Beyond Borders: Investigating Cross-Jurisdiction Transfer in Legal Case Summarization

Santosh T.Y.S.S[1],**Vatsal Venkatkrishna**[2], **Saptarshi Ghosh**[2], **Matthias Grabmair**[1]
[1]Technical University of Munich, Germany
[2]Indian Institute of Technology, Kharagpur, India
{santosh.tokala, matthias.grabmair}@tum.de
vatsalvenkatkrishna@gmail.com, saptarshi@cse.iitkgp.ac.in

Abstract

Legal professionals face the challenge of managing an overwhelming volume of lengthy judgments, making automated legal case summarization crucial. However, prior approaches mainly focused on training and evaluating these models within the same jurisdiction. In this study, we explore the cross-jurisdictional generalizability of legal case summarization models. Specifically, we explore how to effectively summarize legal cases of a target jurisdiction where reference summaries are *not* available. In particular, we investigate whether supplementing models with unlabeled target jurisdiction corpus and extractive silver summaries obtained from unsupervised algorithms on target data enhances transfer performance. Our comprehensive study on three datasets from different jurisdictions highlights the role of pre-training in improving transfer performance. We shed light on the pivotal influence of jurisdictional similarity in selecting optimal source datasets for effective transfer. Furthermore, our findings underscore that incorporating unlabeled target data yields improvements in general pre-trained models, with additional gains when silver summaries are introduced. This augmentation is especially valuable when dealing with extractive datasets and scenarios featuring limited alignment between source and target jurisdictions. Our study provides key insights for developing adaptable legal case summarization systems, transcending jurisdictional boundaries.

1 Introdution

Legal professionals, including lawyers, judges, and paralegals, are often inundated with an overwhelming amount of case judgements containing intricate textual content encompassing multiple issues and arguments, making it time-consuming to read and comprehend every document in detail manually. To address this, many legal information systems offer case summaries authored by legal experts. There has been a lot of academic research over the years to automatically produce the summaries of legal case judgements, which can help reduce the human cost effectively (Bhattacharya et al., 2021; Shukla et al., 2022; Agarwal et al., 2022).

Traditional legal case judgement summarization methods have primarily focused on extractive summarization, where crucial sentences are selected from source documents. Initially, unsupervised approaches were prominent, tailored to capture legal discourse nuances without requiring training data. More recently, supervised extractive summarizers have emerged, relying on expert-annotated summaries for training. Recognizing the limitations of extractive methods in providing comprehensive context and coherence, researchers have shifted towards abstractive summarization, leveraging recent advances in unsupervised pre-training (e.g., Devlin et al. 2019; Lewis et al. 2020), legal domain-specific continued pre-training (e.g., Paul et al. 2022; Chalkidis et al. 2020; Gururangan et al. 2020), and summarization task-specific pre-training (e.g., Zhang et al. 2020).

The task of summarizing legal case judgements presents distinctive challenges due to several factors, such as the legal domain vocabulary being slightly different in different jurisdictions (e.g., country / court), as well as the complex sentence and discourse structures influenced by various writing styles that are in use in different jurisdictions. The current strategy to design a legal case summarization system for a *new* jurisdiction (which we call *target jurisdiction*) either involves unsupervised methods or collecting expert annotated datasets to fine-tune supervised summarization models. Though these supervised models usually show improved performance compared to unsupervised methods, this often demands a large number of expert summaries, making it costly and less adaptable to new jurisdictions. This prompts us to ask *how we can build an effective summarization system for a target jurisdiction without data*

annotation efforts. In this regard, we seek to investigate whether models trained on some other jurisdiction for which training data exists (which we call as *source jurisdiction*) can generate better summaries than unsupervised methods for the target jurisdiction. Thus, we evaluate the cross-jurisdiction generalizability of different legal case summarization systems and further propose an adversarial training-based setup to improve the cross-jurisdiction transfer performance, paving the way for the development of summarization systems that are generalizable in real-world legal scenarios. We summarize our three primary research questions:

RQ1: In situations where reference summaries are unavailable to train supervised models for a specific target jurisdiction, can supervised summarization models, trained using data from a different jurisdiction (source jurisdiction) outperform unsupervised methods in generating more effective summaries? What criteria should guide the selection of the most suitable source jurisdiction for a specific target jurisdiction? Which models exhibit better cross-jurisdiction summarization capabilities?

RQ2: Can we leverage unlabeled judgment data (only documents, not reference summaries) from the target jurisdiction to improve transfer performance during training supervised summarization models on source jurisdiction? This approach is akin to unsupervised domain adaptation, where unlabelled target inputs are used when fine-tuning on a labelled source corpus. We employ an adversarial domain discriminator method (Ganin et al., 2016) to learn the jurisdiction-agnostic feature representations to facilitate cross-jurisdictional transfer.

RQ3: Can extractive *silver summaries* on the target jurisdiction data, *obtained using unsupervised summarization algorithms without manual human annotation*, further improve transfer performance in addition to unlabelled target jurisdiction data? These silver summaries aid the decoder to grasp semantic nuances in target jurisdiction, complementing adversarial domain discriminator.

To answer these questions, we conducted a comprehensive experimental analysis on three datasets from different jurisdictions. We find that supervised models trained on a different source jurisdiction can yield better summarization in the target jurisdiction, compared to unsupervised models. Further, we observe that the choice of source dataset is highly influenced by jurisdictional similarity rather than the dataset size or abstractiveness (RQ1). We witness an improvement in transfer performance using domain discriminator-based adversarial training. However, this improvement is contingent on the choice of backbone model. For instance, we observe a decline in transfer performance when applied to Legal pre-trained Pegasus in contrast to improvement in general pre-trained BART model (RQ2). Finally, we notice an improvement when adding silver summaries in scenarios when the dataset is extractive and there is less similarity between source-target jurisdictions (RQ3).

2 Related Work

Legal Case Summarization: Traditional works in this area have primarily employed extractive methods known for their faithfulness to the input case document (Bhattacharya et al., 2019). These encompass both unsupervised methods (Bhattacharya et al., 2021; Polsley et al., 2016; Farzindar, 2004; Saravanan et al., 2006; Mandal et al., 2021) and supervised approaches (Agarwal et al., 2022; Liu and Chen, 2019; Zhong et al., 2019; Xu et al., 2021; Mandal et al., 2021), employing diverse strategies such as knowledge engineering of domain-specific features, redundancy handling using MMR, joint multi-task learning with Rhetorical Role Labeling, leveraging argument structures, integrating document-specific catchphrases.

While extractive approaches excel in faithfully representing source document content, they face inherent challenges, including incomplete or incorrect discourse, coreference resolution issues, and a lack of context, impacting the readability of generated summaries (Zhang et al., 2022). Hence, there is a growing interest in exploring abstractive summarization methods for legal case summarization (Shukla et al., 2022; Ray et al., 2020; Schraagen et al., 2022). Shukla et al. 2022 applied transformer-based pre-trained abstractive models like BART, Legal-LED, and Legal-Pegasus to legal case summarization. To address long legal documents, Moro and Ragazzi 2022 proposed chunking input documents into semantically coherent segments, enabling these pre-trained transformer-based models to summarize lengthy documents without truncation. Recent efforts also aim to enhance factual accuracy in generated summaries by employing entailment modules to select faithful candidates (Feijo and Moreira, 2023). Additionally, there are efforts to understand the argumentative structure in the legal documents to improve the performance of

legal case summarization models (Xu and Ashley, 2022; Elaraby and Litman, 2022; Elaraby et al., 2023). In contrast to the approaches mentioned, which train and evaluate summarization models within specific jurisdictions, our work focuses on evaluating the cross-jurisdiction generalizability of legal case summarization systems.

Cross Domain/Dataset Generalization: This assessment in NLP has been undertaken across various tasks, including constituent parsing (Fried et al., 2019), reading comprehension (Talmor and Berant, 2019), named entity recognition (Fu et al., 2020), summarization (Chen et al., 2020), sentiment analysis (Yu and Jiang, 2016) and relation extraction (Bassignana and Plank, 2022), among others. Recently, Thakur et al. (Thakur et al., 2021) created a robust benchmark called BEIR (Benchmarking IR), for evaluation of retrieval model generalization capabilities across various domains.

In the legal domain, Kumar et al. (Kumar et al., 2022) explored the cross-domain transferability of text generation models across four legal domains, including congressional bills, treaties, legal contracts, and privacy policies. However, the tasks they investigated within each domain exhibited significant variations in terms of generated text, such as titles in the privacy policy domain and actual summaries in congressional bills. In contrast, our work specifically focuses solely on summarizing legal case judgments. We consider each jurisdiction as a distinct domain characterized by its unique vocabulary and discourse writing style. Furthermore, we explore strategies to enhance cross-jurisdictional transfer performance through adversarial learning-based domain adaptation.

Domain Adaptation: Domain adaptation mitigates covariate shifts between source and target data distributions (Ruder, 2019). Unsupervised domain adaptation (UDA) leverages labeled source data and unlabeled target data, extensively studied in NLP and computer vision (Daumé III, 2009; Sun et al., 2016; Wang et al., 2019; Shah et al., 2018; Ganin et al., 2016; Shen et al., 2018). Recently, DA has extended to information retrieval (Xin et al., 2022; Wang et al., 2022; Yu et al., 2022), aided by benchmarks like BEIR. Domain adaptation often employs domain discriminator-based adversarial training (Ganin et al., 2016) to align source and target distributions within an encoder, reducing distribution disparities. In the legal domain, this technique has been applied to case outcome classification, treating legal articles as separate domains to assess transfer performance across violation predictions for different articles (Tyss et al., 2023).

We employ DA to create jurisdiction-invariant representations in legal case summarization. However, relying solely on a jurisdiction-agnostic encoder representation, as commonly used in classification tasks, is insufficient as in the context of summarization, the decoder must also capture semantic nuances to generate domain-specific summaries. Therefore, we explore the use of silver summaries derived from unsupervised methods for target domain data.

3 Datasets

There are very few publicly available English legal case summary corpora available. We use the following three datasets from prior works.

(i) UK-Abstractive dataset (abbreviated as 'UK-Abs') (Shukla et al., 2022): This dataset includes 793 case documents and their summaries from the UK Supreme Court, spanning from 2009 onwards. The abstractive summaries are sourced from the official press-released version available on the court's website. We use the provided split, with 693 documents for training and 100 for testing.

(ii) Indian-Abstractive dataset (IN-Abs) (Shukla et al., 2022): This dataset comprises case-summary pairs from Indian Supreme Court judgements and is obtained from the website of the Legal Information Institute of India. These abstractive summaries, also known as headnotes, are available for 7130 cases, out of which 7030 are used for training and the remaining 100 for testing.

(iii) BVA-Extractive (BVA-Ext) (Zhong et al., 2019): It comprises case-summary pairs focusing on single-issue Post-Traumatic Stress Disorder decisions from the US Board of Veterans' Appeals (BVA). It encompasses a total of 112 decisions, each accompanied by an expert-annotated gold-standard extractive summary. Within this dataset, 92 cases within the training set are associated with one annotated extractive summary each. Additionally, the test set comprises 20 cases, for which four distinct extractive summaries are provided by multiple annotators. In all our experiments, for each document in the test set, we report the averaged metric score across these four reference summaries.

Dataset characteristics: To gain deeper insights into model performance, we measure the following dataset characteristics. **Compression Ra-**

tio (Grusky et al., 2018) indicates the word ratio between the summary and the case document. **Coverage** (Grusky et al., 2018) indicates the percentage of words in the summary that are part of an extractive fragment with the input, and quantifies the extent to which a summary is derivative of an input text. **Density** (Grusky et al., 2018) quantifies how well the word sequence of a summary can be described as a series of extractions. It is derived from the average length of the extractive fragment to which each word in the summary belongs. **Copy Length** (Chen et al., 2020) denotes the average length of segments in summary copied from the source document. **Repetition** (See et al., 2017) indicates the proportion of n-grams repeated in the summary itself. **Novelty** (See et al., 2017) denotes the proportion of n-grams present in the summary that are *not* in the source. High values of coverage, density, and copy length suggest a more extractive dataset, while novelty reflects the level of abstractiveness in the reference summaries. The lower the compression ratio, the more precise capture of critical aspects from the article text is required, which presents a greater challenge in summarization.

Table 1 reports the characteristics of the three datasets. Barring the completely extractive BVA-Ext, UK-Abs tend to have slightly higher coverage than IN-Abs. By contrast, IN-Abs has more copy length and density, suggesting a higher presence of verbatim copy segments when compared to UK-Abs. Conversely, UK-Abs is characterized by a higher 3-gram novelty score, which, coupled with its low compression ratio, indicates its challenging nature. Additionally, IN-Abs tends to include more 2-gram repeated phrases within the summaries.

Similarity between the datasets from different jurisdictions: We seek to quantify the similarity between jurisdictions on both lexical and semantic levels. On the lexical level, we calculate the **vocabulary overlap** (%) between the jurisdictions from both the documents and summaries. The vocabulary for each domain is created by considering the top 1K most frequent words (excluding stopwords) (Gururangan et al., 2020) from the documents and the summaries. At the semantic level, we employ a pre-trained GPT-2 model (Radford et al., 2019) to compute **perplexity scores** using a strided sliding window approach with a stride of 512. We gauge the similarity between jurisdictions based on the order of perplexity scores among the jurisdictions, which provides insight into their semantic resemblance.

From Table 2 and Table 3, it becomes evident that IN-Abs holds more resemblance to UK-Abs compared to BVA-Ext. This observation can be attributed to two key factors. Firstly, India's historical ties to the UK have led to a substantial influence on its legal system, stemming from British colonial rule. Secondly, both UK-Abs and IN-Abs datasets originate from Supreme Courts dealing with a broad spectrum of legal matters due to their comprehensive jurisdiction. In contrast, the BVA corpus, focusing specifically on single-issue PTSD decisions, might exhibit more narrowly focused language. With respect to BVA-Ext, UK-Abs tend to be more similar than IN-Abs on both similarities.

4 Models & Evaluation Metrics

We use the following representative models from different legal case summarization methods in Shukla et al. 2022 for our study. We briefly describe each method and refer the reader to the aforementioned work for implementation details.

Unsupervised Extractive Methods: (i) Domain-agnostic methods which employ graph-based salient sentence identification, namely LexRank (Erkan and Radev, 2004) and Reduction (Mihalcea and Tarau, 2004), matrix factorization-based LSA (Steinberger et al., 2004) and TF-IDF based Luhn (Luhn, 1958). (ii) Legal Domain-specific methods include LetSum (Farzindar, 2004), CaseSummarizer (Polsley et al., 2016), which ranks sentences based on their TF-IDF weights coupled with legal-specific features, Maximal Marginal Relevance (MMR) (Zhong et al., 2019) uses iterative selection of predictive sentences and finally uses MMR to select the final summary sentences.

Supervised Extractive Methods: SummaRuNNer (Nallapati et al., 2017) uses binary classification for sentence selection to determine whether it should be part of the summary. To build the training data for the classifier from a source document d and its abstractive reference summary s, each sentence in s is matched with the top sentences from d, chosen based on the highest average ROUGE-1, 2 and L scores. The extractive pseudo-reference summary is created by combining all selected sentences and are then used for supervised training of the model.

Abstractive Methods: We use BART (Lewis et al., 2020) pre-trained on general English cor-

Table 1: Statistics of the case summarization datasets. Sum.:Summary; Doc.:Document

	UK-Abs	IN-Abs	BVA-Ext
Train size	693	7030	92
Test Size	100	100	20
Compression Ratio	0.129	0.224	0.152
Avg. #Tokens in Sum.	1268	1067	259
Avg. #Tokens in Doc.	14599	5408	2548
Copy Length	2.958	3.725	39.86
Coverage	0.968	0.943	1.00
Density	8.129	10.314	53.91
Novelty (3-gram)	0.44	0.375	0.00
Repetition (2-gram)	0.019	0.021	0.013

pora, and the 'Legal Pegasus' model[1] that is further trained on documents from litigations of U.S. Securities and Exchange Commission (SEC) after its pre-training phase. We assess their performance with and without fine-tuning. To accommodate the 1024-token input limit, we chunk the input document, generate summaries for each chunk, and merge them, similar to what was done in Shukla et al. 2022. To create reference summaries for each chunk for fine-tuning, we map each summary sentence to its most similar counterpart in the document using cosine similarity between mean token-level Sentence-BERT (Reimers and Gurevych, 2019) embeddings and concatenating them as references for each document chunk. We also use the Legal-Longformer Encoder-Decoder (Legal-LED) model[2] which is trained on SEC legal corpus and is tailored specifically for long documents with 16,384 tokens through sparse attention mechanism (Beltagy et al., 2020).

Hybrid Extractive-Abstractive Methods: We also use both vanilla and fine-tuned models of BERT-BART (Bajaj et al., 2021), wherein the document length is reduced by selecting salient sentences using a BERT-based extractive summarization model followed by a BART model to generate the final summary.

Evaluation metrics: For evaluating the quality of generated and extracted summaries, we employ ROUGE-L F-score (R-L) (Lin, 2004) and BERTScore (BS) (Zhang et al., 2019) and finally report the average scores across all the test instances of the target jurisdiction.

[1] https://huggingface.co/nsi319/legal-pegasus

[2] https://huggingface.co/nsi319/legal-led-base-16384

Table 2: Perplexity score of the datasets using GPT2

Dataset	Perplexity
UK-Abs	16.91
IN-Abs	17.81
BVA-Ext	14.74

Table 3: Vocabulary overlap (%) between datasets.

	UK-Abs	**IN-Abs**	**BVA-Ext**
UK-Abs	100	49.7	30.7
IN-Abs	49.7	100	25.9
BVA-Ext	30.7	25.9	100

5 RQ1: Cross-Jurisdiction Generalizability

Unlike previous studies (Bhattacharya et al., 2019; Shukla et al., 2022) that primarily evaluated these summarization algorithms in in-domain settings, where fine-tuning and evaluation occur on the same dataset, we focus on assessing their generalization capabilities across domains/jurisdictions. Unsupervised models can be applied directly to summarize documents of any dataset (jurisdiction). For supervised models, we train the model on a specific dataset (source jurisdiction) and evaluate its performance on other datasets (target jurisdictions) in a blind zero-shot manner.

Results (Table 4): Among unsupervised methods (Block 1 in Tab.4), domain-specific algorithms CaseSummarizer and MMR consistently performed the best, especially in terms of Rouge-L score. Supervised extractive model SummaRuNNer did not consistently yield better cross-jurisdiction performance compared to unsupervised models, owing to its lack of pre-training except in case of BVA-Ext target using UK-Abs as source (Block 2 in Tab. 4).

Abstractive & Hybrid models, due to their pre-training, demonstrated better performance compared to unsupervised methods even without any fine-tuning, as reflected over UK-Abs and IN-Abs primarily due to their abstractive nature of datasets and pre-training objective (Block 3 in Tab.4). LED is semi-pretrained as it has been initialized by repeatedly copying BART without undergoing end-to-end general pre-training and then it has been fine-tuned on SEC legal corpus to create Legal-LED, making it overfit to that specific domain, as reflected in its lower zero-shot performance.

Finally, abstractive & hybrid models fine-tuned on datasets from source jurisdictions generally per-

Table 4: ROUGE-L F scores (R-L) and BERT Score (BS) of all methods across three datasets. Each test dataset is represented by a column, while training dataset used for supervision is indicated in Train column, if applicable. Best results in each group and overall are bolded in black and magenta respectively.

	Test →	**UK-Abs**	**IN-Abs**	**BVA-Ext**
Model	**Train ↓**	**R-L/BS**	**R-L/BS**	**R-L/BS**
Unsupervised, Extractive Models				
LexRank	-	0.427/**0.836**	0.439/**0.850**	0.356/0.862
LSA	-	0.361/0.823	0.410/0.841	0.308/0.854
Luhn	-	0.435/0.828	0.414/0.843	0.353/0.860
Reduction	-	0.424/0.829	0.423/0.847	0.364/**0.865**
CaseSummarizer	-	0.443/0.835	0.476/0.836	**0.386**/0.829
LetSum	-	0.361/0.772	0.362/0.797	0.298/0.807
MMR	-	**0.467**/0.834	**0.485**/0.847	0.384/0.825
Supervised, Extractive Models				
SummaRuNNer	UK-Abs	-	**0.410/0.832**	**0.405/0.861**
	IN-Abs	**0.430/0.834**	-	0.327/0.846
	BVA-Ext	0.378/0.782	0.374/0.839	-
Abstractive & Hybrid Models without fine-tuning				
BART	-	**0.517**/0.833	0.486/0.839	0.324/0.841
Legal-Pegasus	-	**0.517**/0.831	**0.522/0.841**	**0.386**/0.830
Legal-LED	-	0.240/0.821	0.292/0.812	0.232/0.832
BERT-BART	-	0.501/**0.836**	0.482/0.838	0.329/**0.844**
Abstractive & Hybrid Models with fine-tuning				
BART	UK-Abs	-	**0.522/0.845**	**0.341/0.842**
	IN-Abs	**0.532/0.837**	-	0.306/0.834
	BVA-Ext	0.519/0.835	0.503/0.840	-
Legal-Pegasus	UK-Abs	-	**0.532/ 0.847**	**0.451/0.861**
	IN-Abs	**0.533/ 0.838**	-	0.434/0.854
	BVA-Ext	0.522/0.834	0.522/0.842	-
Legal-LED	UK-Abs	-	**0.418/0.838**	0.321/**0.834**
	IN-Abs	**0.479/0.832**	-	**0.339**/0.831
	BVA-Ext	0.233/0.817	0.289/0.812	-
BERT-BART	UK-Abs	-	**0.513/0.846**	**0.342/0.842**
	IN-Abs	**0.515/0.837**	-	0.308/0.836
	BVA-Ext	0.508/0.836	0.501/0.839	-

formed well compared to the unsupervised methods, emphasizing the importance of legal adaptation during fine-tuning in addition to their pre-training either generally as with BERT/BART-BERT or legal-specific as with Legal-Pegasus (Block 4 in Table 4). They all displayed greater generalization stability across different source datasets irrespective of the relation between the source and the target. Overall, Legal-Pegasus stood out as the most robust model across all the datasets attributed to its legal pre-training.

Choice of Source Jurisdiction: For both UK-Abs and IN-Abs, training on each other's data yielded better results compared to training on BVA-Ext (e.g., BART applied on UK-Abs test set, achieves 0.532/0.837 when trained on IN-Abs, compared to 0.519/0.835 when trained on BVA-Ext) due to their larger size, abstractive nature, and jurisdictional similarity between In-Abs and UK-Abs, as stated earlier in Section 3. Interestingly, for BVA-Ext, UK-Abs proved to be a better source domain than IN-Abs (for most supervised methods), despite the latter's larger size and more extractive nature. This transferability is mainly attributed to the similarity between jurisdictions, as observed in Section 3. This result shows that, while choosing the source jurisdiction (whose data will be used to train supervised summarization models), blindly choosing the jurisdiction with the largest dataset size is not an optimal approach. Rather, the source jurisdiction

should be chosen based on its lexical and semantic similarity with the target jurisdiction. To this end, the metrics we used in Section 3 can be useful to quantify the similarity between jurisdictions.

Main Takeaways: Our findings reveal that supervised models trained on a different source jurisdiction can yield better summarization in the target jurisdiction, compared to unsupervised models. Specifically, pre-trained models exhibit superior cross-jurisdictional generalizability due to their pre-training. Legal-oriented training further enhances their generalization capabilities. The choice of the source dataset significantly influences the development of an effective system for a specific target jurisdiction, with jurisdictional similarity playing a critical role than dataset size and abstractiveness.

6 RQ2: Leveraging Unlabelled Target Jurisdiction Corpus

In this section, we explore whether we can improve the cross-domain performance of supervised abstractive summarization models using the *unlabeled* judgement documents from the target jurisdiction, while training the models on source jurisidiction judgement-summary pairs. To this end, we adopt *domain-adversarial training* from unsupervised domain adaptation literature (Ganin et al., 2016). Note that, we still do *not* use any reference summary from the target jurisdiction.

In this technique, we additionally introduce an *auxiliary jurisdiction classifier* that tries to discriminate the source jurisdiction embeddings from the target ones. The encoder of the abstractive model is updated not only to facilitate the decoder in producing summaries for source jurisdiction documents, but is also adversarially trained to confuse the jurisdiction classifier to learn jurisdiction-invariant feature representations. Put differently, we want our models to learn how to summarize the given input document with minimal encoding of the jurisdiction-specific information contained in the texts, facilitating model to generalize to documents from target jurisdictions.

In our adaptation scenario, we have access to labelled judgement-summary pairs of a source jurisdiction $\{x_s, y_s\}$ and unlabelled judgement corpus from target jurisdiction $\{x_t\}$. Let θ_e and θ_d be the encoder E and decoder D parameters of a summarization model. We introduce an auxiliary jurisdiction classifier J with parameters θ_j, implemented as a linear layer, which takes the document representation from the encoder and performs binary classification to identify the document's jurisdiction. We train the classifier in an adversarial fashion to maximize the encoder's ability to capture information required for the summarization task while minimizing its ability to predict the jurisdiction. Instead of the two-step adversarial min-max objective of GAN (Goodfellow et al., 2014), Ganin et al. 2016 proposed to jointly optimize all the components using a Gradient Reversal Layer (GRL). The GRL is inserted between the encoder and jurisdiction classifier, where it acts as the identity during the forward pass but, during the backward pass, scales the gradients flowing through by $-\lambda$, making the encoder receive the opposite gradients from the jurisdiction classifier. Here, hyperparameter λ is the *adaptation rate* controlling the influence of the jurisdiction classifier on the encoder during training. Thus, the overall objective function can be compactly represented as:

$$\begin{aligned} \arg \min_{\theta_e,\theta_d,\theta_j} & L_1(D(E(x_s)), y_s) + \\ & \lambda L_2(J(GRL(E(x_?))), y_j) \end{aligned} \tag{1}$$

where the first term represents the actual loss term for the summarization task, computed over the labelled article-summary of source jurisdiction $\{x_s, y_s\}$ and the second term represents the binary cross entropy loss computed for classification of jurisdiction for both source and target documents $x_?$ ($x_?$ can be x_s or x_t) and it belongs to, between the source or target jurisdiction (y_j).

Experimental Setup: We select BART, Legal-Pegasus, and BERT-BART for these experiments, given their robust cross-domain performance in Section 5. For each pair of source-target dataset, we apply this setup to all three abstractive models. We adopt the hyperparameters from Shukla et al. 2022. The jurisdiction classifier takes the mean of token-level embeddings obtained from the encoder as the document representation. The adaptation rate is given by $\lambda = \frac{2}{1+\exp(-\gamma p)} - 1$, where $p = \frac{t}{T}$ and t and T represent the current training steps and total steps respectively. We fine-tune γ within [0.05, 0.1] using the validation set.

Results (Table 5): On comparing the models trained with the GRL (RQ2) to their non-GRL counterparts (blind zero-shot transfer, as in RQ1), we notice an improvement with addition of GRL for both BART and BERT-BART across the three test sets, indicating that the models gained more robust and domain-invariant representations through

Table 5: ROUGE-L F-score (R-L) and BERT Score (BS) of three abstractive summarization models trained under different settings. Each target (test) dataset is represented by a column and and 'S' columns denote the source dataset for fine-tuning. An empty configuration indicates that the model is fine-tuned on the source dataset and evaluated on the target dataset (RQ1). 'GRL' indicates models fine-tuned in an adversarial fashion (RQ2). 'Silver' indicates models trained with silver summaries of the target dataset obtained through an unsupervised MMR algorithm (RQ3). Entries marked with * are statistically significantly higher than the baseline (the empty configuration) using 95% confidence interval by Wilcoxon signed rank test.

	Target →		**UK-Abs**		**IN-Abs**		**BVA-Ext**
Model Config		**S**	**R-L/BS**	**S**	**R-L/BS**	**S**	**R-L/BS**
BART		IN	0.532/0.837	UK	0.522/0.847	UK	0.341/0.842
	GRL		**0.536/0.838**		**0.526/0.848**		0.460/0.868 *
	Silver		0.529/0.838		0.507/0.832		**0.471/0.872** *
		BVA	0.519/0.835	BVA	0.503/0.840	IN	0.306/0.834
	GRL		0.544/0.837 *		0.513/0.842 *		0.440/0.864 *
	Silver		**0.546/0.839** *		**0.515/0.842** *		**0.471/0.871** *
BERT-BART		IN	0.515/0.838	UK	0.513/0.846	UK	0.342/0.842
	GRL		**0.520/0.838** *		**0.522/0.848***		0.461/0.868 *
	Silver		0.516/0.838		0.508/0.838		**0.465/0.868** *
		BVA	0.508/0.837	BVA	0.501/0.839	IN	0.308/0.836
	GRL		0.522/0.838 *		0.504/0.840		0.440/0.857 *
	Silver		**0.527/0.838** *		**0.509/0.843** *		**0.465/0.869** *
Legal-Pegasus		IN	0.533/0.836	UK	0.532/0.847	UK	0.451/0.861
	GRL		0.528/0.834		0.524/0.838		0.394/0.829
	Silver		**0.534/0.836**		**0.537/0.848** *		**0.469/0.868** *
		BVA	0.522/0.834	BVA	0.522/0.842	IN	0.434/0.854
	GRL		0.517/0.828		0.521/0.842		0.326/0.818
	Silver		**0.524/0.835**		**0.524/0.844**		**0.465/0.862** *

adversarial training, ultimately enhancing their transferability. For instance, for BART trained on IN-Abs and tested on UK-Abs, the R-L/BS scores for non-GRL variant are 0.532/0.837, and they improved to 0.536/0.838 for the GRL variant. This improvement is greatly pronounced when we consider BVA-Ext as the test set (e.g., 0.341/0.842 to 0.460/0.868 in the case of UK-Abs as the train dataset for BART). However, in the case of Legal-Pegasus, which already demonstrated strong cross-domain generalizability (RQ1) due to its pre-training on legal texts, we notice a decline in its performance with addition of GRL (eg., 0.533/0.836 to 0.528/0.834 in case of IN-Abs as train, UK-Abs as test). We attribute this to what we refer to as 'representation erasure'. It seems that the incorporation of the GRL and jurisdiction discriminator inadvertently leads to the dilution of generalizable knowledge that the model had acquired during its pre-training phase. This erasure effect is most pronounced in BVA-Ext as test dataset (e.g., 0.451/0.861 to 0.394/0.829 when UK-Abs is used as train dataset), which aligns with Legal-Pegasus pre-training data in terms of its jurisdiction (US) specific features compared to others.

Main Takeaway: Adversarial learning can enhance general pre-trained models (e.g., BART, BERT-BART) making them proficient at acquiring domain-invariant representations. But this strategy can also result in decreased performance for legal pre-trained models like Legal-Pegasus, attributable to representation erasure.

7 RQ3: Leveraging Silver Summaries of Target Jurisdiction

In this section, we explore whether incorporation of silver summaries from the target jurisdiction, obtained via *unsupervised* extractive summarization algorithms without the need for human annotations, can enhance effectiveness of cross-domain transfer. We employ the same domain-adversarial training framework, with a slight modification accounting for availability of judgment-summary pairs from both the source jurisdiction (expert-annotated) and the target jurisdiction (extractive silver summaries obtained via unsupervised algorithms). The first

term in overall objective, as described in Eqn. 1, is now used for target judgement-summary pairs as well. In other words, the x_s in first term of Eqn. 1 can now be $x_?$ (i.e., either x_s or x_t). This modification also strengthens the robustness of the adversarial learning approach. In the previous setup, the decoder's updates were solely influenced by source jurisdiction summaries, potentially leading to overfitting to source jurisdiction semantics and limiting its generalizability to new jurisdictions. The inclusion of silver summaries helps the decoder align with the semantics of the target jurisdiction, thereby facilitating more effective transfer performance.

Experimental Setup: Similar to RQ2, we replicate the setup using three abstractive models BART, Legal-Pegasus and BERT-BART for each dataset pair, designating one as the source jurisdiction with expert-annotated summaries and the other as the target jurisdiction with silver summaries. We obtain the silver summaries for the target jurisdiction using the MMR algorithm.[3]

Results (Table 5): With respect to BART and BERT-BART on UK-Abs (IN-Abs) as the target dataset, augmenting models with silver summaries yield improvements using BVA-Ext as train (eg., 0.544/0.837 to 0.546/0.839 on UK-Abs test, BVA-Ext train for BART), while models trained on IN-Abs (UK-Abs) did not exhibit enhancement. This can be attributed to the semantic similarity between UK-Abs and IN-Abs, allowing effective learning without additional silver summaries from each other. However, while using BVA-Ext as train dataset, the semantics specific to UK-Abs (IN-Abs) necessitate the inclusion of their silver summaries to facilitate the learning of target semantics.

With respect to Legal-Pegasus on both UK-Abs and IN-Abs as test set, we witness improvements in all training setups. For BVA-Ext as the test set, incorporation of silver summaries led to greater performance improvements across all three models and both source datasets consistently (e.g., 0.460/0.868 to 0.471/0.872 on UK-Abs train set on BART). This lends to the nature of BVA-Ext dataset being extractive getting boost from addition of silver summaries which are extractive.

Main Takeaways: Adding silver summaries significantly improved performance, particularly for extractive datasets like BVA-Ext and when dealing with less similar jurisdictions (e.g., using BVA-Ext for training and IN-Abs/UK-Abs for testing). This, along with our RQ2 findings, highlights the efficacy of silver summaries in mitigating representational erasure in legal pre-trained models, as seen in the improved performance of Legal-Pegasus.

[3]We also experimented with CaseSummarizer to obtain the silver summaries and observed that MMR showed better results than CS, as the trend in RQ1 Tab. 4.

8 Case Study

Tables 7 and 8 (in the Appendix) illustrate some errors found in the summaries generated by different configurations of BART models for a specific Indian Supreme Court judgement. One notable error involves Indian jurisdiction-specific terms such as I.P.C. (Indian Penal Code). Models with blind-zero transfer (RQ1) and those with access to the source text only in GRL (RQ2) often output terms like K.P.C. or K.C. (in place of I.P.C.), demonstrating a limitation in the semantics of the decoder to understand Indian-specific jargon. However, upon adding silver summaries (RQ3), the decoder becomes exposed to such jargon, leading to the correct output of I.P.C. Another common error across all models is incomplete sentences and coherence issues. In some cases, the model abruptly shifts to a new aspect without completing the previous one, potentially resulting in misrepresentations of the input document. These observations highlight the challenges in long text summarization, and the need for future improvement.

9 Conclusion

Our study reveals the following practical insights on building an effective legal summarization system for a target jurisdiction without annotated data: (i) Fine-tuning on non-target datasets outperforms unsupervised methods, but success depends on the similarity between source and target jurisdictions. (ii) When one has access to similar source data (e.g., IN-Abs for target UK-Abs), using general pre-trained models like BART and BERT-BART, combined with adversarial training, enhances transfer. (iii) When access to only non-similar source data (e.g., BVA-Ext for UK-Abs) is available, augmenting models with silver summaries from the target jurisdiction improves transfer. (iv) When employing adversarial learning with legal pre-trained models like Legal-Pegasus, it is important to be mindful of representational erasure, which can be mitigated by incorporating silver summaries.

Limitations

While our research contributes valuable insights into the cross-jurisdictional generalizability of legal case summarization systems, it is crucial to acknowledge certain limitations that affect the applicability of our findings. Our experiments are conducted on three specific legal datasets from diverse jurisdictions. The extent to which our observations hold true for a broader range of legal systems remains an open question. The legal domain is vast and varied, and different jurisdictions may exhibit unique characteristics that impact the generalizability of summarization models. The scarcity of publicly available legal summarization datasets limits the diversity and size of our training and evaluation data. Hence we are constrained with use of these three, only publicly available legal case summarization datasets in English. This constraint could impact the robustness and generalizability of our proposed methods and insights.

Our evaluation primarily relies on established summarization metrics such as ROUGE and BERTScore. While these metrics have been used in many prior works on legal document summarization, and are known to provide a quantitative measure of summarization quality, they may not fully capture the nuanced legal content, context, and intricacies essential for legal professionals. A potential avenue for further research could be developing additional legal domain-specific evaluation metrics. Another significant limitation of our study is the absence of direct participation or validation by legal experts in the assessment of summarization outputs, which we could not perform due to lack of access to legal experts.

Ethics Statement

We use publicly available datasets from prior works, which are obtained from the web. Though the case documents are not anonymized, we do not foresee any harm beyond their availability. We acknowledge the potential presence of biases within the training data, which may inadvertently be reflected in the generated summaries. To deploy these models in a production system, one must thoroughly check for such biases by comprehensively evaluating summarization performance across relevant groups (e.g., gender and race). Legal systems can even carry historical biases, and training models on biased datasets may perpetuate or exacerbate existing inequalities. As such, we urge vigilance in scrutinizing dataset biases and committing to fair and unbiased representation.

An inherent challenge in the development of automatic summarization models for legal decisions lies in potential performance variations across different partitions with in the same legal domain. For instance, in contexts like the Board of Veterans' Appeals (BVA) as discussed in Agarwal et al. 2022, cases involving rarely occurring disabilities or specialized legal and military situations may lead to suboptimal summaries due to sparsity in the training data. This variability could disproportionately impact groups that should be treated equally if their characteristics coincide with these less frequent legal configurations. Engaging domain experts to curate datasets with better representation across different types of injuries and legal phenomena can be a proactive step. This can enhance the model's understanding of uncommon or group-related legal contexts, potentially mitigating disparities in performance.

References

Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. 2022. Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1857–1872.

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, and Andrew McCallum. 2021. Long document summarization in a low resource setting using pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 71–80.

Elisa Bassignana and Barbara Plank. 2022. Crossre: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne,*

Germany, April 14–18, 2019, Proceedings, Part I 41, pages 413–428. Springer.

Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 22–31.

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.

Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Cdevalsumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3679–3691.

Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mohamed Elaraby and Diane Litman. 2022. Arglegalsumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. *arXiv preprint arXiv:2306.00672*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Atefeh Farzindar. 2004. Atefeh farzindar and guy lapalme,'letsum, an automatic legal text summarizing system in t. gordon (ed.), legal knowledge and information systems. jurix 2004: The seventeenth annual conference. amsterdam: Ios press, 2004, pp. 11-18. In *Legal knowledge and information systems: JURIX 2004, the seventeenth annual conference*, volume 120, page 11. IOS Press.

Diego de Vargas Feijo and Viviane P Moreira. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1):91–113.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330.

Jinlan Fu, Pengfei Liu, and Qi Zhang. 2020. Rethinking generalization of neural models: A named entity recognition case study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7732–7739.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vinayshekhar Bannihatti Kumar, Kasturi Bhattacharjee, and Rashmi Gangadharaiah. 2022. Towards cross-domain transferability of text generation models for legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 111–118.

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

Chin-Yew Lin. 2004. Looking for a few good metrics: Rouge and its evaluation.

Chao-Lin Liu and Kuan-Chun Chen. 2019. Extracting the gist of chinese judgments of the supreme court. In *proceedings of the seventeenth international conference on artificial intelligence and law*, pages 73–82.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

A. Mandal, Paheli Bhattacharya, Sekhar Mandal, and Saptarshi Ghosh. 2021. Improving legal case summarization using document-specific catchphrases. In *International Conference on Legal Knowledge and Information Systems*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11085–11093.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022. Pre-training transformers on indian legal text. *arXiv preprint arXiv:2209.06049*.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Oliver Ray, Amy Conroy, and Rozano Imansyah. 2020. Summarisation with majority opinion. In *JURIX*, pages 247–250.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

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

Murali Saravanan, Balaraman Ravindran, and Shivani Raman. 2006. Improving legal document summarization using graphical models. *Frontiers in Artificial Intelligence and Applications*, 152:51.

Marijn Schraagen, Floris Bex, Nick Van De Luijtgaarden, and Daniël Prijs. 2022. Abstractive summarization of dutch court verdicts using sequence-to-sequence models. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 76–87.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

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.

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.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 1048–1064.

Josef Steinberger, Karel Jezek, et al. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4(93-100):8.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Santosh Tyss, Oana Ichim, and Matthias Grabmair. 2023. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019. 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.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360.

Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2022. Zero-shot dense retrieval with momentum adversarial domain invariant representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4008–4020.

Huihui Xu and Kevin Ashley. 2022. Multi-granularity argument mining in legal texts. *arXiv preprint arXiv:2210.09472*.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 236–246.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. Coco-dr: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1479.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shiyue Zhang, David Wan, and Mohit Bansal. 2022. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *arXiv preprint arXiv:2209.03549*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D Ashley, and Matthias Grabmair. 2019. Automatic summarization of legal decisions using iterative masking of predictive sentences. In *Proceedings of the seventeenth international conference on artificial intelligence and law*, pages 163–172.

A Implementation Details

We follow the hyper parameters for all of the baseline models as outlined by Shukla et al. 2022. All the implementational details and code of the unsupervised baselines are provided in their github repository[4]. For BART, we employ a learning rate of 2e-5. We train the model end-to-end for 3 epochs with a batch size of 1. We limit maximum input and output length to 1024 and 512 respectively. For Legal Pegasus, we use a learning rate of 5e-5 and with batch size of 1 for 2 epochs, with maximum input and output length of 512 and 256 respectively. For Legal LED, we use a learning rate of 1e-3 trained for 3 epochs with a batch size of 4, with input and output length of 16384 and 1024 respectively. All models are trained with Adam optimizer (Kingma and Ba, 2014).

B Segment-wise Evaluation

Each summary in UK-Abs dataset is segmented into three segments – 'Background to the Appeal', 'Judgement', and 'Reasons for Judgement' –allowing segment-wise evaluation to assess the quality of summarization for each rhetorical segment. In line with Shukla et al. 2022, we present the ROUGE-L recall scores for each rhetorical segment, as available in the UK-Abs dataset (stated in Sec. 3). These scores are computed by comparing the generated summary to the corresponding golden summary under each segment.

Table 6 showcases the results for BART and BERT-BART models trained on IN-Abs and BVA-Ext source datasets, considering both non-GRL, GRL, and silver variants. Notably, when trained with BVA-Ext, the incorporation of silver summaries yields better scores across all segments than GRL and non-GRL variants. Conversely, for IN-Abs, the GRL variant exhibits strong performance across both models than silver summary variants. These segment-level trends also align with document-wise trends in Table 5, where performance of silver-enhanced models dropped when trained on IN-Abs but improved when trained on BVA-Ext (UK-Abs column of Tab. 5).

[4]https://github.com/Law-AI/summarization

Table 6: Segment-wise ROUGE-L Recall scores on the UK-Abs dataset. BGA, Jud., Rea. indicate 'Background to the Appeal', 'Judgement', and 'Reasons for Judgement' respectively.

Source →		**IN-Abs**			**BVA-Ext**		
Model Config		**BGA**	**Jud.**	**Rea.**	**BGA**	**Jud.**	**Rea.**
BART		0.319	0.488	0.283	0.315	0.449	0.295
	GRL	**0.377**	**0.528**	**0.340**	0.318	0.452	0.301
	Silver	0.348	0.492	0.293	**0.338**	**0.471**	**0.302**
BART-BERT		0.312	0.474	0.266	0.309	0.445	0.281
	GRL	**0.337**	**0.493**	**0.294**	0.313	0.45	0.286
	Silver	**0.337**	0.479	0.286	**0.334**	**0.47**	**0.294**

Config	Train	Summary Snippet	Explanation for Errors
Blind transfer (RQ1)	UK-Abs	... under section 302 K.P.C. ..... and had a bath there. He The murder was committed ..... mere possession of a sword so stained be not sufficient to establish conclusively. The extra judicial confession of the	It mentions as K.P.C., but it has to be I.P.C. Incomplete sentence: It starts with "He" and digresses to the next sentence. Incoherent discourse: It doesn't mention what can't be established. The closest sentence in the source document is "... not sufficient to establish conclusively that the person who possessed it so shortly after the murder of a person with whom he had enmity, had committed the murder..."
	BVA-Ext	.... under section I.C., by the Session .... cattle shed in village Bhadurpur Ghar. With respect to the learned Judges, theseobservations are not very consistent. To us, it seems that in the middle of June when the chari and sugar cane crop would not have been very	It mentions as I.C instead of I.P.C. and misses actual section number. This sentence in the source document is in context to Ujagar Singh's extrajudicial confession, and not the details of the crime. Incoherent discourse: It digresses into the details of the extrajudicial confession without describing the remaining evidence against the appellant

Table 7: Examples of errors in summaries produced by BART model in zero-shot blind transfer (RQ1) for the Indian Supreme Court judgement https://indiankanoon.org/doc/1811974/, trained with different jurisdiction datasets.

Config	Train	Summary Snippet	Explanation for Errors
GRL (RQ2)	UK-Abs	... under Section 302 K.P.C. ..... murder. There is no good reasons for Ujagar Singh to state falsely. If the statement made	It mentions as K.P.C instead of I.P.C Incomplete sentence: It doesn't mention what Ujagar Singh states falsely.
	BVA-Ext	section I.C., by the Session Judge..... the appellant caused injuries to Sheo Sahai, who was sleeping in his cattle shed in village Bhadurpur Ghar. With respect to the learned Judges, these observations are not very consistent. The learned Sessions Judge has discussed the criticism urged .....it seems that in the middle of June when the chari and sugar cane crop would not have been very high, it is improbable Appeal dismissed. The evidence	It mentions as I.C instead of I.P.C. and misses actual section number. This sentence in the source document is in context to Ujagar Singh's extrajudicial confession, and not the details of the crime. Incomplete sentence: It doesn't describe what is improbable and moves to next sentence.
Silver (RQ3)	UK-Abs	Criminal Appeal No. 1782 of 60 and Referred No.125 of 1960, section 302 I.P.C.....High Court erred in rejecting the statement of Ujagar Singh about the appellant 's confessing to him that The extra judicial	Incomplete sentence, which doesn't describe what the appellant confessed.
	BVA-Ext	... in Criminal Appeal No.125 of 1960, section 302 I.P.C..... for the appellant has argued	Missing complete details: The source document states "...Criminal Appeal No. 1782 of 60 and Referred No. 125 of 1960..."

Table 8: Examples of errors in summaries produced by BART model in GRL setting (RQ2) and silver summary setting (RQ3) for the Indian Supreme Court judgement https://indiankanoon.org/doc/1811974/, trained with different jurisdiction datasets.