Zero-shot Transfer of Article-aware Legal Outcome Classification for European Court of Human Rights Cases

Santosh T.Y.S.S[1]**, Oana Ichim**[2]**, Matthias Grabmair**[1]
[1]School of Computation, Information, and Technology;
Technical University of Munich, Germany
[2]Graduate Institute of International and Development Studies, Geneva, Switzerland
{santosh.tokala, matthias.grabmair}@tum.de
oana.ichim@graduateinstitute.ch

Abstract

In this paper, we cast Legal Judgment Prediction on European Court of Human Rights cases into an article-aware classification task, where the case outcome is classified from a combined input of case facts and convention articles. This configuration facilitates the model learning some legal reasoning ability in mapping article text to specific case fact text. It also provides an opportunity to evaluate the model's ability to generalize to zero-shot settings when asked to classify the case outcome with respect to articles not seen during training. We devise zero-shot experiments and apply domain adaptation methods based on domain discrimination and Wasserstein distance. Our results demonstrate that the article-aware architecture outperforms straightforward fact classification. We also find that domain adaptation methods improve zero-shot transfer performance, with article relatedness and encoder pre-training influencing the effect.

1 Introduction

Legal Judgment Prediction (LJP) has recently gained considerable attention in the mainstream NLP community (e.g., Aletras et al. 2016; Chalkidis et al. 2019, 2021, 2022b; Santosh et al. 2022, 2023). In LJP, the outcome of a case should be classified/predicted based on a textual description of case facts. In actual legal reasoning, legal practitioners (e.g., advocates, judges) determine relevant rules from the sources of law (e.g., statutes, regulations, precedent) that are relevant to the case at hand. They then carry out an analysis to determine which rules apply to the case at hand, and deduce the outcome of the case by applying them. Subsuming case facts under elements of rules given in legal sources plays a critical role in this process. Many current LJP approaches (e.g., Aletras et al. 2016; Chalkidis et al. 2019, 2022b; Santosh et al. 2023) tackle this as a straightforward classification problem with the textual descriptions of case fact as the sole input. This reliance on the model learning statistical correspondences from case fact descriptions directly to outcomes neglects the role of legal sources in this relationship. As a consequence, the model may learn sub-optimal fact-outcome patterns that are informed by the case distribution in the data rather than learning to align facts with the legal source text containing applicable rules. The models may also attend to outcome-correlating distractors present in the dataset rather than engage in the legal fact-vs-law reasoning that is required of legal practitioners for a proper justification of the outcome (Santosh et al., 2022).

This work seeks to remedy this incomplete inference and enable the model to learn more authentic reasoning between rules and case facts by casting LJP into an article-aware classification setting and subjecting it to a zero-shot transfer challenge. Article-aware classification has been explored on Chinese criminal case corpora (Wang et al., 2018, 2019b; Yue et al., 2021; Chen et al., 2022). Similarly, Holzenberger et al. 2020 has modeled statutory reasoning by classifying US tax law provisions concatenated with textual case descriptions. We build on this prior work in two ways. First, we develop and evaluate our model on a public dataset (Chalkidis et al., 2022b) of cases by the European Court of Human Rights (ECtHR), which hears complaints by individuals about possible infringements of their rights enshrined in the European Convention on Human Rights (ECHR) by states. To the best of our knowledge, this is the first work applying article-aware case outcome prediction setting to human rights adjudication. Our approach pairs case fact descriptions with candidate ECHR articles and assigns a binary target label depending on whether the article has been alleged/deemed to have been violated, or not. Our results show that the article-aware classification model outperforms the traditional classification setup by a small but consistent margin.

605

Second, we subject the model to a zero-shot transfer task. Models trained on case facts alone cannot produce inferences about convention articles they did not observe during training. By contrast, human judges can conduct outcome analysis with new/amended legal provisions because they are trained to understand the rules they contain and apply them to case facts in an expertise-informed way, even in the absence of secondary sources (e.g., commentaries to the rule, etc.). Article-aware classification allows an emulation of this process by means of a zero-shot benchmarking task on articles unseen at training time. We compare two conditions where (1) the model either has no access to the target articles, or (2) it is allowed to 'read' the target articles but is not given any prediction outcome labels for case-target article pairs.

We experiment with domain adaptation by means of a domain discriminator (Ganin et al., 2016) and Wasserstein distance (Shen et al., 2018). Our results show that this improves performance on unseen articles compared to a vanilla model. We study the impact of law-specific pre-trained encoders on this zero-shot transferability compared to the standard language pre-trained one. Intuitively, we observe that our models perform better in zero-shot transfer if the target/unseen articles are semantically related to articles seen at training time.

It should be noted that, despite these tasks being typically referred to as instances of 'legal judgment prediction', ECtHR fact statements are typically not finalized until the decision outcome is known, making the task effectively one of retrospective classification rather than prediction (Medvedeva et al., 2021). While this does lead to distracting and confounding phenomena (see our prior work in Santosh et al. 2022), the dataset remains a useful resource for the development of NLP models that analyze these fact statements for text patterns that correspond to specific convention articles as drafted by the court. Consequently, in this paper we hence speak of our models as engaging in *case outcome classification (COC)*.

Our main contributions in this paper are[1]:

- We cast LJP/COC on ECtHR cases as an article-aware classification task by pairing case fact descriptions with candidate articles. Assuming a frozen pre-trained encoder network, our article-aware prediction model out-

performs straightforward fact classification.

- We conduct zero-shot transfer benchmarking of article-aware COC models. We find this to be a difficult testing task for the generalization of COC models. We show that domain adaptation using a domain discriminator and a Wasserstein distance method improves generalization.

- We conduct auxiliary experiments validating that article relatedness positively affects transfer performance and show an interaction between domain adaptation and domain specific encoder pre-training.

2 Related Work

Legal Judgement Prediction: LJP/COC as an NLP task has been studied using corpora from different jurisdictions, such as the ECtHR (Chalkidis et al., 2019, 2021, 2022b; Aletras et al., 2016; Liu and Chen, 2017; Medvedeva et al., 2020; SAYS, 2020; Medvedeva et al., 2021; Santosh et al., 2023) Chinese Criminal Courts (Luo et al., 2017; Zhong et al., 2018; Yang et al., 2019; Yue et al., 2021; Zhong et al., 2020), US Supreme Court (Katz et al., 2017; Kaufman et al., 2019), Indian Supreme Court (Malik et al., 2021; Shaikh et al., 2020) the French court of Cassation (Şulea et al., 2017b,a), Brazilian courts (Lage-Freitas et al., 2022), the Federal Supreme Court of Switzerland (Niklaus et al., 2021), UK courts (Strickson and De La Iglesia, 2020) and German courts (Waltl et al., 2017)

Early works (Aletras et al., 2016; Şulea et al., 2017a,b; Virtucio et al., 2018; Shaikh et al., 2020; Medvedeva et al., 2020) used bag-of-words features. More recent approaches use deep learning (Zhong et al., 2018, 2020; Yang et al., 2019). Large pre-trained transformer models have since become the dominant model family in COC/LJP (Chalkidis et al., 2019; Niklaus et al., 2021), including legal-domain specific pre-trained variants (Chalkidis et al., 2020; Zheng et al., 2021) that have been employed for the benchmark ECtHR corpus we use in this paper (Chalkidis et al. 2021, 2022b).

Prior work on Chinese criminal case corpora case extends fact-based classification by providing the text of legal source articles as additional input. Luo et al. 2017 used an attention-based neural network which jointly models charge prediction and relevant article extraction in a unified framework whose input includes the text of legal articles. Sim-

---

[1] Our code is available at https://github.com/TUMLegalTech/zeroshotLJP

ilarly, Wang et al. 2018, 2019b; Chen et al. 2022; Yue et al. 2021 employ matching mechanism between case facts and article texts. To the best of our knowledge, ours is the first work to adapt article-aware prediction to the ECtHR corpus, which is situated in the in human rights litigation domain. Going beyond previous works, we further benchmark the zero-shot transfer performance of such models, providing a test bed to evaluate their capability to process article texts they have not seen during training time and applying them to case facts towards classifying allegations/outcomes.

Domain Adaptation (DA): In transfer learning, the field of domain adaptation (DA) addresses the covariate shift between source and target data distributions (Ruder, 2019). It is tackled under three different settings: (1) Semi-supervised DA (Bollegala et al., 2011; Daume III and Marcu, 2006) where labels for the source and a small set of labels for the target domain are available, (2) unsupervised DA (Ganin et al., 2016; Blitzer et al., 2006) where only labels for the source domain and unlabelled target data are given, and (3) Any Domain Adaptation / Out of Distribution generalization (Ben-David et al., 2022; Volk et al., 2022) where only labeled source data is given. In this work, we distill the existing public LexGLUE ECtHR dataset into a new benchmark on more challenging unsupervised and any domain adaptation settings for COC to emulate legal reasoning involving previously unseen convention articles.

DA variants have been benchmarked for various NLP tasks, such as Question answering (Yu et al., 2018), duplicate question detection (Shah et al., 2018), sentiment analysis (Li et al., 2017; Ganin et al., 2016), dependency parsing (Sato et al., 2017), relation extraction (Wu et al., 2017), POS tagging (Yasunaga et al., 2018), named entity recognition (Jia et al., 2019), event trigger identification (Naik and Rose, 2020), machine reading comprehension (Wang et al., 2019a), and machine translation (Yang et al., 2018). To the best of our knowledge, this work is the first to benchmark domain adaptation in COC/LJP. While previous works typically involve short text, COC on ECtHR data involves case facts and articles, both of which typically are long documents.

Methods proposed for domain adaptation can be categorized into four types: (a) Instance-based data selection methods (Jiang and Zhai, 2007; Remus, 2012) which employ similarity metrics to sample source data points to match the distribution of the target domain and train models based on obtained subsamples from the source domain, (b) Pseudo-labeling approaches (Ruder and Plank, 2018; Rotman and Reichart, 2019) which train a classifier based on source data initially and use it to predict labels on unlabeled target data towards further adapting the model, (c) Pivot-based methods (Blitzer et al., 2006; Ziser and Reichart, 2017) which aim to map different domains to a common latent space (where the feature distributions are close) by employing auto encoders and structural correspondence learning, and (d) Loss-based methods (Ganin and Lempitsky, 2015; Shen et al., 2018) which employ domain adversaries aiming to minimize the discrepancies between source and target data distributions. In this woork, we employ loss-based approaches using a domain discriminator (Ganin et al., 2016) and Wasserstein distance (Shen et al., 2018) to enable domain adaptation for our COC models.

3 Dataset, Tasks & Settings

We use the LexGLUE ECtHR dataset provided by (Chalkidis et al., 2022b), which consists of 11k case fact descriptions along with target label information about which convention articles have been alleged to be violated (task B), and which the court has eventually found to have been violated (task A). The dataset is chronologically split into training (2001–2016), validation (2016–2017), and test set (2017-2019) with 9k, 1k, and 1k cases, respectively. The label set includes 10 prominent ECHR articles, which forms a subset of all the rights contained in the convention and its protocols. In both the ECtHR A and B benchmarks, it is assumed that the model classified the target from the fact description alone, which we refer to as the **fact classification variant**.

For our article-aware classification settings, we augment the dataset with the texts of the 10 articles copied from the publicly available ECHR convention document[2]. We formulate the **article-aware prediction variant** for both tasks: Given both the case fact statements and a particular article information, the model should classify the binary outcome of whether an article has been alleged to be violated by the claimant (task B) or found to have been violated by the court (task A).

---

[2]https://www.echr.coe.int/documents/convention_eng.pdf

Our zero-shot transfer task then involves determining violation/allegation from case facts with respect to articles which are not seen during training time. We consider a 'domain' to be a particular convention article (i.e., 10 convention articles form 10 domains). The objective is to train a model on a source domain (seen articles) with the goal of performing well at test-time on a target domain (unseen articles). Following (Yin et al., 2019; Ramponi and Plank, 2020), we propose two settings under zero-shot COC:

Zero-Shot Restrictive / Unsupervised Domain Adaptation (UDA): In this setting, the model is given a pair of case facts and the text of training set articles (i.e., the source domain) along with their corresponding violation/allegation outcome label. In the target domain, it is provided with case facts and article text pairs as well, but the outcome label is withheld. The goal of UDA is to learn an outcome classifier from the outcome labelled source domain which should generalize well on the target domain by leveraging outcome-unlabeled target data. This setting is legally realistic, as the text of new or modified written legal sources is typically known for a given task and available for domain adaptation (e.g., a public administration decision support tool receives an update after relevant legislation has changed).

Zero-shot Wild / Any Domain Adaptation (ADA) / Out of Distribution Generalization: In this setting, the model never sees any article data from the target domain during training, yet should be able to generalize to it. In the legal setting, this corresponds to a model which is required to work with texts of sources only available at query time (e.g., complex retrieval settings where multiple legal sources potentially apply).

We reorganize the dataset to evaluate our zero-shot transfer/adaptation models by splitting the 10 ECHR articles into two non-overlapping groups, such that both contain articles of various frequencies (common, moderate, rare).

- $split\_0$: 6, 8, P1-1, 2, 9

- $split\_1$: 3, 5, 10, 14, 11

We evaluate UDA and ADA on $split\_0$ as source and $split\_1$ as target, and vice-versa.

4 Method

We employ a hierarchical neural model which takes the case fact description x along with the article a as input and outputs a binary outcome (allegation in Task B and violation in Task A) for case x with respect to article a. Our architecture is a modified version of the Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017) incorporating conditional encoding (Augenstein et al., 2016; Rocktäschel et al., 2016) that has been adapted to deal with long input sequences following hierarchical attention networks (Yang et al., 2016). We experiment with two domain adaptation components based on adversarial training: (1) a classification-based domain discriminator and (2) a Wasserstein-distance based method which aims to reduce the difference between the source and the target domain distributions.

4.1 Article-aware prediction Model

Given the facts of the case $x = \{x_1, x_2, \ldots, x_m\}$ where $x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ and the article $a = \{a_1, a_2, \ldots, a_k\}$ where $a_j = \{a_{j1}, a_{j2}, \ldots, a_{jl}\}$, the model outputs a binary label. x_i / a_i and x_{jp} / a_{jp} denote the ith sentence and p^{th} token of the jth sentence of the case facts / article, respectively. m/k and n/l denote the number of sentences and tokens in the ith sentence of case facts / article, respectively. Our model contains an encoding layer, followed by an interaction layer, a post-interaction encoding layer, and a classification header. See Fig. 1 for an overview of our architecture.

4.1.1 Pre-interaction Encoding Layer

Our model encodes the facts of the case x sentence-wise with LegalBERT (Chalkidis et al., 2020) to obtain token level representations $\{z_{i1}, z_{i2}, \ldots, z_{in}\}$. These are aggregated into sentence level representations using token attention:

$$u_{it} = \tanh(W_w z_{it} + b_w) \tag{1}$$

$$\alpha_{it} = \frac{\exp(u_{it} u_w)}{\sum_t \exp(u_{it} u_w)} \ \& \ f_i = \sum_{t=1}^{n} \alpha_{it} z_{it} \tag{2}$$

where W_w, b_w and u_w are trainable parameters. The sentence level representations $\{f_1, \ldots, f_n\}$ are passed through a GRU encoder to obtain context-aware sentence representations of the facts $h = \{h_1, h_2, \ldots, h_m\}$. The analogous article encoder takes a as input and outputs $s = \{s_1, s_2, \ldots, s_k\}$.

4.1.2 Interaction Layer

Interaction between the sentences of the case facts and articles is done via dot product attention between the two sequences of sentences as follows:

$$e_{ij} = h_i^T s_j \quad \& \quad h_i' = \sum_{j=1}^{k} \frac{\exp(e_{ij})}{\sum_{l=1}^{k} \exp(e_{il})} s_j \qquad (3)$$

$$s_j' = \sum_{i=1}^{m} \frac{\exp(e_{ij})}{\sum_{l=1}^{m} \exp(e_{lj})} h_i \qquad (4)$$

where e_{ij} represents the dot product interaction score between the context-aware representations of the ith sentence of the case facts and the jth sentence of the article. h_i' and s_j' represent article-aware representations corresponding to the ith sentence of the case facts and the fact-aware representation corresponding to the jth sentence of the article, respectively. Finally, we obtain interaction-aware sentence representations of the facts $h' = \{h_1', h_2', \dots, h_m'\}$. Similarly for the article, we obtain $s' = \{s_1', s_2', \dots, s_k'\}$

4.1.3 Post-Interaction Encoding Layer

The article-dependent final representation of the case facts is obtained in two steps: (i) we compute the final representation of the article text and (ii) use it as a conditional encoding (Augenstein et al., 2016; Rocktäschel et al., 2016) to obtain the final article-dependent fact representation.

Final representation of article: We first combine the pre-interaction sentence encodings and fact-aware sentence representations of the article:

$$p_i = [s_i, s_i', s_i - s_i', s_i \odot s_i'] \qquad (5)$$

where $\odot$ denotes element-wise product. This representation aims to capture high-order interaction between the pre- and post- interaction elements (Chen et al., 2017). The sentence representations p_i are passed over a non-linear projection and a GRU (as in the pre-interaction encoder) to perform context-level modelling among sentence sequences. The final article representation A is obtained via sentence attention analogous to eq. 2.

Final Representation of Case Facts: Similarly, we pass the combined representation of case facts using pre- and post- interaction similar to Eq. 5 over a non linear projection, a GRU layer, and sentence level attention to the obtain article-dependent final representation of case facts. To ensure conditioning, we initialize the GRU hidden state with the final representation of the articles A. This facilitates capturing the salient case fact information with respect to the specified article.
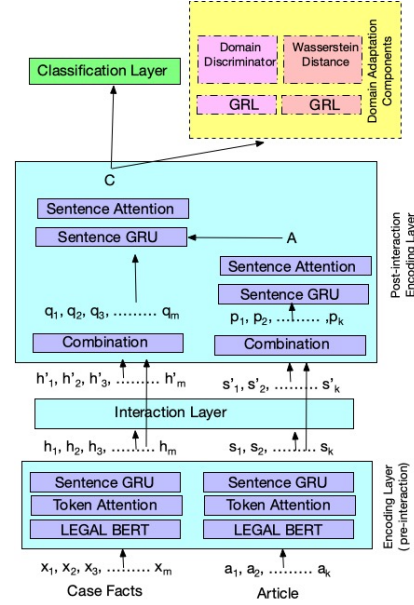

Figure 1: Our article-aware prediction model architecture

4.1.4 Classification Layer

We pass the article-dependent final representation of the case facts through a nonlinear projection to classify the outcome.

4.2 Domain Adaptation Components

Domain Adaptation aims to make models generalize well from a source to a target domain. Both domains are mapped to a common latent space, reducing differences between their distributions and facilitating domain invariant feature representations. In our case of article-aware COC, we regard reasoning with respect to every ECHR article as a domain and seek to learn article-invariant case facts representations. Put differently, we want our model to learn how to read two texts and interrelate them towards an outcome determination (as lawyers do) with minimal encoding of the information contained in the texts into the model itself. This way, the models can achieve generalization capability to adapt and perform reasoning with regard to articles not seen during training time.

4.2.1 Domain Discriminator

We employ a two layer feed forward network as a discriminator which takes the article-dependent case fact representation as input to predict the article (i.e., the domain). We train the discriminator in an adversarial fashion to maximize the model's ability to capture information required for the outcome task while minimizing its ability to

predict the article. This guides the model to generate article-invariant feature representations and improves transferability. Following (Ganin and Lempitsky, 2015; Ganin et al., 2016), we perform a min-max game adversary objective optimization using a gradient reversal layer (GRL) between the feature extractor and discriminator. It acts as the identity during the forward pass but, during the backward pass, scales the gradients flowing through by $-\lambda$, making the feature extractor receive the opposite gradients from the discriminator. The overall objective function reduces to:

$$\arg\min_{\theta_F, \theta_C, \theta_D} [L_c(C(F(x, a)), y_e) + \lambda L_d(D(\text{GRL}(F(x, a))), y_a)]$$

(6)

where L_c, L_d represents the loss function corresponding to classifier and domain discriminator, respectively, λ is the GRL hyperparameter, x is the input, y_e is the outcome label, y_a is the class-id of the article, F, C and D represents feature extractor, classifier, and discriminator with parameters θ_F, θ_C and θ_D, respectively. In case of UDA (where the model has access to the text of target domain articles), we discriminate among all the source and target articles. While in case of ADA, we discriminate among source articles only.

4.2.2 Wasserstein Method (Distance based)

Our second method aims to reduce the Wasserstein distance (Shen et al., 2018) between different domain feature distributions. In a given batch, the final feature representations will be fed into the domain critic (Arjovsky et al., 2017), which is a feed-forward network whose output is a single scalar for each batch element. These scalars are then averaged per domain in the batch, resulting in two numbers representing source and target domains, respectively. Their difference can be considered an approximation of the Wasserstein distance between the two feature distributions and becomes the Wasserstein loss component of the network. If the domain critic neural network satisfies the constraint of the Lipschitz-1 continuous function, we calculate the approximate empirical Wasserstein distance by maximizing the following domain critic loss:

$$L(X_p, X_q) = \frac{1}{n_p} \sum_{x_p \in X_p} f_w(F(x_p)) - \frac{1}{n_q} \sum_{x_q \in X_q} f_w(F(x_q))$$

(7)

where f_w, F denote the Wasserstein domain critic and feature extractor, respectively, X_p and X_q denote datasets from two domains p and q with n_p and n_q samples, respectively.

Table 1: Fact Classification vs Article-aware prediction Performance on Task A and Task B. mic. and mac. indicates micro-F1 and macro-F1 scores, respectively.

	Task B		Task A	
Model	**mac.**	**mic.**	**mac.**	**mic.**
Fact Classification	71.96	77.40	61.21	72.21
Article-aware pred.	**74.14**	**78.49**	**67.09**	**74.77**

During optimization, a gradient reversal layer (Ganin et al., 2016) between the feature extractor and domain critic ensures that (a) the domain critic weights are updated such that the Wasserstein loss becomes maximal, while the encoder weights are updated towards minimizing it. Through this procedure, we encourage the model to learn feature representations that are invariant to the covariate shift between the source and the target domain. Since the Wasserstein distance is continuous and differentiable everywhere, we can train the domain critic end-to-end. In case of UDA, we minimize the distance between the source and the target domains, while in case of ADA, we minimize among the different source domains. To enforce the Lipschitz constraints, we clip the weights of the domain critic within a compact space $[-c, c]$ after each gradient update following (Arjovsky et al., 2017).

5 Experiments & Discussion

5.1 Baseline

For the **fact classification variant**, we employ an architecture similar to the article-aware prediction model but reduced to the case fact based encoding without the interaction mechanism. The output layer is modified to 10 classes and trained against a multi-hot target vector using a binary cross entropy loss. Notably, we freeze the weights in the LegalBERT sentence encoder, both to save computational resources and to reduce the model's susceptibility to shallow surface signals and ensure the comparability of our domain adaptation methods. We describe the detailed hyperparameters for the article-aware prediction model in Appendix Sec. A

5.2 Does Article-aware Classification Perform Better than Fact-only Classification?

Micro-F1 and macro-F1 scores for both tasks A and B with regard to the 10 target articles are given in Table 1. The article-aware model performs better than fact-only classification across the board. In

Table 2: Task B F1 performance of baseline and domain adaptation models

| | | Transfer $0 \rightarrow 1$ | | | | Transfer $0 \leftarrow 1$ | | | |
| | | source : $split\_0$ | | target : $split\_1$ | | source : $split\_1$ | | target : $split\_0$ | |
Setting	Model	mac.	mic.	mac.	mic.	mac.	mic.	mac.	mic.
Baseline	Source only	73.45	75.63	7.32	7.37	70.26	77.10	8.49	9.08
UDA	Domain Disc.	**73.81**	**76.95**	**13.92**	14.94	**70.63**	**77.43**	**22.50**	26.27
	Wasserstein	69.63	74.86	13.17	**18.16**	66.89	75.21	20.78	**30.30**
ADA	Domain Disc.	**73.76**	**76.13**	**9.62**	**10.77**	**69.71**	**76.85**	**9.30**	**10.45**
	Wasserstein	70.17	74.80	9.14	9.89	67.46	75.25	9.26	10.38

particular, we notice a greater improvement in the macro-F1 score, indicating the article-aware classification approach helps the model to improve performance for sparser articles which are not prominently represented in the case distribution. We conjecture that this performance difference can be explained with article-aware classification being subjected to a different training regime. In fact-only classification, a given case's fact text will always be associated with the same multi-hot outcome vector. By contrast, in the fact-aware setting, it will occur multiple times alongside different article texts and the model is forced to predict a single binary outcome variable. This seems to lead the model away from shallow signals towards capturing fact-article correspondence, resulting in a better model. Additionally, the beneficial effect is greater for the harder task of violation classification.

5.3 Does Domain Adaptation Help to Improve Zero Shot Transferability ?

We evaluate UDA and ADA on both Task A and Task B with the two article splits. A baseline *source only* model is trained without domain adaptation using the labelled source data only and tested on the target test data directly. Tables 2 and 3 show the performance of different models with our two splits on task B and A, respectively.

Baseline* vs *Domain Adaptation: From both tables, we observe that the performance of the *source only* model on target data is lower compared to their domain adaptation counterparts with a significant margin. This indicates that, intuitively, models trained on source data without any adaptation do not generalize to unseen articles. This also highlights the need to have domain adaptation components to achieve a generalizable model.

UDA: Under unsupervised domain adaptation, we observe that the Wasserstein distance method performs better on target data than the Domain Discriminator in micro-F1 by a significant margin. It also improved macro-F1 marginally in Task A tar-

get data, but is inferior in Task B. Most strikingly, however, Wasserstein performance on source data is lower than the source only baseline across the board, especially with respect to macro-F1. These observations also indicate that the Wasserstein distance method is able to transfer well to certain articles more than others. This can be attributed to the method influencing feature representations towards a reduction of the mean difference across articles. The distribution of target articles which are closer to the source articles distributions might have gained well. We further validate this hypothesis using an experiment illustrated in sec 5.5. On source data, the Domain Discriminator performed better than the source only model, albeit by very small margins but consistent across the tables.

ADA: On target data, both the Domain Discriminator and Wasserstein distance are comparable across the tables in both metrics. With respect to source data, in task B, the Domain Discriminator performed better than the Wasserstein distance method in both micro and macro F1. Strikingly, in Task A, Wasserstein performance on source data picks up in micro-F1 (slightly even better than source only baseline) but stays behind in macro-F1.

ADA vs UDA: Unsurprisingly, the performance on target data under ADA tends to be lower compared to UDA due to no access to target article information in this setting compared to UDA.

The absolute performance levels on the target data immediately suggest that the zero-shot transfer task we propose is very difficult and the discrepancy of performance between source and target data is still large, even in the case of domain adaptation components. This indicates ample opportunity for further research on neural models capable of reasoning with legal text in a way that transfers well to unseen legal domains. Some of the source-target performance divergence can likely be attributed to the model falling prey to spurious correlations that exist in the data, which is especially prominent in the ECtHR datasets that suffer from fact statements

not being finalized until the case outcome is known (see our prior work in Santosh et al. 2022. Given this limitation, our zero-shot framework serves as a challenging benchmark in the development of legal NLP models that learn to interrelate case facts and legal source text towards supporting domain experts.

5.4 How does Encoder Pre-Training influence Zero-Shot Transferability?

We conduct an additional experiment on Task A with $split\_1$ as source and $split\_0$ as target, where we replace LegalBERT embeddings used in the encoding layer with BERT base embeddings (Kenton and Toutanova, 2019), and report its performance in Table 4. Comparing it to Transfer $0 \leftarrow 1$ in Table 3, we observe that the BERT base model performs worse on target data than the LegalBERT encoder. In particular, the best performing Wasserstein domain adaptation model drops from 26.2 to 16.36, much more than the Domain Discriminator. We leave an exploration of this asymmetric effect of the pre-training regime across different domain adaptation strategies to future work.

Base BERT performs similarly on the source domain. This indicates that even a non-legally pre-trained encoder can be harnessed to reach comparable in-domain performance. However, to generalize to unseen target articles, domain specific pre-training is beneficial. It should be noted that LegalBERT (Chalkidis et al., 2020) has been pre-trained on a collection of ECtHR decisions that may include cases from LexGLUE's test partition, thereby possibly injecting domain-specific information about the target articles into the encoding.

5.5 How does Article Relatedness Affect Zero-Shot Transferability?

To test whether article relatedness between source and target domains affect performance, we experiment with Article P1-1 (Article 1 of Additional Protocol 1 - The Protection of Property) as the target domain. This simulates the realistic scenario of our zero shot setting where the convention is amended with an additional protocol. We then constructed one related and one unrelated source domain based on the suggestion provided by a legal expert (the second author) while ensuring training sets of similar size. The related domain consists of articles 6 (right to a fair trial) and 8 (right to respect for private and family life). The unrelated domain articles comprise articles 2 (right to life), 3 (prohibition of torture, and 5 (right to liberty and security).

We report the performance on Task A for target P1-1 in Table 5. We observe that the related source domain is able to perform better across the board, confirming the intuition that relatedness between source and target is an important factor to be considered when training a model for transferability. As before, we observe that UDA achieves higher performance overall as it has the chance to see article P1-1 during training. Interestingly, we observe the Wasserstein method outperforming the Domain Discriminator for the related source, but vice versa for the unrelated source. We believe this is owed to related articles forming similar feature distributions and thereby making it easy for the Wasserstein distance to facilitate adaptation. This case study suggests the design of domain adaptation components which derive information more from related articles than unrelated ones when transferring to a target article. This raises a related question of how article relatedness could be determined by the model itself rather than a priori by an expert.

6 Conclusion

We cast case outcome classification on ECtHR data into an article-aware architecture. This configuration is inspired by realistic legal reasoning involving both the case facts and convention articles to determine possible allegations/violations. Assuming non-finetuned pre-trained encoders, we observe a performance improvement over a simple fact-only classification model. It also enables us to conduct experiments in zero shot transfer COC with and without access to unlabeled target data during domain adaptation. While we show that domain adaptation techniques are in principle suitable to facilitate generalization, the divergence between source and target domain performance is large and this task variant is very difficult. We further observe that the effectiveness of domain adaptation interacts with law-specific pre-training of transformer-based encoders and with the relatedness of the source and target domains. Overall, this zero-shot COC task formulation opens up new research opportunities towards legal NLP models that are more aligned with expert reasoning.

Limitations

We cast the legal judgment prediction task into an article-aware classification setting and create a

Table 3: Task A F1 performance of baseline and domain adaptation models

Setting	Model	Transfer $0 \rightarrow 1$				Transfer $0 \leftarrow 1$			
		source : $split\_0$		target : $split\_1$		source : $split\_1$		target : $split\_0$	
		mac.	mic.	mac.	mic.	mac.	mic.	mac.	mic.
Baseline	Source only	63.62	71.98	3.14	3.78	67.79	74.57	5.80	8.02
UDA	Domain Disc.	**64.65**	**72.52**	9.52	9.87	**68.19**	**75.32**	14.47	16.51
	Wasserstein	60.26	71.46	**11.04**	**18.20**	63.56	74.89	**15.23**	**26.20**
ADA	Domain Disc.	**64.89**	72.08	7.18	**7.78**	67.12	74.43	6.45	9.34
	Wasserstein	61.78	**72.36**	**7.27**	7.61	65.71	**74.88**	**6.71**	**9.71**

Table 4: Task A F1 Performance in one split using BERT base embeddings (as opposed to Legal Bert)

Setting	Model	source : $split\_1$		target : $split\_0$	
		mac.	mic.	mac.	mic.
UDA	Dom. Disc.	68.01	75.26	13.68	15.21
	Wasserstein	62.15	74.32	14.12	16.36
ADA	Dom. Disc.	67.92	75.32	4.77	7.44
	Wasserstein	66.71	74.95	4.73	7.65

Table 5: Task A F1 target performance on article P1-1 with related and unrelated source domains

Setting	Source Model	Related		Unrelated	
		mac.	mic.	mac.	mic.
UDA	Dom. Disc.	54.13	73.71	43.52	65.72
	Wasserstein	62.35	74.64	34.01	49.62
ADA	Dom. Disc.	42.79	68.46	37.12	56.16
	Wasserstein	43.25	69.91	26.87	38.28

zero-shot benchmark on a corpus of ECtHR cases. Matching between the text of legal sources and case fact descriptions varies greatly between different legal systems and subdomains, and is highly dependent on the textual nature of the case fact and legal sources. Specific to our context, for example, we have discussed the ECtHR fact statements as being influenced by the eventual case outcome and not suitable for prospective prediction in sec 1. COC as article-aware classification in other jurisdictions will likely lead to different levels of task difficulty, absolute performance, and zero shot transferability. In particular, many legal areas require multiple sources to be applied in conjunction to a set of case facts.

Technically, a major hurdle dealing with corpora related to the legal domain is their lengthy nature. We resort to hierarchical models, which are inherently limited in that tokens across long distances cannot directly attend to one another. This restriction of hierarchical models is still underexplored (but see preliminary work in, e.g. Dai et al. 2022; Chalkidis et al. 2022a). Additionally, we freeze the weights in the LegalBERT sentence encoder, both to save computational resources and to reduce the model's susceptibility to shallow surface signals

and ensure the comparability of our domain adaptation methods, in particular with respect to the impact of domain-specific pre-training. We leave an exploration of COC as article-aware classification with fine-tuned encoders for future work.

Ethics Statement

We experiment with a publicly available datasets of ECtHR decisions, which has been derived from the public court database HUDOC[3]. These decisions contain real names of the parties involved without any anonymization. We hence do not consider our experiments to produce any additional harmful effects relating to personal information.

The task of legal judgment prediction raises ethical, civil rights, and legal policy concerns, both general and specific to the European Court of Human Rights (e.g., (Fikfak, 2021) on system bias and court caseload). The main premise of this work is to make incremental technical progress towards enabling systems to work with case outcome information in a way that is aligned with how human experts analyze case facts through an interplay with complex legal sources. We do not advocate for the practical application of COC/LJP systems by courts, but rather explore how their core functionality of processing legal text can be made as expert-aligned as possible. Our research group is strongly committed to research on such models as a means to derive insight from legal data for purposes of increasing transparency, accountability, and explainability of data-driven systems in the legal domain.

We are conscious that, by adapting pre-trained encoders, our models inherit any biases they contain. Similarly, the ECtHR case collection as historical data may contain a data distribution in which sensitive attributes of individuals (e.g., applicant gender) may have some predictive signal for the allegation/violation variable (see, e.g., (Chalkidis et al., 2022c)). We believe the results we observe

---

[3]https://hudoc.echr.coe.int

in our COC experiments to not be substantially related to such encoded bias. However, legal NLP systems leveraging case outcome information and intended for practical deployment should naturally be scrutinized against applicable equal treatment imperatives regarding their performance, behavior, and intended use.

All models of this project were developed and trained on Google Colab. We did not track computation hours.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2011. Relation adaptation: learning to extract novel relations with minimum supervision. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022a. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos.

2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022b. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022c. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406.

Junyi Chen, Lan Du, Ming Liu, and Xiabing Zhou. 2022. Mulan: A multiple residual article-wise attention network for legal judgment prediction. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–15.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126.

Veronika Fikfak. 2021. What future for human rights? decision-making by algorithm. *Decision-making by algorithm (September 3, 2021). Strasbourg Observers*, 19.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Nils Holzenberger, Andrew Blair-stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *NLLP@KDD*.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization 3rd int. In *International Conference on Learning Representations*.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting brazilian court decisions. *PeerJ Computer Science*, 8:e904.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2237–2243.

Zhenyu Liu and Huanhuan Chen. 2017. A predictive performance comparison of machine learning models for judicial cases. In *2017 IEEE Symposium series on computational intelligence (SSCI)*, pages 1–6. IEEE.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. 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 4046–4062.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ICAIL*.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.

Aakanksha Naik and Carolyn Rose. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855.

Robert Remus. 2012. Domain adaptation using domain similarity-and domain complexity-based instance selection for cross-domain sentiment analysis. In *2012 IEEE 12th international conference on data mining workshops*, pages 717–723. IEEE.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention.

Guy Rotman and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054.

T. Y. S. S Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023. Leveraging task dependency and contrastive learning for case

outcome classification on european court of human rights cases. *arXiv preprint arXiv:2302.00768*.

T.y.s.s Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for European court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 71–79.

JURI SAYS. 2020. Prediction system for the european court of human rights. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334, page 277. IOS Press.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063.

Rafe Athar Shaikh, Tirath Prasad Sahu, and Veena Anand. 2020. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:2393–2402.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 2020 the 3rd international conference on information science and system*, pages 204–209.

Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the law area and decisions of french supreme court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Avinante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd annual computer software and applications conference (COMPSAC)*, volume 2, pages 130–135. IEEE.

Tomer Volk, Eyal Ben-David, Ohad Amosy, Gal Chechik, and Roi Reichart. 2022. Example-based hypernetworks for out-of-distribution generalization. *arXiv preprint arXiv:2203.14276*.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in germany's tax law. In *International conference on electronic participation*, pages 89–99. Springer.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019a. Adversarial domain adaptation for machine reading comprehension. 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 2510–2520.

Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019b. Hierarchical matching network for crime classification. In *proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 325–334.

Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling dynamic pairwise attention for crime classification over legal articles. In *the 41st international ACM SIGIR conference on research & development in information retrieval*, pages 485–494.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4085–4091.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised domain adaptation for neural machine translation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 338–343. IEEE.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923.

Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 682–690.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: a circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

Yftah Ziser and Roi Reichart. 2017. Neural structural correspondence learning for domain adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410.

A Implementation Details

We employ a maximum sentence length of 256 and document length (number of sentences) of 50. Our word level attention context vector size is 300. The sentence level GRU encoder dimension is 200 (i.e. 400 bidirectional), and the sentence level attention vector dimension is 200. The entailment classifier hidden layer also has size 200. Domain discriminator and critic have two layered networks with hidden layers of size 200 and 100. The entailment classifier is trained with a binary cross entropy loss while the domain discriminator is trained with cross entropy loss over a one-hot domain vector. The model is optimized end-to-end using Adam (Kingma and Ba, 2015). The dropout rate (Srivastava et al., 2014) in all layers is 0.1. To handle data skewness in the entailment setup, we employ a custom batch sampler which ensures every batch contains 4 different articles as well as 2 positive and 2 negative instances per article. Our batch size is 16. We employ a learning rate scheduler based on loss plateau decay. For adversarial training using GRL, following (Ganin and Lempitsky, 2015), we set the λ in gradient reversal to be $\lambda = \frac{2}{1+\exp(-\gamma p)} - 1$ where $p = \frac{t}{T}$, where t and T denote current training step and total training steps. γ is determined using a grid search over [0.05, 0.1, 0.15, 0.2]. We employ a 10 class domain discriminator (5 from source and 5 from target) in the case of UDA and a 5 class discriminator in the case of ADA. We reduce the mean between instances of a particular article of source and target in the case of UDA. In the case of ADA, we reduce the mean between instances of different articles in the source domain.