-LADA** (Chen et al., 2020) proposes a local additivity based data augmentation method which uses the back-translation technique.

Semi-supervised RE We compared our *Jointprop* framework with the following 6 representative semi-supervised relation models:

- **Mean-Teacher** promotes the model's variants to generate consistent predictions for comparable inputs.
- **DualRE** (Lin et al., 2019) trains a prediction and retrieval module in conjunction to choose samples from unlabeled data.
- **MRefG** (Li et al., 2021) constructs reference graphs to semantically relate unlabeled data to labelled data.
- **MetaSRE** (Hu et al., 2021a) constructs pseudo labels on unlabeled data using meta-learning from the successfulness of the classifier module as an extra meta-objective.

- **GradLRE** (Hu et al., 2021b) is the state-of-the-art approach that encourages pseudo-labeled data to mimic the gradient descent direction on labelled data and boost its optimization capabilities via trial and error.
- **Gold labels** train annotated (i.e., sampled 5%, 10% or 30% training data) and unlabeled data with their gold labels indicating the model upper bound.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.