A Multi-Format Transfer Learning Model for Event Argument Extraction via Variational Information Bottleneck

Jie Zhou[1†*], Qi Zhang[2†], Qin Chen[2], Qi Zhang[1], Liang He[2], and Xuanjing Huang[1]

[1]School of Computer Science, Fudan Univerisity, Shanghai, China
[2]School of Computer Science and Technology, East China Normal University, Shanghai, China
{jie_zhou, qz, xjhuang}@fudan.edu.cn
qzhang@stu.ecnu.edu.cn; {qchen, lhe}@cs.ecnu.edu.cn

Abstract

Event argument extraction (EAE) aims to extract arguments with given roles from texts, which have been widely studied in natural language processing. Most previous works have achieved good performance in specific EAE datasets with dedicated neural architectures. Whereas, these architectures are usually difficult to adapt to new datasets/scenarios with various annotation schemas or formats. Furthermore, they rely on large-scale labeled data for training, which is unavailable due to the high labelling cost in most cases. In this paper, we propose a multi-format transfer learning model with variational information bottleneck, which makes use of the information especially the common knowledge in existing datasets for EAE in new datasets. Specifically, we introduce a shared-specific prompt framework to learn both format-shared and format-specific knowledge from datasets with different formats. In order to further absorb the common knowledge for EAE and eliminate the irrelevant noise, we integrate variational information bottleneck into our architecture to refine the shared representation. We conduct extensive experiments on three benchmark datasets, and obtain new state-of-the-art performance on EAE.

1 Introduction

Event Extraction (EE) has received widespread attention in recent years, which aims to obtain structured information (e.g., trigger, event types, arguments, and argument roles) from large unstructured text corpora (Lu et al., 2021; Zhang et al., 2022). Event argument extraction (EAE) plays a crucial role in EE. Recently, deep learning-based models have obtained tremendous success in this task. However, these methods rely on a large-scale labeled dataset for training, which is time-consuming and labor-intensive due to the complexity of event extraction.

ACE2005 Example	Conflict.Attack	
And to the south , **British**[Place] **forces**[Target] continue their **attack**[trigger] on targets around Basra .	trigger	attack
	Target	forces
	Place	Basra
	Attacker	-
	Victim	-
	Instrument	-

WIKIEVENTS Example	Conflict.Attack.DetonateExplode	
For example , Ms . Davis has identified a man whose photo matches that of a \" John Doe # 2 \" sought immediately after the Murrah **Building**[Target] **attack**[trigger]. He appears to be a Palestinian by the name of Hussain Hashem Al Hussaini ...	trigger	attack
	Target	Murrah Building
	Place	-
	Attacker	-
	ExplosiveDevice	-
	Instrument	-

Figure 1: An example with a different format.

In this paper, we aim to answer the question "Can we transfer the knowledge from the existing complex event extraction datasets with different formats?". There are several event extraction datasets, such as ACE 2005 (Doddington et al., 2004), RAMS (Ebner et al., 2020), and WikiEvents (Li et al., 2021). These datasets contain abundant event types and semantic roles that may possess overlap knowledge and help to improve the performance of new datasets or low-resource extraction. As shown in Figure 1, both ACE2005 and WikiEvents datasets contain the same "attack" event type with inconsistent names. Additionally, some shared argument roles (e.g., "Target", "Attacker", "Place", and "Instrument") are labeled in both two datasets. All this information shows that the event knowledge can be transferred between two datasets.

However, the transfer between different event argument extraction is a challenging task. **(C1)** One challenge is that the formats of various datasets are inconsistent due to the complex structure of event records. Thus, it is hard to find a unified model to extract arguments with different formats. More specifically, 1) Two datasets may have dif-

*Corresponding authors; †Equal contribution.

ferent event types, which have various argument structures; 2) The same event type or argument type in two datasets may have different names. For example, the event names are “Conflict Attack" and “Conflict Attack Detonate Explode" in ACE2005 and WikiEvents respectively (Figure 1); 3) The argument roles set of the same event type may be different in various datasets. For instance, the argument role “Victim" and “ExplosiveDevice" for event “Attack" only appear in ACE2005 and WikiEvents, respectively (Figure 1). **(C2)** The another challenge is that the annotation among different datasets may exist a gap, which brings noise for transfer learning. Two datasets may have significant semantic differences, as they may belong to different domains. In addition, the annotation guidelines may be contradictory among various datasets. Our experiments also show that merging two datasets simply may reduce the performance.

Previous works mainly regard the argument extraction as a sequence labeling, which can not transfer to new event argument types (Yang et al., 2018). Then, a machine reading comprehension problem (MRC) based model is proposed to extract the arguments using natural questions (Liu et al., 2020; Du and Cardie, 2020). Recently, prompt-learning (Schick and Schütze, 2020; Liu et al., 2021b) based models (Ma et al., 2022; Chen et al., 2020) and generation-based models (Chen et al., 2020; Du et al., 2021; Li et al., 2021) are utilized for event argument extraction. These studies inspire us to design a unified model that can extract arguments with different formats for EAE. Moreover, some researches investigate cross-lingual event extraction (Subburathinam et al., 2019) and zero-shot event extraction (Chen et al., 2020; Feng et al., 2020), which are under zero-shot setting. In other words, these studies train on the source language or domain and transfer it to the target domain, where the target domain has no training data. Different from them, we train our model on both the source and target datasets with different formats where the format-shared knowledge is essential.

To deal with the above challenges, we propose a multi-format transfer learning model for EAE via information bottleneck, denoted as `UnifiedEAE`, which can leverage all event extraction datasets with heterogeneous formats. First, we adopt a Shared-Specific Prompt (`SSP`) framework to capture format-shared and format-specific knowledge to extract arguments with different formats. Then, to better capture the format-shared representation, we incorporate the variational information bottleneck (VIB) into the format-shared model (`SharedVIB`). VIB has been widely used to forget the irrelevant information and retain the vital information for prediction (Li and Eisner, 2019; Tishby et al., 2000). We leverage it to enhance the model to learn the format-shared knowledge. We conduct a series of experiments on three publicly available datasets and obtain new state-of-the-art performance. Our `UnifiedEAE` can also improve the performance of low-resource EAE effectively. Furthermore, the results show that our model can capture the format-shared knowledge and forget the noise among various datasets.

In summary, the main contributions of this paper are summarized as follows.

- We design a unified architecture that can learn both the format-shared and format-specific knowledge from various EE datasets with heterogeneous formats.
- The information bottleneck technology is utilized to enhance the model to learn the format-shared knowledge among different datasets by eliminating the irrelevant information and reserving the format-shared knowledge.
- Extensive experiments on three datasets show the great advantages of our model. Also, our model performs well on low-resource event argument extraction.

2 Related Work

2.1 Event Argument Extraction

Event extraction can be split into two subtasks, event identification and event argument extraction (EAE) (Zhang et al., 2020; Chen et al., 2015; Lin et al., 2022). We focus on the EAE task, which aims to extract the arguments based on the given event type and trigger (Wei et al., 2021; Ma et al., 2022). Wei et al. (2021) added constraints with each argument role to take the interaction into account. Data augmentation is adopted for event argument extraction (Liu et al., 2021a). To avoid the error propagation and learn the relationships among the subtasks, end to end model performs two subtasks jointly (Zhang et al., 2019; Wadden et al., 2019; Li et al., 2021). Several studies regard event argument extraction as a machine reading comprehension problem (MRC), which extracts the arguments

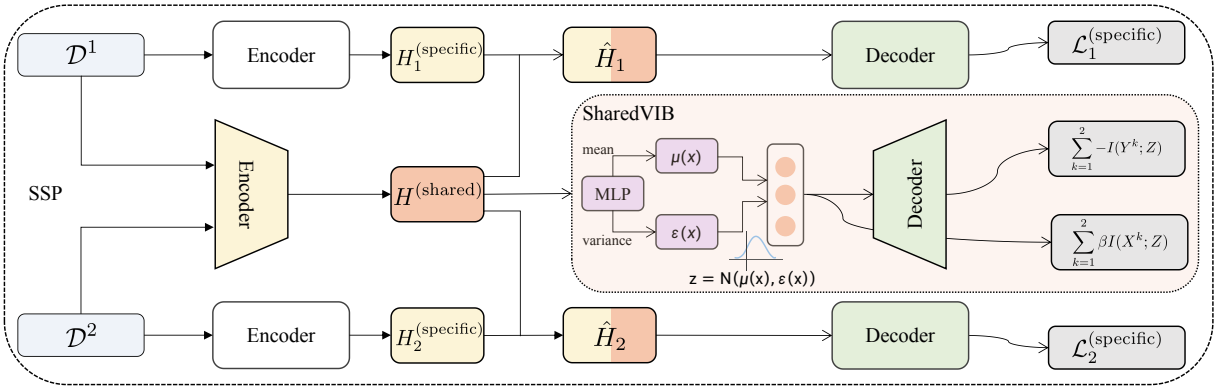


Figure 2: The framework of our `UnifiedEAE` model. To learn both the format-shared and format-specific representations (i.e., $\hat{H}_1$ and $\hat{H}_2$), we introduce a shared-specific prompt (`SSP`) model (white background). Then, we design a `SharedVIB` module (pink background) to better capture the format-shared representation (e.g., $H^{(\text{shared})}$) by forgetting format-specific information ($\sum_{k=1}^{2} \beta I(X^k; Z)$) and retaining format-shared knowledge ($\sum_{k=1}^{2} -I(Y^k; Z)$) via information bottleneck.

based on natural questions (Liu et al., 2020; Du and Cardie, 2020). Recently, prompt-learning (Schick and Schütze, 2020; Liu et al., 2021b) based models (Ma et al., 2022; Chen et al., 2020) and generation-based models (Chen et al., 2020; Du et al., 2021; Li et al., 2021) are utilized for event argument extraction. In this paper, we aim to transfer the knowledge of the existing event extraction datasets to the target dataset, which is not well studied since the complexity of this task.

2.2 Transfer Learning for NLP

To reduce the requirements of labeled data, transfer learning has attached great attention in the field of natural language processing (Liu et al., 2017; Ruder et al., 2019; Raffel et al., 2020; Zhou et al., 2020). Liu et al. (2017) proposed an adversarial multi-task learning framework to learn the shared and private representation. Cross-lingual event extraction aims to transfer the knowledge from the source language to the target language (Subburathinam et al., 2019). Zero-shot transfer learning is also explored on semantic role labeling (SRL) (Peng et al., 2016), event extraction (Chen et al., 2020; Feng et al., 2020), and abstract meaning representation (AMR) (Huang et al., 2018). Different from them, we focus on transfer learning among event argument extraction datasets with various complex formats where both the format-shared knowledge and format-specific knowledge are important.

2.3 Information Bottleneck

Recently, information bottleneck (IB) has been applied in NLP tasks, such as word cluster (Pereira et al., 1994), dependent parsing (Mahabadi et al., 2021), summarization (West et al., 2019), interpretability (Zhou et al., 2021). Li and Eisner (2019) use IB for compressing the hidden representations of words by removing the task-irrelevant information. Sun et al. (2021) adopted the IB principle for graph structure learning. Variational IB (VIB) is used as a regularization technique to improve the fine-tuning of pre-training language models in low-resource scenarios (Mahabadi et al., 2021). In this paper, we attempt to use VIB to constraint model to learn format-shared information for event argument extraction.

3 Methodology

To transfer the knowledge among the datasets with different formats, we propose a `UnifiedEAE` model for event argument extraction task (Figure 2). `UnifiedEAE` is based on a shared-specific prompt (`SSP`) architecture, which learns both the format-shared and format-specific knowledge from diverse datasets with multiple formats. Then, to enhance the model to learn the format-shared knowledge, we integrate variational information bottleneck into the format-shared model (`SharedVIB`) by removing the format irrelevant information and retaining format invariable knowledge.

Formally, given two event argument extraction datasets, denoted by $\mathcal{D}^1 = \{(X_i^1, Y_i^1)\}_{i=1}^{|\mathcal{D}^1|}$ and $\mathcal{D}^2 = \{(X_i^2, Y_i^2)\}_{i=1}^{|\mathcal{D}^2|}$ where $|\mathcal{D}^1|$ and $|\mathcal{D}^2|$ are the number of samples in dataset $\mathcal{D}^1$ and $\mathcal{D}^2$. For each sample $(X, Y) \in \mathcal{D}^1 or \mathcal{D}^2$, the input $X = \{s, e, t, R\}$ contains the sentence s, event

type e, and trigger word t, R denotes the set of event-specific role types, we aims to extract a set of span Y. For the i-th span in Y, $\text{span}_i^{\text{start}}$ and $\text{span}_i^{\text{end}}$ are the start and end indices of the span.

3.1 Shared-Specific Prompt

The shared-specific prompt (SSP) architecture aims to learn both format-shared and format-specific knowledge for EAE. This framework consists of three event argument extractors: two format-specific and one format-shared extractor, which are used to learn the format-specific and format-shared knowledge. We adopt a prompt-based model as the basic extractor to predict multi-format arguments.

Basic Prompt-based Extractor. Following the Ma et al. (2022), we use a BART (Lewis et al., 2020) based prompt model as an event argument extractor. This model consists of an encoder and decoder. The encoder is used to learn the event-aware sentence representation. Then we adopt a decoder model to extract all the argument spans jointly via a prompt template.

Encoder. To consider the position information of event, we insert special token "<t>" and "</t>" before and after the trigger t in the sentence s. Then we input it into BART to obtain the event-aware sentence representation H,

$$\begin{aligned} H_{\text{encoder}} &= \text{BART}_{\text{Encoder}}(s), \\ H &= \text{BART}_{\text{Decoder}}(s, H_{\text{encoder}}), \end{aligned} \tag{1}$$

Decoder. In the decoder, we use a prompt with slots to extract the argument roles at the same time. We use manual template from Li et al. (2021). For example, the prompt is "Person married Person at Place (and Place)" for event type "Life.Marry" in ACE2005. We aim to predict the argument spans for the four argument role slots. We input the prompt p to the BART decoder to obtain the prompt representation.

$$H_p = \text{BART}_{\text{decoder}}(p, H_{\text{encoder}})$$

For the i-th role slot in the prompt, we use the mean pooling of the corresponding tokens' representation from H_p as the role representation r_i. Then, we extract the argument span for role r_i by predicting the start and end index in the text.

$$\begin{aligned} p_i^{(\text{start})} &= \text{Softmax}(r_i w^{(\text{start})} H) \\ p_i^{(\text{end})} &= \text{Softmax}(r_i w^{(\text{end})} H) \end{aligned} \tag{2}$$

where $w^{(\text{start})}$ and $w^{(\text{end})}$ are the learnable parameters.

Finally, the cross-entropy between the predicted start/end probability $p_i^{(\text{start})}/p_i^{(\text{end})}$ and the ground truth,

$$\begin{aligned} \mathcal{L} = \sum_{X \in D} \sum_{i=1}^{|R|} &\text{CrossEntropy}(p_i^{(\text{start})}, \text{span}_i^{(\text{start})}) \\ &+\text{CrossEntropy}(p_i^{(\text{end})}, \text{span}_i^{(\text{end})}) \end{aligned} \tag{3}$$

where $\text{span}_i^{\text{start}}/\text{span}_i^{\text{end}}$ are the start and end index of the i-th argument role's span, $|R|$ is the length of roles set R in X.

For dataset $\mathcal{D}^1$ and $\mathcal{D}^2$, we use two independent prompt-based extractors to learn format-specific sentence representations $H_1^{(\text{specific})}$ and $H_2^{(\text{specific})}$ calculated by Equation 1. Then, the third extractor is adopted to learn the format-shared sentence representation $H^{(\text{shared})}$ among two datasets. To predict the event argument based on both format-specific and format-shared knowledge, we combine specific representation $H_k^{(\text{specific})}, k \in \{1, 2\}$ and shared representation $H^{(\text{shared})}$ via a gate mechanism (Hochreiter and Schmidhuber, 1997).

$$\begin{aligned} g_k &= \sigma\left(W_{g_k} \cdot \left[H_k^{(\text{specific})}, H^{(\text{shared})}\right] + b_{g_k}\right) \\ \hat{H}_k &= g_k * H_k^{(\text{specific})} + (1 - g_k) * H^{(\text{shared})} \end{aligned}$$

where W_{g_k}, b_{g_k} are the trainable parameters, σ is a sigmoid active function.

Then, we extract argument span by replacing H in Equation 2 with $\hat{H}_k$. In this way, we can predict the arguments based on both the format-specific and format-shared knowledge. Then, we obtain the format-specific loss $\mathcal{L}_{\text{SSP}} = \mathcal{L}_1^{(\text{specific})} + \mathcal{L}_2^{(\text{specific})}$, where $\mathcal{L}_1^{(\text{specific})}$ and $\mathcal{L}_2^{(\text{specific})}$ are the loss for dataset $\mathcal{D}^1$ and $\mathcal{D}^2$.

3.2 Shared Knowledge Learning via VIB

We hope the shared model in shared-specific prompt architecture to learn the format-shared knowledge while forgetting the format-specific knowledge. However, we do not add objectives to enhance the model to do this. Inspired by (Li and Eisner, 2019), we integrate variational information bottleneck (VIB) into our shared model (SharedVIB) to capture the format-shared knowledge while eliminating the format-specific information.

Particularly, the information bottleneck aims to learn a compressed representation Z, which maximizes mutual information with output Y and minimizes mutual information with input X. In this paper, we tend to let Z retain the information about $Y^k, k \in \{1, 2\}$ and remove the irrelevant information in $X^k, k \in \{1, 2\}$.

$$\sum_{k=1}^{2} \beta I(X^k; Z) - I(Y^k; Z)$$

The shared model performs both dataset $\mathcal{D}^1$ and $\mathcal{D}^2$ at the same time to learn the format-shared knowledge. Then, we let the model to forget the format-specific information in $X^k, k \in \{1, 2\}$ by minimizes mutual information between Z and $X^k, k \in \{1, 2\}$.

It is challenging to compute the mutual information $I(Y^k; Z)$ and $I(X^k; Z)$ directly. The same as (Li and Eisner, 2019), we use variational inference to compute a variational upper bound for $I(X^k; Z)$ as follow,

$$\overbrace{\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{z\sim p(z|x)}\left[\log \frac{p(z \mid x)}{q(z)}\right]\right]}^{\text{upper bound}} - \overbrace{\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{t\sim p(z|x)}\left[\log \frac{p(z \mid x)}{p(z)}\right]\right]}^{\mathrm{I}(X^k;Z)}$$
$$= \mathop{\mathbb{E}}_{x}\left[\mathrm{KL}\left(p(z) \| q(z)\right)\right] \geq 0$$

We optimize the upper bounder of $I(X^k; Z)$ to minimize it. We use reparameterzation method (Kingma and Welling, 2014) to sample Z from the latent distribution according to $p(z|x)$,

$$p(z \mid x) = \mathcal{N}\left(z \mid f^{\mu}(x), f^{\Sigma}(x)\right)$$

where $f^{\mu}(x) = H^{(shared)} \cdot W^{\mu}$ and $f^{\Sigma}(x) = H^{(shared)} \cdot W^{\Sigma}$ are the mean and variance of the latent Gaussian distribution, W^{μ} and W^{Σ} are the learnable parameters. Thus, we estimate $I(X^k; Z) = \mathrm{KL}\left(p(z \mid x) \| q(z)\right)$. For $q(z)$, we let it be a standard diagonal normal distribution.

For $I(Y^k; Z)$, we calculate the variational lower bound,

$$\overbrace{\mathop{\mathbb{E}}_{y,z\sim p}\left[\log \frac{p(y \mid z)}{p(y)}\right]}^{\mathrm{I}(Y^k;Z)} - \overbrace{\mathop{\mathbb{E}}_{y,z\sim p}\left[\log \frac{\psi(y \mid z)}{p(y)}\right]}^{\text{lower bound}}$$
$$= \mathop{\mathbb{E}}_{z\sim p}\left[\mathrm{KL}\left(p(y \mid z) \| \psi(y \mid z)\right)\right] \geq 0$$

Here, we use the decoder model in our shared-specific prompt (Section 3.1) as $\psi(y \mid z)$ by replacing H in Equation 2 with the sampled Z. Thus, the loss for optimizing $\psi(y \mid z)$ on $D^k, k \in \{1, 2\}$ is the same as Equation 3 by replacing the sampled Z with H Equation 2, denoted as $\mathcal{L}^{(shared)}$.

Thus, the loss function for `SharedVIB` is,

$$\mathcal{L}_{\text{SharedVIB}} = \sum_{k=1}^{2}\left(\mathcal{L}_k^{(\text{shared})} + \beta \sum_{X\in D^k} \mathrm{KL}\left(p(z \mid x) \| q(z)\right)\right)$$

Finally, the total loss for our `UnifiedEAE` is,

$$\mathcal{L} = \mathcal{L}_{\text{SSP}} + \mathcal{L}_{\text{SharedVIB}}$$

Table 1: Data statistics of RAMS and WikiEvents

Dataset	Split	#Doc	#Event	#Argument
RAMS	Train	3194	7329	17026
	Dev	399	924	2188
	Test	400	871	2023
WikiEvents	Train	206	3241	4542
	Dev	20	345	428
	Test	20	365	556

Table 2: Dataset statistics of ACE2005

Dataset	Split	#Sent	#Event	#Argument
ACE2005	Train	17,172	4202	4859
	Dev	923	450	605
	Test	832	403	576

4 Experimental Setups

4.1 Datasets

Our experiments are conducted on the three widely-used datasets for event argument extraction task: ACE2005 (Doddington et al., 2004), RAMS (Ebner et al., 2020) and WiKiEvents (Li et al., 2021). The ACE2005 dataset is a sentence-level extraction dataset that defines 33 different event types and 35 semantic roles. The split of training, validating, and testing sets is the same as (Wadden et al., 2019). The RAMS dataset focuses on a document-level argument extraction task, including 139 event types and 65 semantic roles. The WikiEvents dataset is another document-level argument extraction dataset, 246 documents are provided, with 50 event types and 59 argument roles. Our experiments are conducted under the annotations of their conventional arguments. The statistics of the datasets are listed in Table 1 and Table 2.

4.2 Evaluation Metric

Following Ma et al. (2022), we adopt two popular evaluation metrics. (1) Argument Identification F1

Table 3: The main results of event argument extraction. The best results are marked with **bold**.

	ACE2005		RAMS		WikiEvents		
	Args-I	Args-C	Args-I	Args-C	Args-I	Args-C	Head-C
FEAE	-	-	53.50	47.40	-	-	-
DocMRC	-	-	-	45.70	-	43.30	-
OneIE	65.90	59.20	-	-	-	-	-
EEQA	68.20	65.40	46.40	44.00	54.30	53.20	56.90
BART-Gen	59.60	55.00	50.90	44.90	47.50	41.70	44.20
EEQA-BART	69.60	67.70	49.40	46.30	60.30	57.10	61.40
PAIE	73.60	69.80	54.70	49.50	68.90	63.40	**66.50**
UnifiedEAE	**76.06**	**71.85**	**55.46**	**49.94**	**69.84**	**64.00**	66.30
UnifiedEAE (Zero-shot)	42.25	34.60	10.88	8.49	30.27	25.90	40.72
UnifiedEAE (Single)	72.77	68.82	53.32	48.29	68.31	63.40	66.16
UnifiedEAE (Multiple)	74.34	70.80	54.62	49.09	67.63	62.66	66.28

score (Arg-I): we consider an argument span is correctly identified when the predicted offset fits the golden-standard span. (2) Argument Classification F1 score (Arg-C): if both the span and the argument role type are matched with the golden standard, we consider the argument is correctly classified. For the WikiEvents dataset, we also additionally evaluate Argument Head F1 score (Head-C) that only considers the matching of the headword of an argument, the same as (Li et al., 2021).

4.3 Baselines

To investigate the effectiveness of our model, we compare our approach with the following state-of-the-art baseline models.

- ONEIE (Lin et al., 2020) is a joint neural model to extract the globally optimal IE result.
- BART-Gen (Li et al., 2021) proposes a document-level neural event argument extraction model by regarding this task as conditional generation based on event templates.
- EEQA (Du and Cardie, 2020) proposes an end-to-end model and translates event extraction task into a question answering (QA) task.
- FFAE (Wei et al., 2021) constructs the EAE task as a QA-based algorithm and uses the intra-event argument interaction to improve the performance.
- DocMRC (Liu et al., 2021a) utilizes implicit knowledge transfer and explicit data augmentation based on a QA-based method.
- EEQA-BART (Ma et al., 2022) replaces the BERT with BART for event extraction.
- PAIE (Ma et al., 2022) utilizes prompt tuning for extracting argument extraction so that it can take the best advantages of pre-trained language models.
- UnifiedEAE is our full model, which trains on two datasets and transfers to one of them. For UnifiedEAE (Zero-shot), we train our model on two datasets and test on the third dataset via format-shared extractor.
- UnifiedEAE (Multiple) trains on the merged dataset, which removes the SSP from our UnifiedEAE model. In other words, it is a basic prompt-based extractor in Section 3.1. Different from UnifiedEAE (Multiple), UnifiedEAE (Single) trains and tests on the same dataset without transferring.

4.4 Implementation Details

We initialize the weight in encoder-decoder architecture with pre-trained BART base models (Lewis et al., 2020). We use Adam optimizer with the learning rates of 2e-5. The max encoder sequence length is 500, and the max decoder length is 80. The dropout is 0.1. The reported results on the test set are based on the parameters that obtain the best performance on the development set.

5 Results and Analyses

To investigate the efficacy of UnifiedEAE model, we compare our model with the mainstream baselines (Section 5.1). Then, we do the ablation studies

Table 4: The performance of transfer learning between RAMS and WikiEvents.

	RAMS		WikiEvents		
	Args-I	Args-C	Args-I	Args-C	Head-C
UnifiedEAE	**55.05**	49.71	**67.68**	**62.79**	**68.26**
w/o SharedVIB	54.65	48.79	65.90	61.05	67.42
w/o SSP	54.87	**49.92**	63.60	59.27	67.41
UnifiedEAE (Single)	53.32	48.29	68.31	63.40	66.16

Table 5: The performance of transfer learning between ACE2005 and WikiEvents.

	ACE2005		WikiEvents		
	Args-I	Args-C	Args-I	Args-C	Head-C
UnifiedEAE	**76.06**	**71.85**	**69.84**	**64.00**	66.30
w/o SharedVIB	75.12	71.48	68.76	63.27	**67.76**
w/o SSP	74.34	70.80	67.63	62.66	66.28
UnifiedEAE (Single)	72.77	68.82	68.31	63.40	66.16

Table 6: The performance of transfer learning between ACE2005 and RAMS.

	ACE2005		RAMS	
	Args-I	Args-C	Args-I	Args-C
UnifiedEAE	**71.65**	**68.00**	**55.46**	49.94
w/o SharedVIB	67.59	62.96	55.12	**50.02**
w/o SSP	62.61	59.62	54.62	49.09
UnifiedEAE (Single)	72.77	68.82	53.32	48.29

to verify the performance of the parts consisting of our model from two views, the model structure and dataset transferring (Section 5.2). We also explore the effectiveness of transfer learning on low-resource EAE (Section 5.3) and case studies are given (Section 5.4).

5.1 Main Results

In this section, we compare our framework with several prior competitive baselines (Table 3). Note, for UnifiedEAE and UnifiedEAE (w/o SSP), we report the best results of transferring over each two datasets (e.g., ACE2005 and RAMS, ACE2005 and WikiEvents, RAMS and WikiEvents).

From the table, we find the following observations. **First**, we observe that our model consistently outperforms the state-of-the-art baseline in terms of Args-I and Args-C. Compared with the best baseline PAIE, our approach outperforms it with an improvement of 2.46% in terms of Args-I over ACE2005, which indicates the effectiveness of multi-format transfer learning. **Second**, our model can leverage the knowledge from other datasets effectively. Our UnifiedEAE model outperforms UnifiedEAE (w/o SSP), which trains a basic prompt-based extractor using merged data directly. Moreover, UnifiedEAE (Single) sometimes performs better than UnifiedEAE (w/o SSP). These indicate that merging two datasets directly for training may bring noise and lead to small gains or even drops. **Third**, we apply our transfer learning framework to implement zero-shot. It is capable of extracting new event argument roles that are unseen in the training phase using the format-shared model. From the results, we can observe that transferring between ACE2005 and WikiEvents achieves a good performance because they have many similar event types and argument roles.

5.2 Ablation Studies

We do the ablation studies to further investigate the effectiveness of the main components in our model from two perspectives, model structure and resource information. The results are shown in Table 4, 5 and 6.

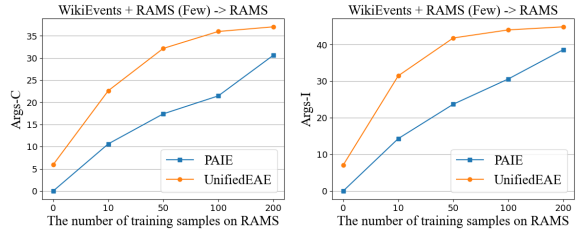
From a model structure view, we remove the SharedVIB (w/o SharedVIB) and SSP (w/o SSP) from our model respectively. We observe that each component can help boost the performance of EAE. Particularly, **1)** Removing SSP will cause about four points decline in terms of argument identification and classification when transferring the knowledge of RAMS to WikiEvents (Row 1,3 in Table 4). This justifies that directly merging datasets may bring noises, which results in the degradation of test data. Our model can learn both the format-shared and format-specific knowledge, which improves the performance effectively. **2)** Our SharedVIB strategy can enhance the model to learn the format-shared knowledge. Removing SharedVIB from our model will reduce the performance in most cases. For example, UnifiedEAE obtains more than 4 points improvement compared with the one without SharedVIB when transferring RAMS to ACE2005.

To investigate the effectiveness of transfer learning among different resources, we evaluate our model on each two datasets. As we mentioned above, UnifiedEAE (Single) trains and tests on the same dataset and UnifiedEAE without SSP trains on the merged datasets. We can find that not all the transferring can improve the performance since it may contain noise for the target dataset. For example, transferring from RAMS to ACE2005 caused more than six points drop for both Args-I and Args-C compared with UnifiedEAE (Single) that only trains on ACE2005 (Row 3 and 4 in Table 6). Our model can reduce the influence of noise effectively by learning both format-shared

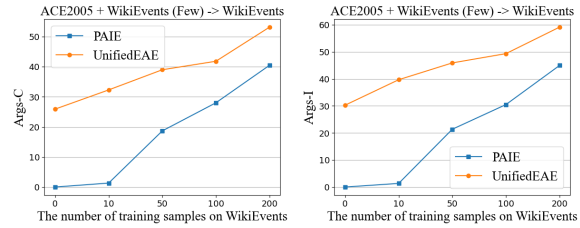
Train	Test	Business.Start-Org	Role	Shared	Specific	w/o SSP
ACE2005+ WikiEvents	ACE2005	**Founded**[trigger] by former mayor **Gholamhossein Karbaschi**[Agent] , **Hamshahri**[Org] was quick to become **Iran**[Place] 's biggest - selling daily with a circulation of 450,000 . It also built up healthy finances , carrying scores of pages of private advertisements daily.	Org	Hamshahri √	Hamshahri √	Gholamhossein Karbaschi ×
			Place	Iran √	Iran √	Iran √
			Agent	Gholamhossein Karbaschi √	Gholamhossein Karbaschi √	- ×

Train	Test	Personnel.Nominate	Role	Shared	Specific
RAMS+ WikiEvents	ACE2005	Suzanne I mean , I 'd like to s- -- I 'd like to see the **Greens**[Agent] **run**[trigger] David **Cobb**[Person] again.	Agent	Greens√	-
			Person	David Cobb√	-

Figure 3: Examples Visualization. We show the results of format-shared and format-specific extractors.



(a) Transfer from WikiEvents to RAMS (b) Transfer from WikiEvents to RAMS



(c) Transfer from ACE2005 to WikiEvents (d) Transfer from ACE2005 to WikiEvents

Figure 4: The results of low-resource event argument extraction via transfer learning.

and format-specific knowledge.

5.3 Low-Resource EAE

Furthermore, we explore the performance of low-resource event argument extraction via transfer learning (Figure 4). We transfer the knowledge from the source dataset (e.g., WikiEvents) to the target dataset (e.g., RAMS) with few samples in the target dataset. In our experiments, we train our model with 0, 10, 50, 100, and 200 samples. From the results, we obtain the following observations. First, `UnifiedEAE` significantly outperforms the state-of-the-art PAIE model in terms of both Args-C and Args-I over two datasets. Particularly, our model achieves over 30 points with only ten examples on WikiEvents in terms of F1, while the PAIE model is almost not working. Second, `UnifiedEAE` obtains better performance with fewer samples compared with PAIE. `UnifiedEAE` uses ten examples and performs even better than PAIE with 100 examples. Third, `UnifiedEAE` with 200 samples achieves the comparable results with the existing baselines (e.g., BART-Gen, EEQA-BART) trained on full training datasets (3241 samples) on WikiEvents. All these findings indicate that our model captures the format-shared and format-specific knowledge and transfers the format-shared information effectively.

5.4 Case Studies

To make it easier to understand how our `UnifiedEAE` model works, we visualize two examples on ACE2005 in Figure 3. We find that our `UnifiedEAE` model transfers the knowledge effectively. 1) The format-shared module extracts the arguments correctly by learning the overlapping knowledge among multiple formats' datasets. However, `UnifiedEAE` (w/o `SSP`), which trains on the merging dataset directly, can not predict "Org" and "Agent". 2) We also train our model on RAMS and WikiEvents and test it on ACE2005 using a format-shared extractor, which is under a zero-shot setting. We find our format-shared model can extract the event roles for the event "Personnel.Nominate" though it does not appear in the training dataset.

6 Conclusions and Future Work

In this paper, we propose a unified event argument extraction (`UnifiedEAE`) model to transfer the knowledge among multi-format datasets. First, a shared-specific prompt architecture is introduced to extract the event arguments with multiple formats based on both format-shared and format-specific representations. Then, to enhance the

model to capture the format-shared knowledge effectively, we integrate the information bottleneck into our architecture. Variational information bottleneck is leveraged to eliminate the format-specific information and retain the format-shared knowledge. We conduct extensive experiments on three EAE datasets and compare our model with several strong baselines. The results show that our `UnifiedEAE` model outperforms the state-of-the-art baselines. Furthermore, the ablation studies show `SharedVIB` can capture the format-shared effectively. Our model also obtains good results on low-resource event argument extraction. In further work, we would like to adopt our model to other complex tasks, such as relation extraction, and named entity recognition.

Acknowledge

The authors wish to thank the reviewers for their helpful comments and suggestions. This work was partially funded by National Natural Science Foundation of China (No. 61976056, 62076069), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

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.

Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, and Benjamin Van Durme. 2020. Reading the manual: Event extraction as definition comprehension. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 74–83.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. *arXiv preprint arXiv:2004.13625*.

Xinya Du, Alexander M Rush, and Claire Cardie. 2021. Grit: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Rui Feng, Jie Yuan, and Chao Zhang. 2020. Probing and fine-tuning reading comprehension models for few-shot event extraction. *arXiv preprint arXiv:2010.11325*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of NAACL*, pages 894–908.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. 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 2744–2754.

Jiaju Lin, Qin Chen, Jie Zhou, Jian Jin, and Liang He. 2022. Cup: Curriculum learning based prompt tuning for implicit event argument extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4245–4251. International Joint Conferences on Artificial Intelligence Organization.

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.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651.

Jian Liu, Yufeng Chen, and Jinan Xu. 2021a. Machine reading comprehension as data augmentation: A case study on implicit event argument extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2725.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. *arXiv preprint arXiv:2202.12109*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2021. Variational information bottleneck for effective low-resource fine-tuning. *arXiv preprint arXiv:2106.05469*.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 392–402.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1994. Distributional clustering of english words. *arXiv preprint cmp-lg/9408011*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. 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 313–325, Hong Kong, China. Association for Computational Linguistics.

Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and Philip S Yu. 2021. Graph structure learning with variational information bottleneck. *arXiv preprint arXiv:2112.08903*.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of EMNLP-IJCNLP*, pages 5784–5789.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4672–4682.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. 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 3752–3761.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

Qi Zhang, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. 2022. Enhancing event-level sentiment analysis with structured arguments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1944–1949.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jie Zhou, Yuanbin Wu, Qin Chen, Xuan-Jing Huang, and Liang He. 2021. Attending via both fine-tuning and compressing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2152–2161.