Unsupervised Domain Adaptation for Joint Information Extraction

Nghia Trung Ngo[1], **Bonan Min**[2*] **and Thien Huu Nguyen**[1]
[1] Department of Computer Science, University of Oregon, Eugene, OR, USA
[2] Amazon AWS AI Labs
{nghian@,thien@cs}.uoregon.edu,
bonanmin@amazon.com

Abstract

Joint Information Extraction (JIE) aims to jointly solve multiple tasks in the Information Extraction pipeline (e.g., entity mention, event trigger, relation, and event argument extraction). Due to their ability to leverage task dependencies and avoid error propagation, JIE models have presented state-of-the-art performance for different IE tasks. However, an issue with current JIE methods is that they only focus on standard supervised learning setting where training and test data comes from the same domain. Cross-domain/domain adaptation learning with training and test data in different domains have not been explored for JIE, thus hindering the application of this technology to different domains in practice. To address this issue, our work introduces the first study to evaluate performance of JIE models in unsupervised domain adaptation setting. In addition, we present a novel method to induce domain-invariant representations for the tasks in JIE, called **D**omain **A**daptation for **J**oint **I**nformation **E**xtraction (DA4JIE). In DA4JIE, we propose an Instance-relational Domain Adaptation mechanism that seeks to align representations of task instances in JIE across domains through a generalized version of domain-adversarial learning approach. We further devise a Context-invariant Structure Learning technique to filter domain-specialized contextual information from induced representations to boost performance of JIE models in new domains. Extensive experiments and analyses demonstrate that DA4JIE can significantly improve out-of-domain performance for current state-of-the-art JIE systems for all IE tasks.

1 Introduction

An information extraction (IE) system extracting structured information from unstructured text typically involves four major tasks: event trigger detection (ETD), event argument extraction (EAE), entity mention extraction (EME), and relation extraction (RE). Recently, the advance of large-scale pre-trained language model has made it possible to replace the classical pipeline approaches (Li et al., 2013; Chen et al., 2015), which suffer from error propagation, with a single transformer-based model performing all four tasks jointly, i.e. Joint Information Extraction approaches (JIE) (Lin et al., 2020; Nguyen et al., 2021). While effective in standard supervised learning scenario, these modern JIE systems fail to address the practical setting where training data (i.e., the source) and testing data (i.e., the target) come from different domains with different distributions. Such discrepancies pose a major challenge due to both the intrinsic variations of linguistics (e.g., lexical and semantic shifts) as well as extrinsic factors such as how textual datasets are collected and annotated. The problem is further exacerbated when the models aim to jointly learn multiple tasks, facing various kinds of domain shifts simultaneously. For example, in a *Die* event where a *Person* entity mention is

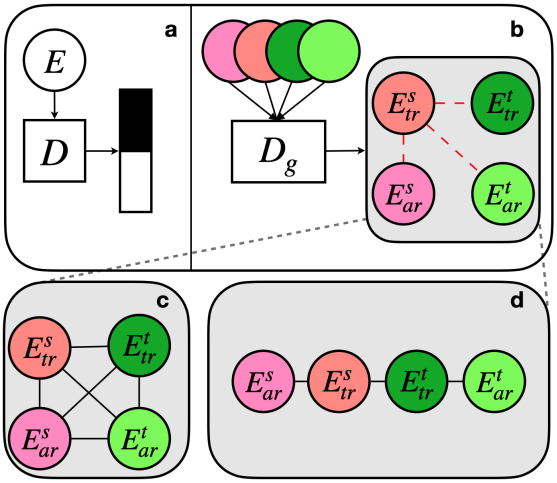

Figure 1: Top figures demonstrate the difference between DANN (**a**) and IrDA (**b**). Bottom figures are the relation graphs following the strict uniform alignment of standard DA methods (**c**) and the *chain* connection of IrDA (**d**).

[*] Work done at Raytheon BBN Technologies (prior to joining AWS AI).

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5894–5905
December 7-11, 2022

a *Victim* event argument, documents recording this type of event in medical records may express these instances in a significantly distinct manner compared to when new anchors report similar tragic incidents.

To address domain difference for IE, a major approach involve unsupervised domain adaptation (UDA) where models leverage additional unlabeled data in target domain together with labeled training set from source domain to improve the performance on target domain. As such, the majority of existing UDA methods have focused on transfer learning between source and target domains for a single IE task (Long et al., 2015; Ganin et al., 2016; Kumar et al., 2018). While recent work also aims to generalize previous approaches to multi-domains setting (Dai et al., 2020; Wright and Augenstein, 2020), the scenario where the considered task involves multiple objectives (as in JIE) with different input distributions remains unexplored. In particular, classical UDA approaches for IE often rely on simplification assumption on the factorization of the joint input-output distribution of an IE task to categorize and solve a specific domain shift problem. An example includes covariate shift where the discrepancy is assumed to be only in the marginal input distribution whereas predictive dependency remains unchanged (Kull and Flach, 2014). However, in JIE, this assumption does not hold and thus necessitates new domain adaptation methods to address UDA for JIE.

To this end, our work introduces a new UDA method for JIE, called DA4JIE. At the core of DA4JIE is an Instance-relational Domain Adaptation (IrDA) module that seeks to simultaneously align instance representations for all downstream tasks in JIE in the source and target domains. Inspired by Graph-relational Domain Adaptation (GrDA) proposed by (Xu et al., 2022) for heterogeneous domain adaptation, we view event trigger and entity mention instances of each domain as domain nodes on a domain-instance relational graph, whose adjacency matrix controls the relationship between domain-specific representations (Fig. 1a). In particular, an edge connecting two instances implies that their representations should be aligned, which is equivalent to their pairwise relationship containing no information to identify their domains. This is achieved by an adversarial learning process on pairwise node relationships. Specifically, a graph discriminator is employed to recover the domain-instance graph via the adjacency structure. Conversely, the text encoder for JIE would prevent the discriminator from doing so. IrDA is a generalization of the standard domain-adversarial training method (Ganin et al., 2016) that enforces strict uniform alignment (fully-connect relational graph) as depicted in Fig. 1c. In contrast, our approach assumes a *chain* connection across instance nodes (Fig. 1d) that reflects the true relationship among instance types, allowing flexible and effective adaptation to new domains for JIE.

In addition, to improve task performance, previous JIE systems have leveraged specialized linguistic structures extracted from input sentences in a heuristic and direct manner, e.g., using heuristic-based dependency graphs between instances in different tasks in JIE (Lin et al., 2020; Veyseh et al., 2020b; Nguyen et al., 2021). However, this approach is not suitable for domain adaptation as it further introduces more domain-specific context-dependency information into the learned representations. To address this problem, we incorporate a novel Context-invariant Structure Learning module (CiSL) into the instance encoding process. CiSL uses graph transformer networks (GTN) (Yun et al., 2019) to fuse different types of context-independent graphs into a single context-invariant graph (CiG) for each input sentence. Here, instance node features are combined with contextual representation to encourage the model to use domain-invariant information for downstream tasks. In addition, by viewing each input sentence as a graph with word-level nodes to induce word representations, we obtain richer instance representations for JIE by aggregating word-level representations. As such, our method also proposes a novel a CiG-conditioned pooling operation to enhance instance representations for classification tasks and boost the overall adaptation performance for JIE.

Finally, we provide extensive evaluation of the proposed UDA method for JIE on the ACE-05 dataset (Walker et al., 2005). The experimental results demonstrate the advantages of DA4JIE that achieves state-of-the-art (SOTA) performance when being adapted to multiple target domains.

2 Related Work

2.1 Joint Information Extraction

Classical methods for IE manually engineered linguistic features to capture the dependency between IE tasks, including Integer Linear Programming

for Global Constraints (Roth and Yih, 2004), Structured Perceptron (Miwa and Sasaki, 2014; Judea and Strube, 2016), and Graphical Models (Yu and Lam, 2010; Yang and Mitchell, 2016). The advance of deep learning and large-scale language models (Devlin et al., 2019) has greatly enhanced the representation ability of modern IE models, enabling them to jointly solve multiple tasks via the shared contextual embeddings. These joint models focused on different sets of IE tasks such as EME and RE (Zheng et al., 2017; Fu et al., 2019; Luan et al., 2019; Veyseh et al., 2020a), and ETD and EAE (Nguyen et al., 2016; Zhang et al., 2019; Nguyen and Nguyen, 2019). Recently, some efforts have been made to address the four tasks all together by introducing specialized structures and regularizations to model the joint instance distribution across tasks (Lin et al., 2020; Nguyen et al., 2021, 2022). Our work continues in their direction, but in UDA setting which is more difficult but also much more practical compared to the standard supervised learning setting.

2.2 Unsupervised Domain Adaptation

The main line of research on UDA approaches the domain shift problem by learning domain-invariant representations, which is either achieved by explicitly reducing the distance between source and target feature space measured by some distribution discrepancy metric (Long et al., 2015; Zellinger et al., 2017), or by adversarial training in which the feature extractor is trained to fool a domain classifier, both are jointly optimized to arrive at an aligned feature space (Ganin et al., 2016). We focus on applying the latter in transformer-based model (BERT) for IE tasks. In particular, there has been several prior works addressing UDA setting for a singular IE task, including event trigger identification (Naik and Rosé, 2020), event detection (Ngo et al., 2021; Trung et al., 2022), and relation extraction (Fu et al., 2017). However, a method specifically tackles joint task learning in UDA is still absent from the literature to the best of our knowledge. Our IrDA is the first to explicitly take into account multiple representations of different tasks when transferring between source and target domains.

3 Model

3.1 Problem Statement

The JIE problem composes of four tasks EME, ETD, RE, and EAE. Given an input sentence, a unified model is used to optimized a linear combination of each task objective. In particular, EME aims to detect and classify entity mentions (names, nomials, pronouns) according to a set of predefined (semantic) entity class (e.g., Person). Similarly, ETD seeks to identify and classify event triggers (verbs or normalization) that clearly evoke an event in a given set of event classes (e.g., Attack). Note that event triggers can involve multiple words. Next, RE objective is to predict the semantic relationship between two entity mentions in the sentence. Finally, in EAE, given an event trigger, the systems need to predict the roles that each entity mention plays in the corresponding event. Entity mentions are thus also called event argument candidates in this work. Noted that the sets of relations and roles are pre-determined and include a special class of *None* to indicate negative category.

In UDA setting, data comes from two different domains. For training, we have a labeled source dataset $\mathbf{S}$ consisted of N^s samples and an unlabeled set $\mathbf{T}$ of N^t samples drawn from target domain. The goal is to leverage both datasets to optimize model performance on test data from target domain. At each iteration, a mini-batch consists of samples from both $\mathbf{S}$ and $\mathbf{T}$ is sampled, the former are used to learn the main downstream tasks using their true labels, while the latter are employed to impose a domain-invariant constraint on the extracted features.

3.2 JIE Architecture

The following encoding process is applied to data from both domains, thus we omit the domain index in notations for brevity. Given an input sentence $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ with n words, the model first identifies span of an instance, which can be an entity mention or an event trigger, in $\mathbf{w}$ and then compute its representation for downstream tasks. In particular, following Lin et al. (2020), two conditional random field (CRF) layers, one for event triggers and another for event mentions, take in as input word-level contextual representation sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ ($\mathbf{x_i} \in \mathbb{R}^{\mathbf{h}}$ is obtained by averaging the word-pieces' hidden vectors of w_i returned by the transformer encoder, e.g., BERT). The CRFs output the best BIO tag se-

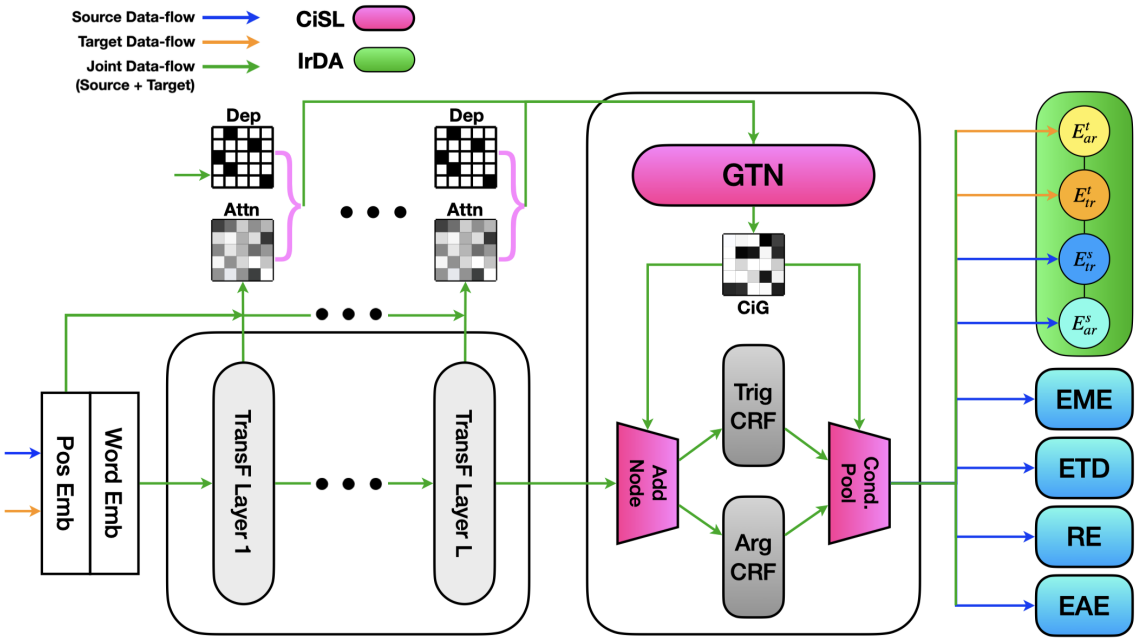

Figure 2: The overall architecture of our framework DA4JIE. First, input sentences from both source and target domains go through the same transformer encoder to compute their contextual representations. Concurrently, the **CiSL** module (pink) extracts the attention probability matrices at each layer to create attention graphs, using position embeddings as node features. These graphs are used to augment the dependency graphs, which are then fused across layers by a GTN to create a context-invariant graph. The node features of which are combined with the contextual representations as input for the instance span detection task using CRF layers. Next, the instance representations are computed based on the outputted spans conditioned on the context-invariant graph. Finally, source instances are used to optimize the encoder for the main JIE tasks (blue), while the **IrDA** module (green) takes the representations from the corresponding instance types for the nodes in the type-relational graph to calculate the discriminator loss.

quences (Chiu and Nichols, 2016) to indicate event trigger and entity mention/event argument spans (i.e., no label prediction yet) in w, which are then used to compute their representations $\mathbf{E}_{tr}$ and $\mathbf{E}_{ar}$ (each can contain multiple instances) by aggregating information from words in the corresponding spans. Finally, separate task-specific feed-forward networks are used to calculate label scores from $\mathbf{E}_{ar}$, $\mathbf{E}_{tr}$, $(\mathbf{E}_{ar}, \mathbf{E}_{ar})$ (i.e., pairs of entity mention spans), and $(\mathbf{E}_{tr}, \mathbf{E}_{ar})$ (i.e., pairs of entity mentions and event triggers) in cross-entropy losses for EME, ETD, RE and EAE respectively. Note that entity mentions/event arguments and event triggers are commonly called "instances" for the tasks in JIE.

For UDA, we follow the domain-adversarial training process in DANN (Ganin et al., 2016). The same encoder E is used to compute instance representations for JIE from input sentences in source and target domains. The source representations are then fed into a classification head F for main task learning. Concurrently, a domain discriminator D is employed taking as input representations of unlabeled samples from both domains to predict their corresponding origins. By pushing E to both minimize the main task losses and maximally misdirect D, the resulting representations will be both discriminative for the tasks at hand and indistinguishable to the domain classifier to boost performance in the target domain.

3.3 Instance-relational Domain Adaptation

Existing domain adaptation methods such as DANN tend to view all domains equally and ignore any topological structure among different domains to align them all perfectly. Recently, Xu et al. (2022) propose Graph-relational Domain Adaptation to generalize DANN to multi-domains adaptation setting by introducing a domain graph that captures domain heterogeneity. Each node of the graph represents a domain and a relation between two domains can be captured by an edge. By tailoring the adaptation of domains to a domain graph that reflects the true domain relationships, GrDA relaxes the uniform alignment to adapt more flexibly across domains. We adopt GrDA to solve the problem of UDA for multiple tasks in JIE by treating each of the task (i.e., EME and ETD) in two domains as a node in the type-relational graph $\mathbf{G}_r = (\mathbf{V}_r, \mathbf{A}_r)$ (i.e., `type` here refers to a combination of a task and a domain). Specifically, the vertex set $\mathbf{V}_r$ con-

sists of four nodes $\mathbf{E}^s_{ar}$, $\mathbf{E}^s_{tr}$, $\mathbf{E}^t_{ar}$, and $\mathbf{E}^t_{tr}$, and the adjacency matrix $\mathbf{A}_r \in \mathbb{R}^{4\times4}$ dictates which pair of types should be aligned by setting the value of corresponding position to 1. We assume a *chain* connection in the respective mentioned order for $\mathbf{G}_r$ (i.e., in Fig. 1d), on which detailed analysis will be provided later to justify the assumption. IrDA performs a minimax optimization similar to that of DANN with the following objective:

$$\min_{E,F} \max_{D} L^s_c(E,F) - \lambda L^{s-t}_d(D_g, E),$$

where λ is a balancing term and L^s_c is the combined loss for label prediction for JIE tasks in source domain, which depends on the encoder E and classification head F. Different from DANN where the discriminator predicts the domain identity given representations in the domains (Fig. 1a), the discriminator objective L^{s-t}_d aims to reconstruct the type-relational graph $\mathbf{G}_r$ given the encoding of data from different types (Fig. 1b). In particular, the graph discriminator D_g computes the pairwise relationship $\hat{a}_{ij}$ between two instance representations e_i and e_j (described in next section) for types i and j: $\hat{a}_{ij} = e_i^T e_j$, which is then used as input to the discriminator loss:

$$L^{s-t}_d = \sum_{i,j} l(\hat{a}_{ij}),$$
$$l(\hat{a}_{ij}) = -a_{ij} log\ \sigma(\hat{a}_{ij}) - (1 - a_{ij}) log\ (1 - \sigma(\hat{a}_{ij})),$$

where a_{ij} is the value of the edge between type i and j from the adjacency matrix $\mathbf{A}_r$. Intuitively, D_g aims to recover the relation graph $\mathbf{G}_r$ via the adjacency structure $\mathbf{A}_r$, while E seeks to prevent it from doing so. At equilibrium, the representations of two connected types will provide no information regarding their connection in $\mathbf{A}_r$. In other words, these types are aligned as we cannot infer their origins based on their representations.

3.4 Context-invariant Structure Learning

One problem with domain-adversarial training based methods in general is that they are sensitive to the amount of discrepancy between source and target domains. In particular, DANN's bound on target performance (David et al., 2010a) also depends on the loss of the ideal model to perform the main task on both domains. Accordingly, as the ideal model's loss is often assumed to be negligible for a single prediction task, it is ignored in the modeling process for DANN. However, in our setting with JIE, this simplification might be suboptimal as the combination of multiple tasks might increase the ideal model's loss, thus necessitating approaches to minimize this component for JIE in DA. In fact, if this component is not constrained, DANN will have little alignment effect for the representations while also worsening joint error term (David et al., 2010b; Wu et al., 2019). To this end, our proposal to minimize the ideal model's loss is to learn more transferable representations to facilitate its prediction in different domains. As such, we introduce a Context-invariant Structure Learning (CiSL) mechanism that aim to induce domain-general structures for input texts to better support transferable representation learning for JIE.

CiSL first creates domain-independent structure by combining linguistic and attention graphs extracted from the input sentence. For linguistic graph, we employ dependency trees that prior work found to be useful for IE tasks (Veyseh et al., 2020b). In particular, a graph $\mathbf{G}_d = (\mathbf{V}_d, \mathbf{A}_d)$ is constructed for each sentence based on the output of an off-the-shelf syntactic dependency parser, where $\mathbf{V}_d$ is a set of word-level nodes whose features are obtained by embedding the dependency relations between a word and its governor (embeddings are learnable parameters). The adjacency matrix $\mathbf{A}_d$ is a binary matrix whose cell (i, j) is only set to 1 if w_j is the governor of w_i in the dependency tree. We create augmented versions of $\mathbf{G}_d$ to reduce its sparsity and increase transferability, by merging it with attention graphs extracted from the output of each layer of the transformer encoder. Specifically, define an attention graph at layer l as $\mathbf{G}^l_a = (\mathbf{V}^l_a, \mathbf{A}^l_a)$, which composes of the transformer's position embeddings as node features for word-level nodes in $\mathbf{V}^l_a$ and attention probability matrix as adjacency matrix $\mathbf{A}^l_a$ ($1 \leq l \leq L$, where L is number of encoder's layers). The resulting attention-augmented dependency graphs $\mathbf{G}^l_{da} = (\mathbf{V}^l_{da}, \mathbf{A}^l_{da})$ are computed as follow:

$$\mathbf{A}^l_{da} = \alpha^l_a \mathbf{A}^l_a + \alpha^l_d \mathbf{A}_d\ ,$$
$$\mathbf{Z}^l_{da} = \beta^l_a \mathbf{Z}^l_a + \beta^l_d \mathbf{Z}_d\ ,$$

where $\{\alpha^l_a, \alpha^l_d, \beta^l_a, \beta^l_d\}^L_{l=1}$ are learnable weights, and the $\mathbf{Z}$s are the node representations of corresponding graphs. These graphs are context-independent in the sense that no word embedding information is explicitly included in their node features, whereas their adjacency matrices reflex relation among words that are universal across do-

mains in natural language. Finally, CiSL employs a Graph Transformer Network (Yun et al., 2019) to fuse the attention-augmented dependency graphs across all layers into a single context-invariant graph $\mathbf{G}_{ci} = (\mathbf{V}_{ci}, \mathbf{A}_{ci})$ with $\mathbf{A}_{ci} \in \mathbb{R}^{n \times n}$ and node features $\mathbf{Z}_{ci} \in \mathcal{R}^{n \times h}$.

To incorporate $\mathbf{G}_{ci}$ into the instance representation learning process, we add the node features $\mathbf{Z}_{ci}$ to the contextual representation $\mathbf{X}$, resulting in: $\mathbf{X}_{ci} = \mathbf{X} + \mathbf{Z}_{ci}$ as input to the CRF layers. This encourages downstream tasks to leverage more context-independent information from $\mathbf{Z}_{ci}$ instead of just relying on the domain-specific features $\mathbf{X}$.

Additionally, by viewing the input sentence as a graph with word-level nodes can be pooled obtain instance-level nodes, we introduce a CiG-conditioned pooling operation to create final instance representations for classification tasks. Each of the BIO tag sequence outputted by the CRF layers can be reformulated into a binary assignment matrix $\mathbf{S}_{base} \in \mathbb{R}^{n \times m}$, where $m \leq n$ is the number of spans detected, and $\mathbf{S}_{base_{ij}} = 1$ only if word i lies inside span j. Prior JIE systems simply compute each instance representation by summing its span's words: $e_j = \sum_{i;\, w_i \in span_j} \mathbf{x}_i$, thus solely relying on the text used to express the instance's meaning in the specific context. Accordingly, the previous equation (for each instance type) can also be formulated in matrix form as follow:

$$\mathbf{E} = \mathbf{S}^T \mathbf{X}_{ci} \ , \ where \ \mathbf{S} = \mathbf{S}_{base}.$$

To this end, instead of fixing the assignment matrix, we propose to learn $\mathbf{S}$ by conditioning its on the context-invariant graph $\mathbf{G}_{ci}$ as follow:

$$\mathbf{S}_{ci} = \gamma \odot \mathbf{S}_{base} + \mu \ ,$$

where $(\gamma, \mu) = GCN(\mathbf{Z}_{ci}, \mathbf{A}_{ci}) \in \mathbb{R}^{n \times 2m}$ are the outputs of a graph convolution network (Kipf and Welling, 2017; Nguyen and Grishman, 2018) taking in $\mathbf{G}_{ci}$ as input. Finally, the instance representation for label prediction is computed via:

$$\mathbf{E} = \mathbf{S}_{ci}^T \mathbf{X}_{ci} = (\gamma \odot \mathbf{S}_{base})^T \mathbf{X}_{ci} + \mu^T \mathbf{X}_{ci} \ ,$$

which is able to aggregate information over all words in the sentence through μ, and conversely suppress the role of the domain-related span's words through γ.

4 Experiments

4.1 Dataset, Settings, and Baselines

ACE-05 Following the prior works on JIE (Lin et al., 2020; Nguyen et al., 2021), we evaluate DA4JIE on the ACE-05 dataset (Walker et al., 2005) which provides annotations in 599 documents for entity mentions, event triggers, relations, and argument roles. In particular, there are 33 event classes, 7 entity classes, 6 relation classes, and 22 argument roles. ACE-05 was collected from 6 different domains: `bn`, `nw`, `bc`, `cts`, `wl`, and `un`. For UDA setting, we follow Ngo et al. (2021) and gather data from two closely related domains, `bn` and `nw`, to create a sizable source domain dataset and refer to is it as `in` domain. We use 80% of its documents for training whilst the rest are used for development. For out-of-domain (OOD) setting, each of the other domains is considered a target domain of a single adaptation scenario, where 20% of its documents are reserved for unlabeled training target data and the remainders are utilized as the test dataset. We use the same data processing scripts as in (Lin et al., 2020; Nguyen et al., 2021) for consistency.

			Out-of-domain				
		`in`	`bc`	`cts`	`wl`	`un`	**aDom**
BERT	Trigger-I	78.4	71.4	65.2	62.9	66.3	66.4
	Role-I	64.1	59.5	49.0	46.3	46.9	50.4
	Entity	88.9	80.8	84.0	85.5	80.9	82.8
	Relation-C	64.3	61.7	58.0	52.5	48.0	55.0
	Trigger-C	76.3	68.7	62.4	56.3	64.5	63.0
	Role-C	60.8	55.4	47.9	42.9	43.0	47.3
	aTask	72.6	66.6	63.1	59.3	59.1	62.0
OneIE	Trigger-I	79.1	70.3	68.2	63.2	64.6	66.6
	Role-I	66.2	60.1	51.2	50.6	46.7	52.1
	Entity	89.1	79.5	86.9	85.5	81.5	83.4
	Relation-C	65.6	63.1	56.7	54.7	50.0	56.1
	Trigger-C	77.2	67.5	64.6	56.8	63.4	63.1
	Role-C	62.2	55.7	49.9	47.2	42.6	48.9
	aTask	73.5	66.5	64.6	61.1	59.4	62.9
FourIE	Trigger-I	79.1	70.7	66.0	65.2	64.3	66.6
	Role-I	66.6	60.0	52.6	48.9	49.1	52.6
	Entity	89.1	80.3	84.4	85.4	81.9	83.0
	Relation-C	66.0	63.7	56.6	53.1	52.7	56.5
	Trigger-C	76.9	68.5	63.2	56.4	62.4	62.6
	Role-C	61.8	55.4	51.8	44.5	43.6	48.8
	aTask	73.5	66.9	64.0	59.8	60.1	62.7
DA4JIE	Trigger-I	79.0	72.2	66.0	64.4	66.5	**67.3**
	Role-I	67.3	59.0	54.6	49.8	51.5	**53.7**
	Entity	89.2	82.6	86.0	85.2	83.0	**84.2**
	Relation-C	68.8	65.3	58.7	57.7	54.3	**59.0**
	Trigger-C	76.5	68.7	63.0	57.4	64.1	**63.3**
	Role-C	62.5	55.6	51.9	45.3	44.4	**49.3**
	aTask	**74.2**	**68.0**	**65.0**	**61.4**	**61.4**	**64.0**

Table 1: F1 scores of the models on ACE-05 test data for in-domain (`in`) and out-of-domain (`bc`, `cts`, `wl`, `un`) adaptation settings. The suffixes "-I" and "-C" correspond to the identification performance (only concerning the offset correctness) and identification+classification performance (evaluating both offsets and classes). **aTask** is the average score over the four classification tasks, and **aDom** is the average out-of-domain score for each task.

Baselines We compare DA4JIE with the following current SOTA JIE systems: (i) **BERT** (Devlin et al., 2019) uses a shared Transformer encoder to represent the instances for ETD, EME, EAE, and RE and performs classification for the instances

based on the task-specific label distributions. (ii) **OneIE** (Lin et al., 2020) is same as BERT, but leverages set of predefined global features to capture the cross-subtask and cross-instance interactions. (iii) **FourIE** (Nguyen et al., 2021) creates a graph structure of contextual representations to explicitly capture the interactions between related instances of the four IE tasks in a sentence, while also employing a heuristic dependency between the task instances in a dependency-based regularization to further boost the performance of the models. **OneIE** and **FourIE** are current state-of-the-art models for JIE.

Implementation Details and Hyper-parameters All models are implemented in Pytorch. We leverage the pre-trained BERT-large-cased models and checkpoints from Huggingface repository (Wolf et al., 2020). To achieve a fair comparison with the baselines, we follow the same evaluation script and correctness criteria for entity mentions, event triggers, relations, and arguments as in prior work (Lin et al., 2020). To tune each model over in-domain development data, we use Adam optimizer with learning rates chosen from [5e-5, 1e-4, 5e-4, 1e-3, 5e-3], mini-batch size from [16, 32, 64] of which 50% are unlabeled target data. We use GCNs with 2 or 3 layers and GTN with number of channels in [2, 4, 8]. All of the downstream heads are implemented as 2 or 3 layers feed-forward networks with hidden vectors of size [100, 50] or [200, 100, 50], respectively. The IrDA balancing term λ is picked from the range [0.1, 0.5, 1, 5, 10]. Every model is trained for 50 epochs for each target domain, from which the model with the best average task F1 score on the in-domain development set is then evaluated OOD on the test set of the corresponding target domain. Finally, our reported results are average of three runs using the best hyper-parameter configuration with different random seeds. The selected hyper-parameters for our model from the fine-tuning process include: 3 layers for the GCNs and feed-forward classification heads, GTN with 4 channels, 1e-5 for the learning rate with Adam optimizer, 16 for the batch size, and 1 for the balancing term λ. In this work, we use a single Tesla V100-SXM2 GPU with 32GB memory for all experiments.

4.2 Main Results

Table 1 showcases the UDA results in F1 scores for all tasks in JIE. We observe that the latest systems for JIE such as **OneIE** and **FourIE** provide marginal improvement to the standard **BERT** model. In particular, while the specialized architectures of these models are able to boost in-domain performance as expected, they are not tailored to UDA settings where the focus is on extracting transferable and domain-invariant features. As a result, their effectiveness in out-of-domain (OOD) settings over **BERT** is situational (**OneIE** is good for `cts` and `wl` domains, while **FourIE** is better at adapting to `un` domains). In contrast, **DA4JIE** manages to achieve the best adaptation performance in average across all considered domains. Our model is 2 points higher in F1 scores than **BERT**'s overall, surpassing the current SOTA methods by over 1 point on average. Notably, this improvement is the result of the simultaneous increases in the average performance of all downstream tasks, which is achieved by combining IrDA and CiSL modules as shown in the following section.

		bc	cts	wl	un	aDom
DA4JIE	Trigger-I	72.2	66.0	64.4	66.5	**67.3**
	Role-I	59.0	54.6	49.8	51.5	**53.7**
	Entity	82.6	86.0	85.2	83.0	**84.2**
	Relation-C	65.3	58.7	57.7	54.3	**59.0**
	Trigger-C	68.7	63.0	57.4	64.1	**63.3**
	Role-C	55.6	51.9	45.3	44.4	**49.3**
	aTask	**68.0**	**65.0**	**61.4**	**61.4**	**64.0**
DA4JIE -CiSL	Trigger-I	71.6	65.9	64.1	64.7	66.6
	Role-I	60.1	52.9	49.5	47.9	52.6
	Entity	82.1	74.4	82.9	82.0	80.3
	Relation-C	63.3	54.8	54.9	53.1	56.5
	Trigger-C	69.3	63.8	56.3	62.9	63.1
	Role-C	56.0	51.7	44.1	43.4	48.8
	aTask	67.7	61.2	59.5	60.3	62.2
DA4JIE -IrDA	Trigger-I	66.4	64.8	64.9	66.9	66.9
	Role-I	52.5	49.8	47.1	52.4	52.4
	Entity	87.2	84.6	81.7	83.9	83.9
	Relation-C	55.3	52.9	53.3	56.1	56.1
	Trigger-C	63.4	57.7	63.5	63.3	63.3
	Role-C	51.3	46.0	42.0	48.7	48.7
	aTask	67.2	64.3	60.3	60.1	63.0
DA4JIE -IrDA -CiSL	Trigger-I	71.4	65.2	62.9	66.3	66.4
	Role-I	59.5	49.0	46.3	46.9	50.4
	Entity	80.8	84.0	85.5	80.9	82.8
	Relation-C	61.7	58.0	52.5	48.0	55.0
	Trigger-C	68.7	62.4	56.3	64.5	63.0
	Role-C	55.4	47.9	42.9	43.0	47.3
	aTask	66.6	63.1	59.3	59.1	62.0

Table 2: Performance (F1 scores) for ablation study on the ACE-05 test datasets for different domains.

4.3 Ablation study

We conduct an ablation study to validate the effectiveness of each of our main components by investigating the following variations of our model by removing CiSL, IrDA, and both respectively. The results is shown in Table 2, where we observe that **DA4JIE-IrDA** noticeably boosts per-

formances for all domains compared to BERT (**DA4JIE-IrDA-CiSL**), while **DA4JIE-CiSL** only has positive impact when adapting to bc and un domains, providing little to no improvement on average. This is the result of CiSL making the instance representations more transferable at low-level, thus ensuring the necessary condition for the domain-adversarial training in IrDA to reach equilibrium. By combining both components, **DA4JIE** significantly outperforms other variants, especially when transferring to target domains that are highly dissimilar to source domains (i.e., wl and un).

	bc	cts	wl	un	**aDom**
None	66.6	63.1	59.3	59.1	62.0
Full	67.1	63.7	60.0	60.1	62.7
Pair-Task	67.5	63.2	60.0	61.1	62.9
Pair-Dom	67.0	62.7	59.1	59.5	62.1
Chain	**68.0**	**65.0**	**61.4**	**61.4**	**64.0**

Table 3: Average task scores for domain-adversarial learning analysis. Performance (F1 scores) on the ACE-05 test datasets for different domains.

	aId	**aCls**
CiSL	**60.5**	**64.0**
CiSL-Pool	59.7	63.2
CiSL-Node	59.4	62.3
CiSL-Node-Pool	58.0	62.0
CiSL-Dep	59.8	62.8
CiSL-Attn	59.2	62.5

Table 4: Average identification and classification scores for CiSL analysis. Performance (F1 scores) on the ACE-05 test datasets for different domains. **aId** and **aCls** are the average scores across all new domains, of all identification and classification tasks, respectively.

5 Analysis

5.1 Instance-relational Graph Analysis

We investigate the effect of IrDA with different patterns of relationships in the type-relational graph compared to our *chain* relation in DA4JIE. In Table 3, **Full** refers to the standard DANN approach where all types (i.e., task+domain) are uniformly aligned (Fig. 1c). **Pair-Task** and **Pair-Dom** are models with only a pair of edges in relation graph, the former connects the same task across domains ($\mathbf{E}^s_{ar} - \mathbf{E}^t_{ar}$ and $\mathbf{E}^s_{tr} - \mathbf{E}^t_{tr}$), while the latter has tasks in the same domain linked ($\mathbf{E}^s_{tr} - \mathbf{E}^s_{ar}$ and $\mathbf{E}^t_{tr} - \mathbf{E}^t_{ar}$). Finally, **None** means no adaptation is used and **Chain** corresponds to our assumption in DA4JIE. The results show that **Full** improves over **None**, but underperforms when compared to **Pair-Task** in most new domains. This indicates that the alignment imposed by **Full** is overly strict and not optimal when adapting multiple tasks together. In addition, appropriate connections are required for effective adaptation, as shown by the low scores of **Pair-Dom**, which basically is equivalent to domain-conditioning the representations without adapting between source and target domains.

We argue that **Chain** is robust and substantially outperforms other models across all domains because it reflects the true relationship among the tasks and domains (types) for JIE. In particular, event triggers are restricted and closely related to the predefined event classes which are shared across domains, therefore their representations should be aligned when adapting to new domains. Conversely, event arguments (i.e., entity mentions) are more diverse and context-dependent, thus may significantly differ across domains and should not be directly connected in the relation graph. They are, however, implicitly connected in **Chain** through the event trigger nodes, which implies their representations are "weakly" aligned, as shown by Xu et al. (2022) where GrDA being able to enforce different levels of alignment. Lastly, the pair of trigger-event edges in source and target domains also equate to aligning the representations of trigger-event relation and help transfer model's role classification ability from source to target domain.

5.2 Context-invariant Structure Learning

To determine the role of different components in CiSL module, we analyze their contributions to DA4JIE performance at different levels of downstream tasks. In Table 4, **CiSL-Dep** and **CiSL-Attn** are the models without leveraging dependency graph and attention graph respectively. **CiSL-Pool** just uses the base assignment matrix for pooling, and **CiSL-Node** is the case where node features of the context-invariant graph are removed from the inputs for the CRF layers. Finally, we completely disable the CiSL module in **CiSL-Node-Pool**. From the results, it is clear that both node features and conditional pooling are responsible for the significant improvement of the final model. Particularly, adding the node features is more effective as it also helps boost the performance of identification tasks by making the representations more transferable from source to target domains at low-level. Furthermore, the last two rows in the table indicate that combining different kinds of structures has a positive impact, especially

when they contain universal linguistic information that is general across domains.

6 Conclusion

We present DA4JIE, a novel framework that jointly solves four IE tasks (EME, ETD, RE, and EAE) in UDA setting. In particular, DA4JIE employs Instance-relational Domain Adaptation method that generalizes the standard domain-adversarial training approach to simultaneously align high-level type representations of all downstream tasks between domains. Additionally, we incorporate a Context-invariant Graph learning module into the encoder to encourage the usage of domain-independent information at low-level, thus extracting more transferable features to improve model's performance in new domains. The extensive experiments demonstrate the effectiveness of the proposed framework. In the future, we plan to extend our approach to more general settings such as multi-source domain adaptation with more IE subtasks such as entity/event coreference resolution.

Limitations

We present the first work to tackle the joint information extraction problem in unsupervised domain adaptation setting. Our framework DA4JIE combines Instance-relations Domain Adaptaion method with Context-invariant Structure Learning mechanism, outperforming state-of-the-art systems on ACE-05 consistently across multiple new domains. Despite positive empirical results, there are still several limitations that can be addressed in future works. First, the current model assumes a *chain* connection for the type-relational graph in IrDA. While intuitive and effective for the considered setting in this work, it is only designed manually. A method that explicitly learns to find the optimal connections for the relation graph might be able to produce better performance for our problem. Another issue is the limited kinds of linguistic structures that CiSL uses to create the context-invariant graph. Prior works have successfully improved IE tasks using semantic role labeling (Christensen et al., 2010) and abstract meaning representation (Zhang and Ji, 2021). Integrating structured graphs extracted from these methods is straightforward for DA4JIE and might improve model performance further.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IU-CRC Center for Big Learning. This research is also based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52–60, Los Angeles, California. Association for Computational Linguistics.

Yong Dai, Jian Liu, Xiancong Ren, and Zenglin Xu. 2020. Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020,*

New York, NY, USA, February 7-12, 2020, pages 7618–7625. AAAI Press.

Shai Ben David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010a. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Shai Ben David, Tyler Lu, Teresa Luu, and David Pal. 2010b. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136, Chia Laguna Resort, Sardinia, Italy. PMLR.

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)*.

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

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Alex Judea and Michael Strube. 2016. Incremental global event extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2279–2289, Osaka, Japan. The COLING 2016 Organizing Committee.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Meelis Kull and Peter A. Flach. 2014. Patterns of dataset shift.

Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. 2018. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks.

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.

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.

Aakanksha Naik and Carolyn Rosé. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nghia Trung Ngo, Duy Phung, and Thien Huu Nguyen. 2021. Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, page 4015–4025. Association for Computational Linguistics.

Minh Van Nguyen, Viet Lai, and Thien Huu Nguyen. 2021. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *NAACL-HLT*, pages 27–38.

Minh Van Nguyen, Bonan Min, Franck Dernoncourt, and Thien Nguyen. 2022. Joint extraction of entities, relations, and events via modeling inter-instance and inter-label dependencies. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4363–4374, Seattle, United States. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Nghia Ngo Trung, Linh Ngo Van, and Thien Huu Nguyen. 2022. Unsupervised domain adaptation for text classification via meta self-paced learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4741–4752, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Amir Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020a. Exploiting the syntax-model consistency for neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8021–8032, Online. Association for Computational Linguistics.

Amir Veyseh, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020b. Graph transformer networks with syntactic and semantic structures for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3651–3661, Online. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963–7974, Online. Association for Computational Linguistics.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. 2019. Domain adaptation with asymmetrically-relaxed distribution alignment. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6872–6881. PMLR.

Zihao Xu, Hao He, Guang-He Lee, Bernie Wang, and Hao Wang. 2022. Graph-relational domain adaptation. In *International Conference on Learning Representations*.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.

Xiaofeng Yu and Wai Lam. 2010. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Coling 2010: Posters*, pages 1399–1407, Beijing, China. Coling 2010 Organizing Committee.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning.

Junchi Zhang, Yanxia Qin, Yue Zhang, Mengchi Liu, and Donghong Ji. 2019. Extracting entities and events as a single task using a transition-based neural model. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5422–5428. International Joint Conferences on Artificial Intelligence Organization.

Zixuan Zhang and Heng Ji. 2021. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online. Association for Computational Linguistics.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.