Applied Intelligence
https://doi.org/10.1007/s10489-024-05842-y

LegalATLE: an active transfer learning framework for legal triple extraction

Haiguang Zhang[1] · Yuanyuan Sun[1] · Bo Xu[1] · Hongfei Lin[1]

Accepted: 1 September 2024

Abstract

Recently, the rich content of Chinese legal documents has attracted considerable scholarly attention. Legal Relational Triple Extraction which is a critical way to enable machines to understand the semantic information presents a significant challenge in Natural Language Processing, as it seeks to discern the connections between pairs of entities within legal case texts. This challenge is compounded by the intricate nature of legal language and the substantial expense associated with human annotation. Despite these challenges, existing models often overlook the incorporation of cross-domain features. To address this, we introduce LegalATLE, an innovative method for legal Relational Triple Extraction that integrates active learning and transfer learning, reducing the model's reliance on annotated data and enhancing its performance within the target domain. Our model employs active learning to prudently assess and select samples with high information value. Concurrently, it applies domain adaptation techniques to effectively transfer knowledge from the source domain, thereby improving the model's generalization and accuracy. Additionally, we have manually annotated a new theft-related triple dataset for use as the target domain. Comprehensive experiments demonstrate that LegalATLE outperforms existing efficient models by approximately 1.5%, reaching 92.90% on the target domain. Notably, with only 4% and 5% of the full dataset used for training, LegalATLE performs about 10% better than other models, demonstrating its effectiveness in data-scarce scenarios.

Keywords Natural language processing · Relational triple extraction · Active learning · Transfer learning

1 Introduction

Relational Triple Extraction (RTE) is a cornerstone task in the field of Natural Language Processing (NLP), dedicated to transforming unstructured textual data into structured data formats that can be readily utilized by downstream applications. The structured data facilitates advanced functionalities such as intelligent question answering [1, 2] and the construction of knowledge graphs [3, 4], which are essential for various NLP tasks. RTE encompasses two critical stages: Named Entity Recognition [5, 6], where the goal is to identify and classify entities within the text, and Relation Extraction [7, 8], which is paramount for discerning the semantic connections between the named entities. In the legal domain, the significance of RTE is amplified as it aids in the judicious decision-making process of judges and is instrumental in the assembly of extensive legal knowledge graphs. These graphs are invaluable for legal analytics, providing a structured representation of legal concepts and their interrelations. However, the legal domain poses unique challenges to RTE. The escalating demand for legal NLP applications is often met with the hurdle of data scarcity and the arduous task of annotation [9]. The paucity of legal data, coupled with the inconsistent quality of annotations, introduces a layer of complexity due to the intricate nature of legal language. This complexity renders many existing models inadequate for real-time legal applications, which are increasingly sought

✉ Yuanyuan Sun
syuan@dlut.edu.cn

Haiguang Zhang
haiguang@mail.dlut.edu.cn

Bo Xu
xubo@dlut.edu.cn

Hongfei Lin
hflin@dlut.edu.cn

[1] Department of Computer Science and Technology, Dalian University of Technology, No. 2 Linggong Road, Dalian, Liaoning Province 116024, P.R. China

Published online: 02 October 2024

after in the legal sector for their potential to expedite and enhance legal proceedings.

In the realm of practical legal scenarios, criminal law encompasses over 400 distinct types of charges, with approximately 100 being prevalent. Notably, theft and drug trafficking cases constitute a significant portion of these common occurrences. Traditionally, joint extraction efforts were undertaken for drug-related cases within legal documents. However, the extensive number of cases renders the construction of dedicated datasets for each case a labor-intensive endeavor. Intriguingly, a comparative analysis reveals textual similarities between drug-related and theft cases. These parallels include the criminal relationships between suspects and items involved in crimes (e.g., “A” and “C” in Fig. 1) and the sale and disposal of such items (e.g., “B” and “D” in Fig. 1).

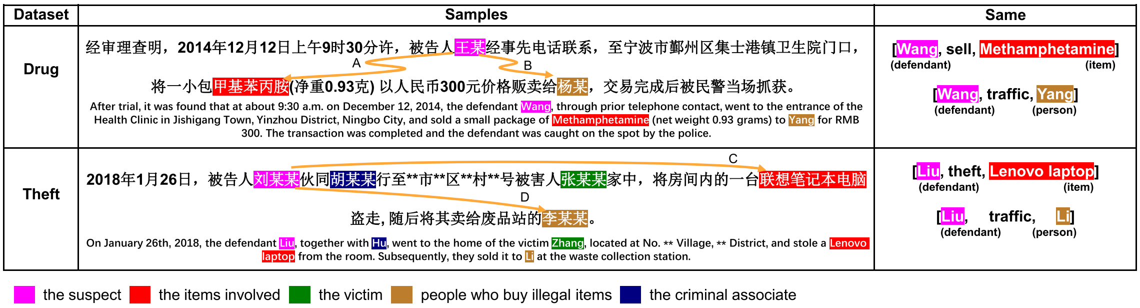
For instance, consider the sentences “被告人王某... 以人民币300元价格贩卖给杨某” and “被告人刘某某... 将其卖给废品站的李某某” in the Drug and Theft sections of Fig. 1, respectively. From a cross-domain perspective, two crucial facts emerge. First, the triple [“王某(Wang)”, traffic, “杨某(Yang)”] aids in extracting the triple [“刘某某(Liu)”, traffic, “李某某(Li)”], and vice versa, due to the remarkable similarity in their (subject, object) pairs: both subjects are defendants, and both objects are persons. Consequently, these entity pairs are highly likely to share the same kind of relation. Second, the mentioned triples along with [“王某(Wang)”, sell, “甲基苯丙胺(Methamphetamine)”] in the Drug section contribute to deducing a new triple [“刘某某(Liu)”, theft, “联想笔记本电脑(Lenovo laptop)”]. This deduction stems from the shared characteristics where both “theft” and “sell” involve subjects and objects as defendants and related items. These cross-domain features, beyond the scope of local features within a single domain, lead us to posit that texts exhibiting similar structural relationships can compensate for inadequate extraction results stemming from data scarcity.

Current methods primarily fill in relation tables using lower-level features extracted from individual token pairs [10, 11], sequential pipeline approaches that are prone to error propagation and exposure bias [12, 13], or a constrained set of signal domain features [14, 15]. Regrettably, these methods often overlook the incorporation of valuable cross-domain features, specifically the domain knowledge about token pairs and relations. Several researchers have explored the integration of transfer learning methods, as indicated by references [16, 17]. However, prevalent transfer schemes often demonstrate suboptimal performance due to significant data disparities and inherent model limitations. In the legal domain, the necessity for expert knowledge in data annotation exacerbates the resource-intensive nature of annotating extensive data. While certain approaches [18] have been investigated, their ability to improve model accuracy through fine-tuning methods is limited, and they do not adequately address the challenge of annotation difficulty, especially in specialized fields requiring in-depth understanding. The scarcity of annotated data in such cases impedes optimal performance.

To surmount these obstacles, we propose an Active Transfer Learning Framework for Legal Triple Extraction, termed LegalATLE, featuring cross-domain-oriented attributes within the Chinese legal domain. Leveraging a RTE model that populates relation tables based on cross-domain associations, LegalATLE aims to identify high-quality training samples using active transfer learning. The end-to-end model facilitates triple extraction, discerning relationships between entities within a sentence. Comprising three key components—an Active Selection Module (ASM), a Feature Extraction Module (FEM), and a Transfer Training Module (TTM)—the model undergoes evaluation using Drug and Theft datasets. Furthermore, to assess LegalATLE’s performance and showcase its efficacy, we have introduced a dataset comprising diverse examples encompassing a range of legal cases. The creation of our dataset entailed gathering theft cases, which were scrutinized to identify their constituent elements. For example, the Theft section of Fig. 1 encompasses suspects (e.g., “刘某某(Liu)”), items involved (e.g., “联想笔记本电脑(Lenovo laptop)”), victims (e.g., “张某某(Zhang)”), accomplices (e.g., “胡某某(Hu)”), and purchasers of illegal items (e.g.,“李某某(Li)”). The selection of cases was based on their relevance to theft-related criminal activities, as well as their availability and suitability for our research objectives. The experimental findings illustrate that this model can align data from two closely related domains into the same semantic space through a transfer mechanism, examining and contrasting the performance of various State-Of-The-Art (SOTA) models with BERT [18] and RoBERTa [19] across source and target domains.

The structure of this study is as follows: Following this introduction, the second section provides a comprehensive survey and synthesis of the extant literature pertinent to our domain. The third section elucidates our proposed approach, detailing the amalgamation of active learning with transfer learning and presenting an overview of the algorithmic framework. The fourth section delves into our experimental setup, where we introduce and characterize a dataset of legal theft-related triples, designed to benchmark the efficacy of our method. Comparative analysis substantiates the superiority of our approach, particularly under conditions of data scarcity. The study culminates with a conclusion that summarizes the salient contributions and limitations of our work, along with a discussion on potential avenues for future research. The contributions of this study are as follows:

- We introduce a novel active transfer learning framework tailored for the legal domain, significantly improving the



Dataset	Samples	Same
Drug	经审理查明，2014年12月12日上午9时30分许，被告人王某经事先电话联系，至宁波市鄞州区集士港镇卫生院门口，将一小包甲基苯丙胺(净重0.93克) 以人民币300元价格贩卖给杨某，交易完成后被民警当场抓获。 After trial, it was found that at about 9:30 a.m. on December 12, 2014, the defendant Wang, through prior telephone contact, went to the entrance of the Health Clinic in Jishigang Town, Yinzhou District, Ningbo City, and sold a small package of Methamphetamine (net weight 0.93 grams) to Yang for RMB 300. The transaction was completed and the defendant was caught on the spot by the police.	[Wang, sell, Methamphetamine] (defendant) (item) [Wang, traffic, Yang] (defendant) (person)
Theft	2018年1月26日，被告人刘某某伙同胡某某行至**市**区**村**号被害人张某某家中，将房间内的一台联想笔记本电脑盗走，随后将其卖给废品站的李某某。 On January 26th, 2018, the defendant Liu, together with Hu, went to the home of the victim Zhang, located at No. ** Village, ** District, and stole a Lenovo laptop from the room. Subsequently, they sold it to Li at the waste collection station.	[Liu, theft, Lenovo laptop] (defendant) (item) [Liu, traffic, Li] (defendant) (person)

🟪 the suspect 🟥 the items involved 🟩 the victim 🟫 people who buy illegal items 🟦 the criminal associate

Fig. 1 Presentation of the legal text structure of the two datasets. The connections “A” to “D” denote entities and their corresponding relationships. “sell(Drug)”: sell drug to. “traffic(Drug)”: traffic in

accuracy of legal RTE. This innovative approach attains 92.90% on the target domain, particularly excelling in scenarios with limited sample sizes.

- Our study encompasses the meticulous creation of a theft-themed dataset for RTE, meticulously annotated to encompass 2,200 legal cases and 7,483 distinct relationship instances. This dataset is unprecedented in its specificity and serves as a robust foundation for the development and assessment of supervised machine learning models for triple extraction, thereby enriching the legal domain’s dataset repository.
- The proposed methodology stands out for its enhanced capability to extract pivotal legal entities and organize information systematically, thereby augmenting the efficiency of the judicial case adjudication process. Its efficacy is particularly pronounced when processing extensive electronic legal case documentation, which underscores its practical utility in streamlining legal proceedings.

2 Related work

2.1 Joint entity and relation extraction for legal documents

RTE has evolved from early pipeline approaches that combined Named Entity Recognition and Relation Extraction to more sophisticated techniques addressing inherent limitations of initial methods. Early methods were plagued by error propagation issues resulting in suboptimal performance. Recent advancements in the general domain include tagging-based methods [12], which have improved accuracy, and novel frameworks [13, 20] designed to reduce redundancy in relationship extraction. Alternative techniques like span-based methods [21–23] and table-filling models [10, 11, 14] have also been introduced, offering increased flexibility and scalability. In legal domain, [24] proposed the method for joint entity and Relational Extraction in legal documents with legal feature enhancement by using a combination of rule and Machine Learning-based methods. Zhang et al. [25] introduced a joint entity and relation extraction for legal documents method based on table filling. Yet, the creation of large-scale evaluation corpora remains a challenge due to the dependency on extensive manual annotation.

2.2 Transfer learning

Transfer learning [16] is a promising strategy when limited data is available in the domain of interest, but a sizeable dataset is available in a related domain. Recently, transfer learning has made significant strides in the field of NLP. Ma et al. [26] considered the task of domain shift from pre-trained BERT to other target domains. Chan et al. [27] introduced an emotion classification method based on transfer learning, where pre-trained language models are transferred to different sentiment analysis tasks, achieving accurate classification. Khurana et al. [28] explored the application of transfer learning in cross-lingual text classification, effectively bridging the gap between high-resource and low-resource languages, which enhanced translation performance. In legal domain, [29] demonstrated that applying transfer learning to Named-Entity Linking networks can achieve good F1-scores on small and large legal datasets. Chen et al. [30] proposed an intelligent law article prediction method to address the data imbalance and missing value problems with transfer learning. Moro et al. [31] introduced a transfer learning approach with extractive and abstractive techniques to cope with the lack of labeled legal summarization datasets. Transfer learning offers an effective solution for addressing data scarcity issues and enhancing model generalization capabilities. However, challenges such as excessive transfer are unavoidable during its application.

2.3 Active transfer learning

Active learning has been extensively studied for its ability to maximize model performance with minimal labeling effort, focusing on selecting the most informative samples from an unlabeled dataset for labeling by an oracle [32, 33]. By combining active learning with transfer learning, researchers have developed active transfer learning strategies that optimize data usage and model effectiveness. These strategies use algorithms to intelligently select samples that are likely to provide the most new information to the model and apply pre-existing knowledge from related tasks to improve learning efficiency. Surveys [34, 35] have proposed integrated active transfer learning frameworks for efficient medical image recognition, significantly reducing the required annotations. Furthermore, novel active multi-source transfer learning algorithms for time series forecasting have been introduced to address the issue of negative transfer in single-source transfer learning [36]. A recent study [37] demonstrated the effectiveness of synergizing active and transfer learning for text categorization, achieving better performance with fewer labeled data points across various datasets.

3 Model framework

Here, we focus on RTE from legal texts, particularly in the context of theft cases within the legal domain. Given a sample x, extracted from a legal judgment and delineating the case facts, the objective of the triple extraction system is to ascertain $(e1, r, e2)$, wherein $e1$ and $e2$ denote entities in x, and r signifies their interrelationship. Accordingly, we adopt a method integrating active learning and transfer learning to tackle the challenge of data scarcity in triple extraction tasks, which consists of three main components: an Active Selection Module (ASM), a Feature Extraction Module (FEM), and a Transfer Training Module (TTM). The model structure is depicted in Fig. 2 and Algorithm 1.

Algorithm 1 LegalATLE framework for legal triple extraction with detailed processing.

```
Input:
$train\_epoch$ - Number of training epochs
$D$ - Dataset; $S$ - Source domain; $T$ - Target domain
$f_{ASM}$ - Section 3.1
$f_{FEM}$ - Section 3.2
$f_{TTM}$ -Section 3.3
$f_{DL}$ - Section 3.4
Output:
$Triple$ - Legal triple
procedure (LegalATLE)
    $D_{aug} \leftarrow chatgpt(D_s)$
    $S_0 \leftarrow D_{aug} \times isr$
    $T_0 \leftarrow D_T \times isr$
    for $i = 0$ to $epochs - 1$ do
        for $j = 0$ to length$(S_i) - 1$ do
            $C_i^S, O_i^{C_S} \leftarrow f_{FEM}(S_i[j])$; $C_i^T, O_i^{C_T} \leftarrow f_{FEM}(T_i[j])$
            $E_{i,j} \leftarrow f_{ASM}(O^{C_S})$
            $\mathcal{L_S} \leftarrow loss\_source(O_i^{C_S})$; $\mathcal{L_T} \leftarrow loss\_target(O_i^{C_T})$
            $\mathcal{L_D} \leftarrow f_{TTM}(C_i^S, C_i^T)$
            $\mathcal{L} \leftarrow f_{DL}(\mathcal{L_S}, \mathcal{L_T}, \mathcal{L_D})$
        end for
        if $(i + 1) \mod sf = 0$ then
            $S_{i+1} \leftarrow select_{Top}(D_{aug} - S_i, E_i, k) \cup S_i$
            $T_{i+1} \leftarrow select_{Random}(D_T - T_i, k) \cup T_i$
        else
            $S_{i+1} \leftarrow S_i$; $T_{i+1} \leftarrow T_i$
        end if
        $Triple_i^S \leftarrow triple(O_i^{C_S})$; $Triple_i^T \leftarrow triple(O_i^{C_T})$
    end for
end procedure
```

3.1 Active selection module (ASM)

Active and transfer learning approaches offer an ability to lower annotation effort by intelligently selecting the most informative examples to annotate [38] or by using existing labeled datasets [16]. However, most active learning approaches usually yield too few samples (on the order of hundreds) to feasibly fine-tune large deep-language models [39]. In terms of transfer learning, fine-tuning on out-of-domain data can lead to detrimental domain shift [26]. Furthermore, fine-tuning can also lead to over-fitting, especially in the case of smaller train sets, and to catastrophic forgetting of knowledge present in the pre-trained model [40].

As is well known, One notable application of machine learning is the generation of text [41]. ChatGPT is an exceptionally high-performing generative model. In this section, to match the quantities of source and target domain samples, we utilize ChatGPT to generate approximate samples for the source domain. We employ a specific template and its generative characteristics to obtain new expressions distinct from the original text. However, not all these are guaranteed to be logically coherent, semantically correct, and contribute positively. Therefore, we select reasonable samples from the generated ones, discard those with semantic deficiencies, and design an active learning module to select suitable samples. The original dataset containing N samples is defined as $D = \{x_1, x_2, \ldots, x_N\}$, and let x_i denote the i-th sample in it. Then, we use ChatGPT to augment each text and generate K new samples, $G = \{x_{N+1}, x_{N+2}, \ldots, x_{N+K}\}$. So, the augmented dataset can then be represented as $D_{aug} = \{x_1, x_2 \ldots, x_{N+K}\}$.

In RTE, active learning typically necessitates a substantial amount of annotated data to train high-performing models. However, the expense of manual annotation is considerable, posing challenges in enhancing model performance with limited annotated data. Furthermore, pool-based active learning introduces additional challenges, such as determining appropriate selection criteria. Various selection criteria may yield different sample selection outcomes, and their effectiveness may vary across distinct scenarios. Given these considerations, a method employing pool-based active learning is applied, which flexibly controls the number of samples from the sample pool and incrementally selects the most suitable

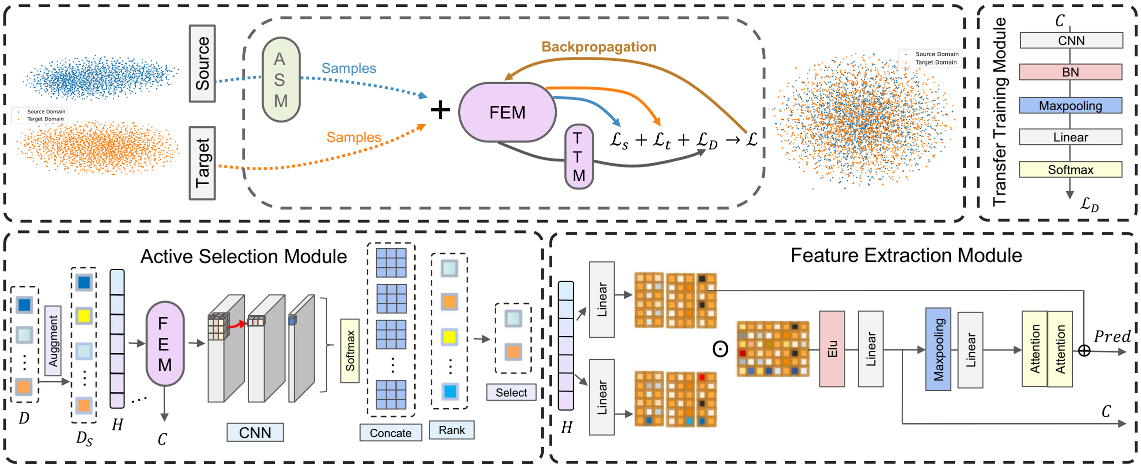


Fig. 2 Schematic overview of LegalATLE. "⊙" represents the Hadamard product. "**+**": new samples from the source and target domains are added to their respective training sets. The orange and blue scatter plots in the upper left corner represent the feature distributions of the source and target domains, respectively

samples during the training process. This approach allows us to fully utilize the limited annotated data to improve the model's performance. Specifically, we adopt uncertainty sampling as the sample selection strategy. A sample's uncertainty is measured by the entropy of its predicted probability distribution across all potential labels. Samples with the highest entropy are deemed the most informative. A sentence of length n is represented as $H = \{w_1, \ldots, w_n\}$. Then we can get the results $Pred$ predicted by FEM, which will be discussed in detail in Section 3.2. The formulas are as follows:

$$Pred_i = FEM(H) \tag{1}$$

Subsequently, we employ a convolution operation on the $Pred$ generated by the model to transform it into a two-dimensional tensor. However, the tensor solely represents the individual features of each sample. To convey the aggregate weight of each sample, we concatenate them and use the softmax function on the final dimension, yielding a probability list P, as follows:

$$Conv_i = Conv3d(Pred_i) \tag{2}$$

$$P_i = \frac{\exp(Conv_i)}{\sum_{j=1}^{K} \exp(Conv_j)}, \quad i = 1, \ldots, K \tag{3}$$

Here, $Conv3d$ represents the three-dimensional convolutional layer. K represents the number of all samples.

Entropy sampling's fundamental concept involves selecting samples with maximum entropy for model updating and conveys the randomness and uncertainty of a dataset. The higher the entropy value of a sample, the greater its randomness and information content. Therefore, we used this method to select suitable samples for training, as follows:

$$E_i = -\sum_{i=1}^{C} P_i \log P_i \tag{4}$$

where C denotes the number of sample categories.

Throughout the training process, the model selects top k samples (the proportion of data selected ($sp \in [0, 1]$)) from the sample pool every n iteration(the selection frequency ($sf \in \mathbb{Z}, sf \in [0, epochs]$)) with the largest entropy values, incorporating them into the training set.

$$S_{i+1} \begin{cases} select_{Top}(D_{aug} - S_i, E_i, k) \cup S_i, & \text{if } epoch \bmod sf = 0, \text{ and } i \in \left[1, \left\lceil \frac{1-isr}{sp} \right\rceil\right] \\ S_i, & \text{otherwise} \end{cases} \tag{5}$$

Here, i represents the current iteration. S_i represents the set after the selected samples are added in the i-th iteration. $isr \in [0, 1]$ represents the initial selection ratio of the model in the source domain. sp and sf and isr are super parameters.

We utilize the select function to derive a subset S_i from D_{aug} , denoted as D_S , and establish the target domain dataset as D_T. To better elucidate our approach, we depict the active learning sample selection structure in the ASM part of Fig. 2 and Algorithm 1.

3.2 Feature extraction module (FEM)

While real-world data is abundant, its annotation for important issues such as theft, drug, and various facets of legal cases are inherently challenging, because of the complexity of privacy protection and text understanding. As the primary component for training and encoding entities and relationships, we employ a structure similar to [11]. Given the multi-token nature of legal entity texts, we use three labels: $labels = \{N/A, EH, ET\}$, where N/A signifying no relationship between two character positions and EH and ET representing the head and tail positions of two entities, respectively. This definition allows for treating Named Entity Recognition and Relation Extraction as a unified task. To maintain a $n \times n$ size table, S is used for dimension transformation. After encoding the sentence text, we obtain:

$$H' = embedding(H) \otimes S \tag{6}$$

where $H' \in \mathbb{R}^{batchsize \times n \times n \times d_k}$, n refers to the maximum count of sentence tokens (including $[CLS]$ and $[SEP]$), d_k signifies the dimension of hidden layer, and $\otimes$ represents tensor product.

Then, to obtain a richer semantic representation, we pass H' through two different multilayer perceptions, yielding:

$$A = W_1 H' + b_1 \qquad B = W_2 H' + b_2 \tag{7}$$

where W_1, W_2, b_1, and b_2 are all optimizable parameters.

Although both layers are fully connected, the knowledge representations they obtain differ. We adopt the Hadamard product method to effectively integrate their information. A and B, both having dimensions $r \times s$. We implement block matrix multiplication for some products to conserve computational resources, as this technique can decrease memory consumption and enhance model training efficiency.

$$C_{i,j} = W(\sum_{k=1}^{n} A_{i,k} \odot B_{k,j}) + b, \{i = 1, 2, \ldots, r, j = 1, 2, \ldots, s\} \tag{8}$$

where $\odot$ represents the Hadamard product (element-wise multiplication) operation for matrices. n represents the number of blocks the matrix is divided into.

Subsequently, we implement a cross-attention mechanism, followed by multiple multi-head attention layers. For a given query vector q and a set of key-value pairs the cross-attention output of q with these pairs can be computed using the following formula:

$$CrossAttention\left(q, \{(k_1, v_1), \ldots, (k_n, v_n)\}\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k/h}} + M\right) V \tag{9}$$

Within this formula, $Q = qW_Q$, $K_i = k_i W_K$, $V_i = v_i W_V$, where Q, K, and V denote query, key, and value, respectively, W_Q, W_K, and W_V are learned parameter matrices. M is a unique mask matrix employed to exclude specific positions during attention weight calculations, $\sqrt{d_k/h}$ serves as a normalization factor for scaling dot product results, while h signifies the number of heads, and d_k refers to the hidden size.

Following several feature extraction iterations, we derive the tensor $P_r \in \mathbb{R}^{batchsize \times n \times r_n}$. The output tensor is a four-dimensional tensor with dimensions, in which the $labels$ indicates relationships between two entities. r_n signifies the product of the number of relation types and the number of labels in the dataset. More specifically, we define a loss function to evaluate the classifier performance in extracting triples from the source/target domain:

$$\mathcal{L}_{extract} = -\sum_{r=1}^{|r_n|}\sum_{i=1}^{n}\sum_{j=1}^{n} y_{r,(i,j)} \log \hat{P}_r(i,j) \left[y_{r,(i,j)} = \mathrm{P}_r(i,j)\right] \tag{10}$$

3.3 Transfer training module (TTM)

Several recent works have studied the use of few-shot instructions and in-context learning for lowering annotation efforts. They, however, focused either on sample-selection strategies [42, 43] or improving few-shot performance of smaller models [44, 45], but did not study transfer from existing pre-labeled datasets. Several works also employed various fine-tuning approaches in low resource settings [39, 46]. So to fully leverage the knowledge from the source domain in the target domain, we propose a domain adaptation method based on transfer learning techniques, named as TTM. We adopt a domain adaptation approach to align the feature distributions between the source and target domains. Next, batches of samples are selected from D_S and D_T, respectively, for the following steps.

Step1.To efficiently extract legal features from different samples across various domains, we employ a convolutional neural network layer [47], by sliding a convolution kernel over the input data and multiplying corresponding elements successively to obtain a locally weighted matrix.

$$Conv_i' = Conv2d(C_i) \tag{11}$$

Here, $Conv2d$ represents the two-dimensional convolutional layer.C denotes the input matrix from the Section 3.2.

Step2.Then, we incorporate the Batch Normalization (BN) layer to normalize the input of each layer in the deep neural network. The purpose of this is to address issues such as gradient vanishing and exploding during training, while also serving as a regularizer to mitigate the risk of overfitting, as:

$$BN = \frac{Conv' - \boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2 + \epsilon}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta} \tag{12}$$

where μ the mean, σ^2 the variance, ϵ a small positive number to prevent zero variance, and γ and β the learnable parameters.

Step3.To reduce dimensionality and alleviate overfitting, the pooling layer is used for downsampling to retain the most important features and is represented by:

$$PL_{i,j} = W(\max_{p=0}^{K-1} \max_{q=0}^{K-1} BN_{s\times i+p, s\times j+q}) + b \tag{13}$$

where K is the size of the pooling kernel and s is the stride. The linear transformation maps an input vector to a two-dimensional vector. W and b are the learnable parameters.

Step4. Finally, we obtain a normalized probability distribution from the K-dimensional vector z, using (14).

$$\text{softmax}(PL)_i = \frac{\exp(PL_i)}{\sum_{j=1}^{K} \exp(PL_j)}, \{i = 1, \ldots, K\} \tag{14}$$

where $\exp(\cdot)$ indicates the exponential function. This expression transforms each element of PL into a non-negative number through the exponential function, then normalizes them into a probability distribution, ensuring $\sum_{i=1}^{K} \text{softmax}(PL)_i = 1$.

3.4 Domain loss

More specifically, we pass samples through the classifier f_{FEM} and the domain classifier f_{TTM} to obtain the predicted label y and the domain label d, respectively. We use domain adversarial training, incorporating the output of the domain classifier as the weight in the loss function for gradient backpropagation and define an adversarial domain classification loss function to measure the classifier $f_{FEM}(x)$'s ability to differentiate between different domains. f_{TTM} is trained to minimize the classification error in distinguishing between the source and target domains. To prevent f_{TTM} from overfitting the source domain, the loss functions for the source and target domains of the triple are as follows:

$$\mathcal{L}_S = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}_{extract}(y_S, f_{FEM}(x_S))$$

$$\mathcal{L}_T = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{extract}(y_T, f_{FEM}(x_T)) \tag{15}$$

$$\begin{aligned}\mathcal{L}_\text{D} = &-\frac{1}{2}\mathbb{E}_{x_S \sim x_S}\left[\log f_{TTM}\left(f_{FEM}\left(x_S\right)\right)\right] \\ &-\frac{1}{2}\mathbb{E}_{x_T \sim x_T}[\log(1 - \ f_{TTM}\left(f_{FEM}\left(x_T\right)\right))]\end{aligned} \tag{16}$$

where m and n represent the number of samples in the source and target domains, respectively. x_S and y_S represent input samples and corresponding labels from the source domain, while x_T and y_T represent those from the target domain. Thus, our total loss function can be summarized as:

$$\mathcal{L} = \alpha_1\mathcal{L}_S + \alpha_2\mathcal{L}_T + (1 - \alpha_1 - \alpha_2)\,\mathcal{L}_D \tag{17}$$

where α_1 and α_2 are the super parameters. By jointly optimizing the $\mathcal{L}_S$, $\mathcal{L}_T$, and $\mathcal{L}_D$, we obtain domain-invariant features that capture triple extraction-related information in both domains. We then fine-tune the extracted features on a small amount of labeled data in the target domain to further enhance the triple extraction performance.

4 Experiment structure

4.1 Dataset building

In legal documents, similar structures are prevalent and constitute a significant portion of legal document. Deep learning has enabled the automatic analysis of specific entity relationships using model-based approaches. In this stage, the careful selection and annotation of the most common and crucial relationships prove to be valuable.

Practically, a case may involve multiple instances of the same relationship or coexist with various distinct or similar relationships. Consequently, to enhance the summarization of primary criminal behaviors and legal relationships involved in cases, we have categorized relationships into four types: "theft", "traffic", "possession", and "accomplice". Each relationship type is defined as follows: "theft" refers to the act of a suspect purloining an item, with the entities being the suspect and the purloined item; "possess" denotes the ownership relationship between the victim and the purloined item, with the entities representing the victim and the purloined item; "traffic" signifies the direction of the suspect's disposal of purloined goods, with the entities being the suspect and the individual procuring the purloined items; "accomplice" epitomizes the collaborative relationship between suspects, with the entities indicating their joint criminal relationship. Due to the large number of these relationships, it is of great significance to consider them as research subjects for legal intelligence.

To augment legal RTE, we meticulously annotated the fundamental criminal relationships in theft case legal texts using the CAILIE [48], which was originally obtained from the China Judgments Online(https://wenshu.court.gov.cn/). Before building, we conducted preprocessing steps, including data cleaning to remove any irrelevant or duplicate cases, tokenization, and handling special characters, to maintain data quality and consistency. The annotation process was carried out by legal volunteers with relevant experience in criminal law, ensuring the accuracy and reliability of the dataset. We annotated all samples involving the aforementioned four relationships, using Label Studio(https://labelstud.io/, guaranteeing that the annotated data is dependable for supervised training to assess the model's efficacy. The resulting dataset consists of 2,200 cases and 7,483 relationship instances, distributed across relationship types as shown in Fig. 3 and Table 2. It is worth noting that in the annotated dataset, several categories are not balanced, with (traffic) accounting for the least and only 1.4% of the data, while (theft), as the main criminal action in theft cases, accounts for nearly half of the proportion.

In the case of a small-sized dataset or a relatively uniform distribution, creating a validation set for model evaluation and adjustment may reduce the size of the training set, or may not provide additional benefits in evaluating the model's performance and generalization ability. Therefore, using only the training set and the test set for model performance and generalization ability evaluation may be a more appropriate and feasible choice. We partitioned the dataset into training and testing sets at a 9:2 ratio, ensuring sufficient data for model training and evaluation. However, it would be beneficial to justify this decision further or consider alternative validation strategies, such as k-fold cross-validation, to ensure that model performance is not overfitted to the training set. It would be helpful to discuss to address potential overfitting issues or whether data augmentation techniques were considered. To better verify the effectiveness of the model and simulate the absence of case samples with relationship labels (positive samples) in legal documents, we added some additional negative samples to the test set to balance the proportion of positive and negative samples to 2:1. We defined this as $Theft_{real}$. To gauge the model's effectiveness, we utilized F_1-score(F_1), $Precision(P)$, and $Recall(R)$ as evaluation metrics. These metrics are widely adopted in information extraction tasks and offer a comprehensive perspective on the model's performance.

Regrettably, the dataset utilized in this study cannot currently be made public due to data sensitivity and copyright restrictions. There is a possibility of releasing the dataset in the future. For any additional inquiries, please feel free to reach out to the authors directly. We appreciate your understanding and interest in our work.

4.2 Experimental setup

In this study, we employ two datasets: the drug-related legal dataset as the source domain and the theft dataset as the target domain. We derive the drug-related legal dataset from [24], while obtaining the theft dataset by re-annotating relationships in the entity dataset. Moreover, we compare the data of entities and relations between the source and target domains, presenting detailed information in Tables 1 and 2.

In Table 1, the specific meanings of the entities are as follows: "NH" represents the suspect, "NDR" denotes the type of drug, "NT" signifies the time of the crime, "NS" indicates the location of the crime, and "NW" stands for the weight of the drug. The meanings of the relations are as follows: "sell drug to" indicates that the suspect sells drugs to another suspect, with the head and tail entities being the buyer and seller, respectively; "traffic in" refers to the suspect illegally transporting or selling drugs, with the head and tail entities being the suspect and the drug, respectively; "provide shelter for" means that the suspect provides a place for another suspect to use drugs, with the head and tail entities being the two aforementioned individuals; "possess" signifies that the suspect possesses or conceals drugs.The total sample size of the drug-related dataset is 1768, of which the training set and test set are 1415 and 353 respectively. Additionally, Table 2 provides the specific meanings of the entities as follows: "NHCS" stands for the suspect, "NHVI" for the victim, "NASI" for the stolen item, "NT" for the time of the crime, "NS" for the location of the crime, "NATS" for the tool used in the crime, "NCGV" for the value of the item, "N0" for the related organization, "NCSP" for the amount obtained from selling the stolen goods, "NCSM" for the amount (Fig. 4) of stolen RMB, and "NHF" for the people who buy illegal items.

In the given context, the statement suggests that based on observations from Fig. 5a, b and c, the two datasets share a certain degree of similarity. This similarity could be in terms of the overall distribution of data and characteristics of individual samples within the datasets. The hypothesis proposed is that because of this observed similarity, the presence of drug-related samples may have a positive impact on the RTE task for theft samples. In other words, the characteristics or patterns identified in drug-related samples might be transferable or beneficial for improving the extraction of relational triples in the context of theft samples. This assumption implies that there could be shared features or relationships between the two types of samples that contribute positively to the extraction task.

Additionally, we have implemented our method using Python and the PyTorch framework. BERT and RoBERTa are utilized as the base model and fine-tuned on our dataset. The detailed parameter settings are shown in Table 3.

4.3 Evaluation standard

To evaluate our model's performance, we employ three metrics: P, R, and F_1. These metrics are widely used in information retrieval and NLP to provide a comprehensive evaluation of model performance. P represents the proportion of true positives among all predicted positives, while R represents the proportion of true positives among all actual positives. F_1 is the harmonic mean of P and R, offering a

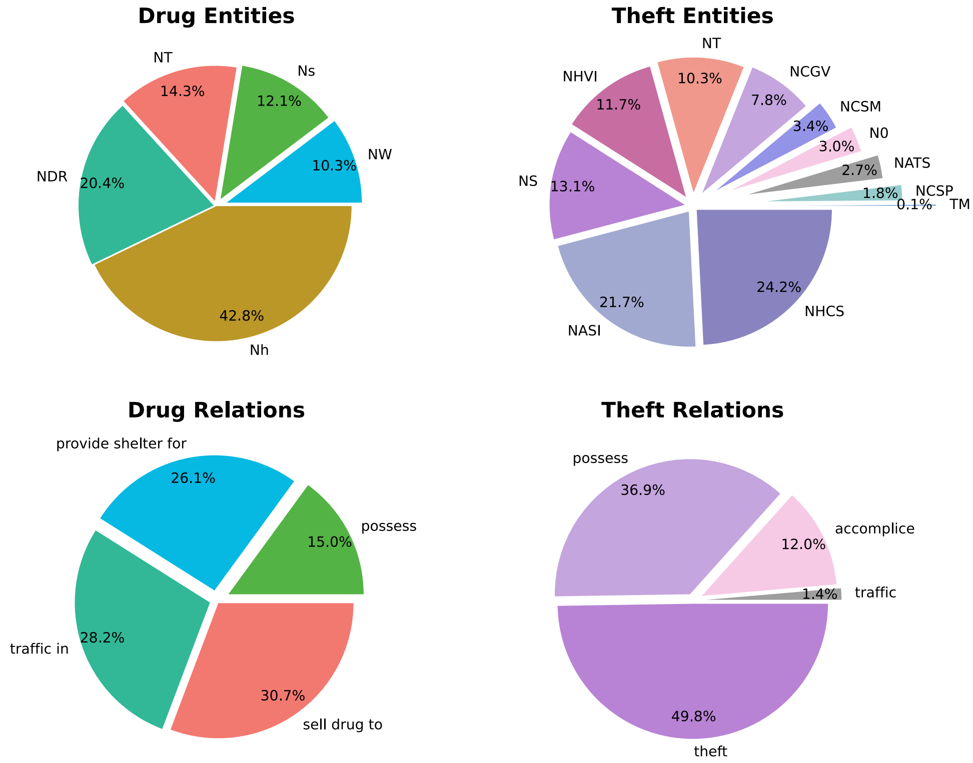


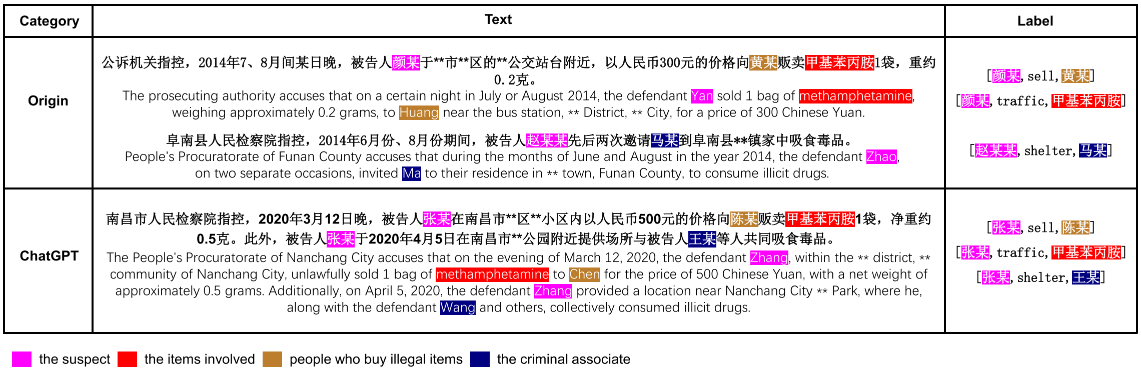
Fig. 3 The proportion of entities and relations in drug and theft datasets

Table 1 Entity, relationship, and average statistics for the drug dataset

Drug	Entity					Relation			
Name	Nh	NDR	NT	Ns	NW	Sell drug to	Traffic in	Provide shelter for	Possess
Count	4293	2751	2665	2154	1951	1231	1130	1043	589
Average	7.81					4.26			
Total	13814					7529			

Table 2 Entity, relationship, and average statistics for the theft dataset

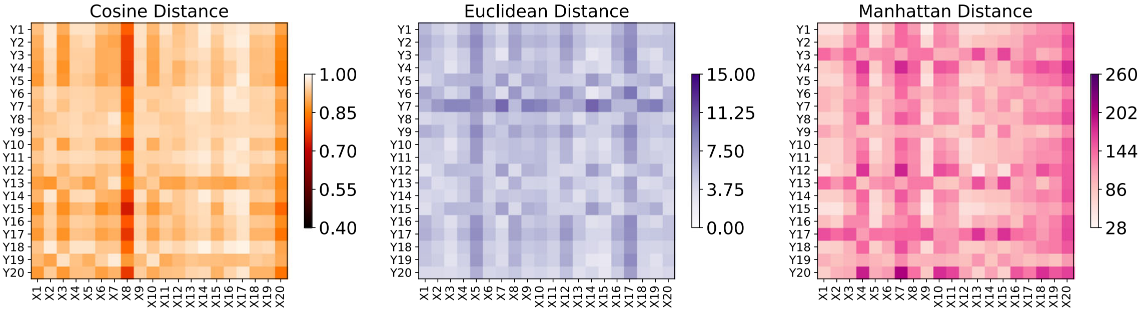
Theft	Entity											Relation			
Name	NHCS	NHVI	NASI	NT	NS	NATS	NCGV	N0	NCSP	NCSM	TM	Theft	Possess	Traffic	Accomplice
Count	6449	3114	5780	2756	3500	732	2086	805	481	913	31	3763	2704	94	922
Average	5.86											1.43			
Total	26647											7483			



Category	Text	Label
Origin	公诉机关指控，2014年7、8月间某日晚，被告人颜某于**市**区的**公交站台附近，以人民币300元的价格向黄某贩卖甲基苯丙胺1袋，重约0.2克。 The prosecuting authority accuses that on a certain night in July or August 2014, the defendant Yan sold 1 bag of methamphetamine, weighing approximately 0.2 grams, to Huang near the bus station, ** District, ** City, for a price of 300 Chinese Yuan. 阜南县人民检察院指控，2014年6月份、8月份期间，被告人赵某某先后两次邀请马某到阜南县**镇家中吸食毒品。 People's Procuratorate of Funan County accuses that during the months of June and August in the year 2014, the defendant Zhao, on two separate occasions, invited Ma to their residence in ** town, Funan County, to consume illicit drugs.	[颜某, sell, 黄某] [颜某, traffic, 甲基苯丙胺] [赵某某, shelter, 马某]
ChatGPT	南昌市人民检察院指控，2020年3月12日晚，被告人张某在南昌市**区**小区内以人民币500元的价格向陈某贩卖甲基苯丙胺1袋，净重约0.5克。此外，被告人张某于2020年4月5日在南昌市**公园附近提供场所与被告人王某等人共同吸食毒品。 The People's Procuratorate of Nanchang City accuses that on the evening of March 12, 2020, the defendant Zhang, within the ** district, ** community of Nanchang City, unlawfully sold 1 bag of methamphetamine to Chen for the price of 500 Chinese Yuan, with a net weight of approximately 0.5 grams. Additionally, on April 5, 2020, the defendant Zhang provided a location near Nanchang City ** Park, where he, along with the defendant Wang and others, collectively consumed illicit drugs.	[张某, sell, 陈某] [张某, traffic, 甲基苯丙胺] [张某, shelter, 王某]

🟪 the suspect 🟥 the items involved 🟫 people who buy illegal items 🟦 the criminal associate

Fig. 4 Comparative analysis of samples before and after ChatGPT enhancement. “sell”: sell drug to. “traffic”: traffic in. “shelter”: provide shelter for



(a) Cosine Similarity (b) Euclidean Distance (c) Manhattan Distance

Fig. 5 Analyzing multifaceted similarities: contrasting source and target domains through various metrics. The lighter the color, the more similarity between the texts. After randomly selecting samples from the theft and drug datasets, the x- and y-axes of graphs (a), (b), and (c) are populated accordingly

comprehensive evaluation of these two metrics. The formulas for calculating these metrics are as follows:

$$F_1 = 2 * P * R/(P + R) \tag{18}$$

$$P = TP/(TP + FP) \tag{19}$$

$$R = TP/(TP + FN) \tag{20}$$

Here, TP denotes the number of true positive samples, FP denotes the number of false positive samples, and FN denotes the number of false negative samples. In the relation triplet extraction task, true positive samples are the relation triplets correctly extracted by the model, false positive samples are the relation triplets incorrectly extracted by the model, and false negative samples are the relation triplets missed by the model. These metrics are used to evaluate the performance of the relation triplet extraction model.

4.4 Experimental outcomes and discussion

To demonstrate the effectiveness of our proposed model, we compare it with several strong baseline models for RTE:

- GRTE [11] is a global, feature-oriented model that utilizes two types of global associations by generating a table feature for each relation, mining global associations from these features, and integrating them back into the table features.
- TPlinker [10] is a one-stage joint extraction model capable of discovering overlapping relations sharing one or both entities while being immune to exposure bias.
- Onerel [14] is a method that casts joint extraction as a fine-grained triple classification problem. Specifically, it consists of a scoring-based classifier and a relation-specific tagging strategy.
- Spn [15] treats joint entity and relation extraction as a direct set prediction problem, solved by networks featured by transformers with non-autoregressive parallel decoding so that the extraction model is not burdened with predicting the order of multiple triples.
- DGCNN [49] introduces a novel operation for learning from point clouds, to better capture local geometric features of point clouds while still maintaining permutation invariance.
- Joint_Feature [24] can extract entities and semantic relations jointly, benefiting from the proposed entity feature and multi-task learning framework.
- Casrel [12] is a method on the NYT and WebNLG datasets based on the BERT backbone, which first identifies all possible head entities in a sentence and then identifies all possible relations and corresponding tail entities for each head entities.
- PRGC [13] decomposes joint extraction into three subtasks: Relation Judgement, Entity Extraction, and Subject-object Alignment, from a novel perspective.

We present the experimental results of LegalATLE and compare them with several methods on benchmark datasets, as shown in Table 4. It is divided into three blocks. The

Table 3 Experimental hyperparameters

Name	Batch size	Learning rate	Hidden size	Weight decay	α_1	α_2
Value	1	1e-5	768	0.01	0.4	0.4

methods in the first two blocks use the Chinese word representation from BERT and RoBERTa, and the last block is the ablation experiment. Our model achieves a significant and consistent performance boost over current SOTA models on the target domain, achieving an F_1 of 87.09 for $Drug$, 92.90 for $Theft$, and 90.95 for $Theft_{real}$. The results demonstrate the effectiveness of LegalATLE for legal relation extraction.

To further assess the effectiveness of LegalATLE for triplet extraction, we conduct ablation studies on the aforementioned datasets. We compare the performance of our method with five baseline methods: (1) active learning without augmented data, (2) without active and selection module, (3) active learning without transfer learning, and (4) a method combining active learning and transfer learning but excluding our proposed domain loss, (5) without active learning and transfer learning.

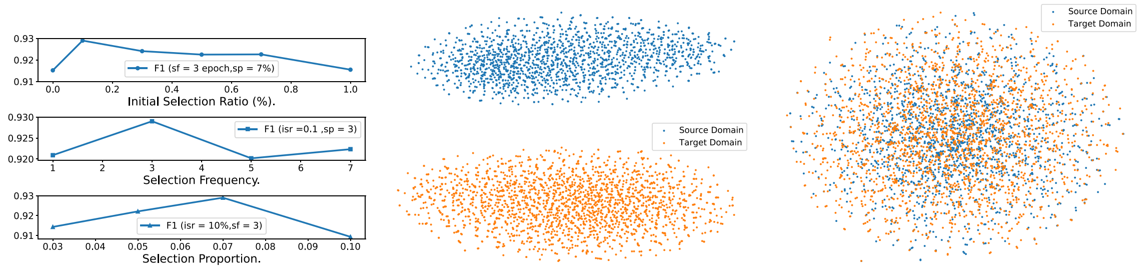
The traditional transfer learning method without active learning achieves a lower F_1 than LegalATLE, indicating the importance of active learning in improving triplet extraction performance. The active learning method without transfer learning achieves a comparable F_1 to the traditional transfer learning method, highlighting the importance of transfer learning for triplet extraction. Finally, the method combining active learning and transfer learning but without our proposed domain loss achieves a lower F_1 than LegalATLE, demonstrating the effectiveness of our domain loss in capturing domain-invariant features for triplet extraction. Additionally, we also apply this framework to the Theft-to-Drug scenario, demonstrating the effectiveness and good generalization capability of LegalATLE. Overall, our research results demonstrate the effectiveness of LegalATLE for triplet extraction, highlighting the importance of incorporating active learning and transfer learning in real-world applications.

In our experiments, we demonstrate significant changes by plotting the scatter distributions of source and target data before and after training as Fig. 6b and c. Before training, the source and target data exhibited separate clusters in the feature space. However, post-training, they converge, indicating that the model training has made the distributions of the two domains more similar. The rationale behind these changes can be attributed to the domain adaptation tech-

Table 4 LegalATLE performance on source and target domains with comparisons to SOTA efficient models trained with BERT and RoBERTa (%)

Methods	Source: Drug			Target: Theft			Target: Theft$_{real}$		
	F_1	P	R	F_1	P	R	F_1	P	R
GRTE [11]♣	88.41	87.87	**88.95**	91.37	**93.46**	89.36	89.46	89.64	89.27
TPlinker [10]♣	84.50	88.50	80.85	90.92	90.20	91.64	87.51	90.94	84.33
Onerel [14]♣	85.23	91.46	79.79	90.40	89.21	91.62	88.14	88.97	87.34
Spn [15]♣	82.94	88.74	77.85	76.17	84.03	69.65	71.19	73.50	69.02
DGCNN [49]♣	81.20	78.99	83.54	75.20	71.78	79.13	72.98	68.55	78.02
Joint_Feature [24]♣	83.10	85.80	80.60	86.82	88.31	85.37	86.16	88.84	83.64
Casrel [12]♣	84.70	89.25	80.59	91.14	92.07	90.24	89.45	87.78	91.19
PRGC [13]♣	89.41	90.37	88.46	89.20	91.60	86.92	89.00	87.53	90.52
LegalATLE♣⋆	87.09	91.63	82.98	**92.90**	92.57	**93.23**	**90.95**	91.06	**90.84**
LegalATLE♣⋆←	**89.48**	**94.00**	85.37	90.72	90.36	91.08	88.82	**92.88**	85.11
GRTE [11]♠	89.42	92.47	86.57	91.69	89.91	93.54	89.58	88.67	90.51
TPlinker [10]♠	86.56	90.78	82.71	90.54	90.40	90.68	89.36	87.23	**91.60**
Onerel [14]♠	83.09	88.64	78.19	90.16	93.80	86.78	88.50	86.42	90.68
Spn [15]♠	84.51	89.87	79.75	76.47	84.76	69.65	75.32	75.19	69.65
DGCNN [49]♠	81.33	83.08	79.65	76.99	76.07	77.94	76.22	73.86	78.73
Joint_Feature [24]♠	86.84	90.61	83.38	89.98	89.20	90.76	87.91	91.51	84.57
Casrel [12]♠	85.07	89.93	80.72	90.65	89.84	91.48	89.86	89.02	90.71
PRGC [13]♠	89.80	91.20	**88.40**	91.20	88.40	**94.20**	89.00	90.38	87.66
LegalATLE♠⋆	**89.81**	**93.23**	86.62	**92.84**	**94.60**	91.14	**90.79**	**92.99**	88.69
LegalATLE$_{w/oAD}$♣⋆	79.97	85.48	75.13	91.88	90.18	93.63	89.54	89.90	89.18
LegalATLE$_{w/oASM}$♣⋆	82.34	86.64	78.46	91.29	92.48	90.12	89.00	91.91	86.27
LegalATLE$_{w/oTTM}$♣	80.54	86.08	75.66	91.19	87.83	94.82	89.45	89.89	89.01
LegalATLE$_{w/oDL}$♣	83.19	87.06	79.65	91.61	91.50	91.72	89.18	90.52	87.88
LegalATLE$_{w/o(ASM,TTM)}$♣⋆	81.13	86.58	76.33	90.62	90.48	90.76	89.24	90.80	87.72

"$_{w/oAD}$" corresponds to active learning without augmented data. "DL" represents domain loss. Bold marks the highest score. "♣" represents using BERT. "♠" represents using RoBERTa. "⋆" represents using transfer method. "←" represents swapping the source and target domains



(a) Active learning super parater-meters discussion.

(b) Initial distribution

(c) Final distribution

Fig. 6 Active learning parameters and distributions shifts

niques employed during our model training. By introducing adversarial learning or other domain adaptation methods during the training process, we successfully reduced the distribution discrepancy between the source and target data in the feature space. This adaptation allows the model to better accommodate the characteristics of the target domain data, thereby improving performance on the target domain.

4.5 Active selection module discussion

As depicted in Fig. 4, we employed ChatGPT to generate novel samples and restructure their sample framework using pre-existing data. Specifically, within the given context, the individuals, locations of incidents, and weights of drugs in the newly generated samples exhibit variations. Moreover, the novel samples introduce an augmented number of triplets, markedly amplifying the volume of information and instigating a reorganization of the semantic structure. Nevertheless, to ascertain the authentic enhancement of semantic information in the samples, a more in-depth computational analysis is imperative. For this purpose, we applied six commonly used quantification standards (details provided Table 5). The findings reveal alterations across the aforementioned metrics in the enhanced text, corroborating my assertion that it has genuinely been enriched and fortified at the semantic level.

To explain the efficiency of the ChatGPT template, we choose three templates to do the test. Figure 7 illustrates the result of that. Moreover, we compare the active learning component of ASM with several methods. Please refer to Table 6. The methodology proposed in this article achieves the highest F_1 among all comparison methods, with a significant improvement of 0.63% compared to the second-best method. Notably, our method also outperforms other methods in terms of P and R. The significant performance improvement can be attributed to the effective selection of informative samples from the source domain and their use in training the model on the target domain. Overall, the experimental results and analysis demonstrate the effectiveness and superiority of LegalATLE for triplet extraction in cross-domain settings.

Besides, we also focus on investigating the impact of hyperparameters in ASM on the model performance. As shown in Fig. 6a, the model achieves optimal performance when the initial selection ratio(isr) is 30%, the selection frequency(sf) is 3 epochs, and the data selection proportion(sp) is 7%. Consequently, different parameter values affect the results. Precisely tuning the hyperparameters to enable the model to compute more accurately is a critical area of research for both our team and the broader scientific community.

4.6 Analysis and discussion of varied sample proportions

To scrutinize the impact of data volume on transfer effects and assess the model's proficiency with small samples, we chose to juxtapose it with the domain adversarial transfer models, encompassing AlexNet [50] DANN [51], ADDA [52], and CDAN [53], as delineated in Table 7.

- AlexNet [50] is a deep convolutional neural network that significantly enhances classification performance on large-scale visual recognition tasks. It achieves this by incorporating multiple layers for feature extraction.

Table 5 Exploring the evolution: comparative metrics analysis before and after ChatGPT augmentation

Indicator	BLEU(%)	ROUGE-1(%)	ROUGE-2(%)	ROUGE-l(%)	BERT-score(%)	2-gram(AVG)	VC(%)
Value(Origin)	–	–	–	–	–	53.84	8.26
Value(ChatGPT)	26.92	37.83	24.52	37.83	95.61	54.01	11.10

"ROUGE-n" with F_1. "VC": vocabulary coverage. "AVG": average value for each sample

T1:	请改写以下句子： {text} Please rephrase the following sentence: {text}
T2:	请根据下面的样本重新改写，并生成一个的新的回答： {text1},{text2} Please rephrase and generate a new answer based on the sample below: {text1},{text2}
T3:	请根据我给的样本改写，并给我一个改写的回答: {text} Please rephrase according to the sample I gave and give me a rephrased answer: {text}

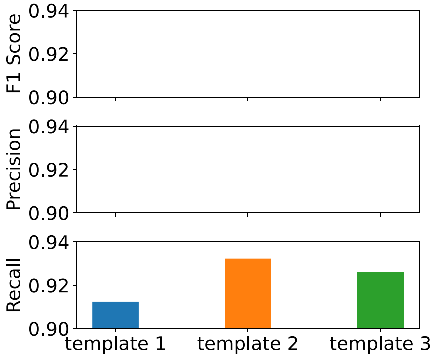


Fig. 7 Comparison of a variety of ChatGPT templates. “T1”: template1, “T2”: template2, “T3”: template3

- DANN [51] is directly inspired by domain adaptation theory, which posits that effective domain transfer requires predictions to be based on features that do not discriminate between the training (Source) and testing (Target) domains.
- ADDA [52] introduces a novel, generalized framework for adversarial adaptation. This framework not only encompasses recent SOTA methods as special cases but also enhances our understanding of previous approaches by offering a broader perspective.
- CDAN [53] is a principled framework that conditions the adversarial adaptation models on discriminative information conveyed in the classifier predictions.

Within the experimental framework, we systematically varied the quantities of source and target domain data across 10 discrete levels, ranging from 1% to 50%. The obtained results underscore that, whether under the BERT or RoBERTa architecture, the Legal model demonstrates markedly superior performance. This advantage is particularly salient in scenarios with diminished data volumes and becomes increasingly apparent as the data volume decreases. At lower data proportions, the LegalATLE model showcases superior adaptability, mitigating overfitting concerns faced by AlexNet and effectively leveraging adversarial training, which might be limited in DANN due to insufficient discriminative domain information. Moreover, ADDA’s performance at lower proportions is constrained by the quality of its generative network, a challenge overcome by LegalATLE. As data proportions increase, LegalATLE’s robust integration of inter-domain information and its ability to harness conditional information surpasses CDAN’s constraints and contribute to its overall optimal performance across a diverse set of proportions, positioning LegalATLE as a comprehensive and resilient solution for domain adversarial transfer tasks.

4.7 Implications of the study

Firstly, we introduce a novel active transfer learning framework specifically designed for legal Recognizing Textual Entailment (RTE), which achieves a notable accuracy of 92.90% in the target domain. This framework excels particularly in scenarios with limited sample sizes, offering a substantial advancement in the accuracy and reliability of RTE models. This theoretical advancement contributes to the broader understanding of transfer learning applications in specialized domains and sets a new benchmark for future research in legal RTE.

Secondly, we have developed a detailed theft-themed dataset comprising 2,200 legal cases and 7,483 distinct relationship instances. This dataset is unique in its specificity and serves as a valuable resource for developing and testing supervised machine learning models for triple extraction. By expanding the dataset repository in the legal field, we provide a robust foundation for more accurate and contextually relevant machine learning applications, benefiting researchers and practitioners who require high-quality data for their work.

Table 6 Comparing active learning sample selection strategies (%)

Method	Random	Uncertainty$_A$	Uncertainty$_B$	Similarity	Diversity$_C$	Diversity$_D$	Our Method
F_1	89.73	91.99	89.30	92.27	91.30	91.08	**92.90**
P	88.41	91.17	89.59	91.41	90.72	92.38	**92.57**
R	91.08	92.83	89.01	93.15	91.88	89.81	**93.23**

Bold marks the highest score. “A”: least confident. “B”: margin sampling. “C”: relation number. “D”: preselect number

Table 7 F_1 comparison between LegalATLE and domain-adversarial models across various data ratios from 1% to 50% (%)

Ratio	1%	2%	3%	4%	5%	10%	20%	30%	40%	50%
AlexNet [50]♣	26.00	34.22	48.44	64.67	70.89	77.17	78.33	80.56	81.78	82.00
DANN [51]♣	31.28	42.88	65.08	68.45	71.47	79.72	80.60	83.28	85.60	87.53
ADDA [52]♣	33.45	39.38	55.31	61.24	67.17	73.10	79.03	84.96	85.89	86.82
CDAN [53]♣	34.82	40.80	66.78	69.75	78.73	79.71	80.69	86.66	87.64	88.62
LegalATLE♣	39.47	50.25	67.56	**79.95**	**81.10**	82.88	**86.18**	88.96	89.23	89.30
AlexNet [50]♠	24.43	34.43	50.92	57.41	65.91	69.40	76.89	80.38	81.88	82.37
DANN [51]♠	30.03	47.30	66.49	68.81	71.33	79.92	81.19	83.47	85.67	87.96
ADDA [52]♠	34.08	49.97	65.87	61.76	57.66	63.55	69.45	75.34	81.24	87.13
CDAN [53]♠	35.94	41.78	67.62	73.46	79.30	80.13	80.97	81.81	82.65	88.49
LegalATLE♠	**40.02**	**59.37**	**74.90**	79.93	80.03	**83.61**	85.92	**89.63**	**89.65**	**89.70**

"♣" represents using BERT. "♠" represents using RoBERTa. GRTE [11] is used as a feature extractor

Finally, the proposed methodology enhances the extraction of key legal entities and systematic organization of information, thereby improving the efficiency of judicial case adjudication. This is particularly useful for processing extensive electronic legal case documentation, helping legal professionals manage large volumes of case data more effectively and streamlining case processing and decision-making.

5 Conclusion

This study introduces LegalATLE, a pioneering method for triple relational extraction within the legal domain, which integrates active learning and transfer learning to outperform existing approaches. The model's Active Selection Module markedly improves the efficiency of sample selection from the source domain, while the Feature Extraction Module adeptly distills pertinent features from both the source and target domains, thereby refining the feature representation. The Transfer Training Module is instrumental in transferring knowledge from the source to the target domain through extensive joint training, which is crucial for the model's enhanced performance.

Our contribution extends to the creation of a manually annotated dataset specific to theft-related cases in the Chinese legal context, tailored for RTE tasks. Empirical results from our experiments substantiate LegalATLE's efficacy, as indicated by its superior performance across a spectrum of evaluation metrics when juxtaposed with traditional models. Despite these advancements, we recognize the need for further empirical validation of our model's effectiveness across a diverse set of criminal offenses. In subsequent research endeavors, we intend to construct a universal text extraction framework capable of processing a wide array of criminal legal documents. Our research will be directed towards broadening the model's transferability to encompass an expansive range of criminal cases, extending beyond theft to include, for instance, cases of intentional injury. The ultimate objective is to facilitate the automated extraction of documents concerning multiple charges, which is essential for constructing an exhaustive legal knowledge graph. This initiative is central to propelling the field of legal informatics towards the sophisticated integration of legal intelligence.

Acknowledgements This work is supported by the National Key Research and Development Program of China (No. 2022YFC3301801), the Fundamental Research Funds for the Central Universities (No. DUT22ZD205).

Author Contributions Conceptualization: Haiguang Zhang and Yuanyuan Sun; Methodology: Haiguang Zhang and Yuanyuan Sun; Formal analysis and investigation: Haiguang Zhang, Yuanyuan Sun and Bo Xu; Writing - original draft preparation: Haiguang Zhang, and Bo Xu; Writing - review and editing: Yuanyuan Sun, Bo Xu and Hongfei Lin; Supervision: Yuanyuan Sun, Bo Xu and Hongfei Lin.

Data Availability and Access The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethical and Informed Consent for Data Used The article was submitted with the consent of all the authors to participate.

References

1. Wang L, Yu K, Wumaier A, Zhang P, Yibulayin T, Wu X, Gong J, Maimaiti M (2024) Genre: generative multi-turn question answering with contrastive learning for entity–relation extraction. Complex Intell Syst 1–15
2. Martinez-Gil J (2023) A survey on legal question-answering systems. Comput Sci Rev 48:100552
3. Yu H, Li H, Mao D, Cai Q (2020) A relationship extraction method for domain knowledge graph construction. World Wide Web 23(2):735–753

4. Yue Q, Li X, Li D (2021) Chinese relation extraction on forestry knowledge graph construction. Comput Syst Sci & Eng 37(3)
5. Li J, Sun A, Han J, Li C (2020) A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng 34(1):50–70
6. Liu P, Guo Y, Wang F, Li G (2022) Chinese named entity recognition: the state of the art. Neurocomputing 473:37–53
7. Guo Z, Zhang Y, Lu W (2019) Attention guided graph convolutional networks for relation extraction. In: Annual meeting of the association for computational linguistics, pp 241–251
8. Zhu H, Tiwari P, Zhang Y, Gupta D, Alharbi M, Nguyen TG, Dehdashti S (2022) Switchnet: a modular neural network for adaptive relation extraction. Comput Electrical Eng 104:108445
9. Sovrano F, Palmirani M, Vitali F et al (2020) Legal knowledge extraction for knowledge graph based question-answering. Front Artif Intell Appl 334:143–153
10. Wang Y, Yu B, Zhang Y, Liu T, Zhu H, Sun L (2020) Tplinker: single-stage joint extraction of entities and relations through token pair linking. In: Proceedings of the 28th international conference on computational linguistics, pp 1572–1582
11. Ren F, Zhang L, Yin S, Zhao X, Liu S, Li B, Liu Y (2021) A novel global feature-oriented relational triple extraction model based on table filling. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 2646–2656
12. Wei Z, Su J, Wang Y, Tian Y, Chang Y (2020) A novel cascade binary tagging framework for relational triple extraction. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 1476–1488
13. Zheng H, Wen R, Chen X, Yang Y, Zhang Y, Zhang Z, Zhang N, Qin B, Xu M, Zheng Y (2021) Prgc: potential relation and global correspondence based joint relational triple extraction. In: Proceedings of the 59th annual meeting of the association for computational linguistics, pp 6225–6235
14. Shang Y-M, Huang H, Mao X (2022) Onerel: joint entity and relation extraction with one module in one step. Proceedings of the AAAI conference on artificial intelligence 36:11285–11293
15. Sui D, Zeng X, Chen Y, Liu K, Zhao J (2023) Joint entity and relation extraction with set prediction networks. IEEE Trans Neural Netw Learn Syst
16. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. Proceedings of the IEEE 109(1):43–76
17. Tripuraneni N, Jordan M, Jin C (2020) On the theory of transfer learning: the importance of task diversity. Advances Neural Inf Process Syst 33:7852–7862
18. Devlin J, Chang M-W, Lee K, Toutanova K (2018) 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, pp 4171–4186
19. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2021) Roberta: a robustly optimized bert pretraining approach, 1218–1227
20. Ren F, Zhang L, Zhao X, Yin S, Liu S, Li B (2022) A simple but effective bidirectional framework for relational triple extraction. In: Proceedings of the Fifteenth ACM international conference on web search and data mining, pp 824–832
21. Dixit K, Al-Onaizan Y (2019) Span-level model for relation extraction. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5308–5314
22. Eberts M, Ulges A (2020) Span-based joint entity and relation extraction with transformer pre-training. In: Proceedings of the 28th international conference on computational linguistics, pp 88–99
23. Zhong Z, Chen D (2021) A frustratingly easy approach for entity and relation extraction. In: Proceedings of the 2021 Conference of the North American Chapter of the association for computational linguistics: human language technologies, pp 50–61
24. Chen Y, Sun Y, Yang Z, Lin H (2020) Joint entity and relation extraction for legal documents with legal feature enhancement. In: Proceedings of the 28th international conference on computational linguistics, pp 1561–1571
25. Zhang H, Qin H, Zhang G, Wang Y, Li R (2023) Joint entity and relation extraction for legal documents based on table filling. In: International conference on neural information processing, Springer, pp 211–222
26. Ma X, Xu P, Wang Z, Nallapati R, Xiang B (2019) Domain adaptation with bert-based domain classification and data selection. In: Proceedings of the 2nd workshop on deep learning approaches for low-resource NLP (DeepLo 2019), pp 76–83
27. Chan JY-L, Bea KT, Leow SMH, Phoong SW, Cheng WK (2023) State of the art: a review of sentiment analysis based on sequential transfer learning. Artif Intell Rev 56(1):749–780
28. Khurana S, Dawalatabad N, Laurent A, Vicente L, Gimeno P, Mingote V, Glass J (2024) Cross-lingual transfer learning for low-resource speech translation. In: IEEE International conference on acoustics, speech and signal processing (ICASSP)
29. Elnaggar A, Otto R, Matthes F (2018) Deep learning for named-entity linking with transfer learning for legal documents. In: Proceedings of the 2018 artificial intelligence and cloud computing conference, pp 23–28
30. Chen Y-S, Chiang S-W, Wu M-L (2022) A few-shot transfer learning approach using text-label embedding with legal attributes for law article prediction. Appl Intell 52(3):2884–2902
31. Moro G, Piscaglia N, Ragazzi L, Italiani P (2023) Multi-language transfer learning for low-resource legal case summarization. Artif Intell Law 1–29
32. Bernhardt M, Castro DC, Tanno R, Schwaighofer A, Tezcan KC, Monteiro M, Bannur S, Lungren MP, Nori A, Glocker B et al (2022) Active label cleaning for improved dataset quality under resource constraints. Nature Commun 13(1):1161
33. Citovsky G, DeSalvo G, Gentile C, Karydas L, Rajagopalan A, Rostamizadeh A, Kumar S (2021) Batch active learning at scale. Adv Neural Inf Process Syst 34:11933–11944
34. Zhou Z, Shin JY, Gurudu SR, Gotway MB, Liang J (2021) Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. Med Image Anal 71:101997
35. Taketsugu H, Ukita N (2023) Uncertainty criteria in active transfer learning for efficient video-specific human pose estimation. In: 2023 18th International Conference on Machine Vision and Applications (MVA), IEEE, pp 1–5
36. Gu Q, Dai Q (2021) A novel active multi-source transfer learning algorithm for time series forecasting. Appl Intell 51:1326–1350
37. Onita D (2023) Active learning based on transfer learning techniques for text classification. IEEE Access 11:28751–28761
38. Farinneya P, Pour MMA, Hamidian S, Diab M (2021) Active learning for rumor identification on social media. Findings of the association for computational linguistics: EMNLP 2021:4556–4565
39. Kasai J, Qian K, Gurajada S, Li Y, Popa L (2019) Low-resource deep entity resolution with transfer and active learning. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5851–5861
40. Fatemi Z, Xing C, Liu W, Xiong C (2023) Improving gender fairness of pre-trained language models without catastrophic forgetting. In: Proceedings of the 61st annual meeting of the association for computational linguistics, pp 1249–1262
41. Ahmad PN, Liu Y, Ullah I, Shabaz M (2024) Enhancing coherence and diversity in multi-class slogan generation systems. ACM Trans Asian Low-Resource Language Inf Process 23(8):1–24
42. Shin J, Kang Y, Jung S, Choi J (2022) Active instance selection for few-shot classification. IEEE Access 10:133186–133195

43. Yu Y, Zhang R, Xu R, Zhang J, Shen J, Zhang C (2023) Cold-start data selection for few-shot language model fine-tuning: a prompt-based uncertainty propagation approach. In: Proceedings of the 61st Annual meeting of the association for computational linguistics, pp 2499–2521
44. Gao T, Fisch A, Chen D (2021) Making pre-trained language models better few-shot learners. In: Proceedings of the 59th annual meeting of the association for computational linguistics, pp 3816–3830
45. Mishra S, Khashabi D, Baral C, Choi Y, Hajishirzi H (2022) Reframing instructional prompts to gptk's language. Findings of the association for computational linguistics: ACL 2022:589–612
46. Lee D-H, Kadakia A, Tan K, Agarwal M, Feng X, Shibuya T, Mitani R, Sekiya T, Pujara J, Ren X (2022) Good examples make a faster learner: simple demonstration-based learning for low-resource ner. In: Proceedings of the 60th annual meeting of the association for computational linguistics, pp 2687–2700
47. Zhang H, Zhang T, Cao F, Wang Z, Zhang Y, Sun Y, Vicente MA (2022) Bca: bilinear convolutional neural networks and attention networks for legal question answering. AI Open 3:172–181
48. Cao Y, Sun Y, Xu C, Li C, Du J, Lin H (2022) Cailie 1.0: a dataset for challenge of ai in law-information extraction v1. 0. AI Open 3:208–212
49. Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM (2019) Dynamic graph cnn for learning on point clouds. ACM Trans Graphics (tog) 38(5):1–12
50. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25
51. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V (2016) Domain-adversarial training of neural networks. J Mach Learn Res 17(59):1–35
52. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7167–7176
53. Long M, Cao Z, Wang J, Jordan MI (2018) Conditional adversarial domain adaptation. Adv Neural Inf Process Syst 31

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.