Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers

Gechuan Zhang[1]**, David Lillis**[1] **and Paul Nulty**[2]

[1] School of Computer Science, University College Dublin, Ireland

[2] Department of Computer Science, Birkbeck, University of London, UK

gechuan.zhang@ucdconnect.ie
david.lillis@ucd.ie
p.nulty@bbk.ac.uk

Abstract

Interdisciplinary Natural Language Processing (NLP) research traditionally suffers from the requirement for costly data annotation. However, transformer frameworks with pre-training have shown their ability on many downstream tasks including digital humanities tasks with limited small datasets. Considering the fact that many digital humanities fields (e.g. law) feature an abundance of non-annotated textual resources, and the recent achievements led by transformer models, we pay special attention to whether domain pre-training will enhance transformer's performance on interdisciplinary tasks and how. In this work, we use legal argument mining as our case study. This aims to automatically identify text segments with particular linguistic structures (i.e., arguments) from legal documents and to predict the reasoning relations between marked arguments. Our work includes a broad survey of a wide range of BERT variants with different pre-training strategies. Our case study focuses on: the comparison of general pre-training and domain pre-training; the generalisability of different domain pre-trained transformers; and the potential of merging general pre-training with domain pre-training. We also achieve better results than the current transformer baseline in legal argument mining.

1 Introduction

Interdisciplinary natural language processing (NLP) has become one of the most important trends in NLP development. For example, processing of legal text has resulted in research topics such as legal topic classification (Nallapati and Manning, 2008), legal information extraction (Chalkidis et al., 2018), judicial decision prediction (Chalkidis et al., 2019), and legal argumentation mining (Mochales and Moens, 2011). Among these, legal argumentation mining is especially difficult, but has strong application potential, given that arguments are among the most important language structures used in the law. The goal of argument mining is to automatically detect arguments from raw text as well as to identify the reasoning relationships between these arguments (Mochales and Moens, 2011). Argument mining systems that help to identify and analyse argumentative text can assist legal professionals to save time and effort when facing modern document systems with huge quantities of files.

However, the shortage of suitable datasets and the high cost of annotating new datasets impede the application of many advanced NLP approaches (such as neural network and deep learning) to legal argument mining, which is a common issue in most interdisciplinary NLP research. Creating and constructing annotated corpora is complex and labour intensive (Lippi and Torroni, 2016; Poudyal et al., 2020). Particularly when the raw text is domain-specific (e.g., legal text), the annotating experts are required to have extensive knowledge of the corresponding field. This leads to a paradoxical situation whereby a domain with enormous quantities of text resources built up over centuries is served by only a small number of suitable corpora that tend to be limited in their scale.

This dilemma in interdisciplinary research may be solved by using transformer frameworks, such as BERT (Devlin et al., 2019): first pre-training (self-supervised learning) on a large group of roughly labelled text, then fine-tuning on downstream tasks with much smaller datasets that do fine-grained feature annotation. Transformers have revolutionised many research fields including legal text processing (Chalkidis et al., 2019; Reimers et al., 2019). This caused us to examine the potential of reducing the burden of annotation in interdisciplinary research through pre-training. In this work, we use legal argument mining as our case study, because it includes not only the general text classification, but also the relation mining on legal text, and has a strong connection with both the legal field (humanities) and the study of argument mining (NLP).

The primary aim of this paper is to explore the extent to which domain pre-training (i.e. pre-training transformers using legal texts) can improve transformers' performance on legal text processing tasks, without the need for large volumes of expensive annotation. Legal text has its own distinct characteristics when compared with general English-language writing, this also motivates an investigation as to whether legal-specific pre-training can improve upon transformers pre-trained on general-purpose corpora (e.g. Wikipedia, books).

Although Poudyal et al. (2020) set a legal argument mining baseline using RoBERTa (Liu et al., 2019), none of the legal-domain pre-trained transformers (Chalkidis et al., 2020; Zheng et al., 2021) have been applied to legal argument mining tasks to date. In this case study, we try to find (a) whether domain pre-trained transformers outperform general pre-trained transformers in interdisciplinary NLP tasks; (b) whether domain pre-trained transformers maintain good generalisability when applied to tasks using another domain-specific dataset without overlap with the pre-train corpora; and (c) whether merging pre-train corpora from different domains can enhance the transformer performance on interdisciplinary downstream tasks.

In our work, we first provide a thorough survey including a wide range of BERT variants which emphasise two pre-train domains (generic and legal) and use different pre-train strategies. Then we evaluate these transformer models on three legal argument mining tasks: (1) argument clause recognition, (2) argument relation mining, (3) argument component recognition. We discuss the potential of using domain-specific pre-training to adapt state-of-the-art transformer models to interdisciplinary research that lacks large annotated datasets, and we analyse what adjustments in domain-specific pre-training may improve transformers' performance in a complex text processing problem like legal argument mining.

2 Argument and Argument Mining

Long before being treated as a research area in NLP, philosophers and rhetoricians had paid special attention to the logic and reasoning processes embodied in human languages. Numerous schemes and theories have been proposed to define and reason about argumentation. In our work, we use the definition given by Walton (2009) that an *argument* is a set of statements (propositions) that includes three parts: conclusion, premises and inference.

As described in (Lawrence and Reed, 2020), *argument mining* is the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language. There are two crucial stages in the framework of argument mining: *argument extraction* and *relation prediction* (Cabrio and Villata, 2018). Argument extraction is the first stage where the arguments (with their internal structures) are identified from the input documents. Relation prediction is where the support or attack relations between the arguments are predicted.

2.1 Structured Argumentation Model

Structured argumentation is one of the main approaches in computational argumentation, which presents an internal structure for each argument, described in terms of some knowledge representation. For structured argumentation models typically applied in argument mining tasks, defining the internal structure of an argument is crucial (Lippi and Torroni, 2015). Such models consider different argument components inside each argument and both internal and external argument relationships:

- **Argument Component** is the smallest unit in structured argumentation model. The argument components connect to each other through the internal relations.

- **Argument Relation** has two different levels in a structured argumentation model: internal and external. Internal argument relations are used to connect elementary argument components into a whole group (i.e., each argument). External argument relations represent the reasoning process between different arguments in a complete text document.

2.2 Walton Argumentation Model

The typical guideline for annotating legal argument mining corpora is Walton's structured argumentation model (Walton, 2009). This has two types of argument components: premises and conclusions. Walton (2009) described each argument as a set of statements (propositions) made up of three parts: a *conclusion*, a set of *premises*, and an *inference* from the set of premises to the conclusion. The conclusion is a claim or a statement which acts as the central component of an argument. The set of premises are the evidences or reasons given to support the conclusion. The inference is the internal argument relation in the Walton argumentation model. For the external argument relations, Walton (2009) defined a set of bipolar relations: an argument can be supported by other arguments, or it can be attacked by other arguments (through raising critical questions about it).

3 BERT-Based Transformers

BERT (Bidirectional Encoder Representations from Transformers) is a contextual word embedding model using the deep transformer architecture (Vaswani et al., 2017) to derive word features (Devlin et al., 2019). It leverages a two-step framework: pre-training and fine-tuning. During pre-training, BERT is trained on a large corpus using self-supervised learning methods. Then, in fine-tuning, the model is tuned on the downstream task's dataset, which is usually much smaller. BERT has achieved state-of-the-art performance on many legal text processing tasks (Chalkidis et al., 2019; Reimers et al., 2019; Poudyal et al., 2020). Since then, several studies have addressed whether pre-training on legal texts can improve transformers' performance on downstream tasks in the same domain (Elwany et al., 2019; Chalkidis et al., 2020; Zhong et al., 2020a,b; Zheng et al., 2021), but none have been evaluated on legal argument mining.

4 BERT Pre-Train Strategies

In order to analyse how different pre-trained transformers perform on legal argument mining tasks, we compare five BERT$_{base}$ (L=12, H=768, A=12, 110M params (Devlin et al., 2019)) variants from two pre-train categories: general pre-trained models using generic English corpora for pre-training, and domain pre-trained models using English legal text in their pre-train corpus. Here we first provide an outline background of each transformer. Then, based on previous studies (Devlin et al., 2019; Liu et al., 2019), we compare these models across three key aspects of pre-train strategy: pre-train corpora selection, pre-train procedure, and text encoding.

4.1 BERT-based Transformers

RoBERTa
A widely used BERT$_{base}$ variant, pre-trained on large generic English corpora (Liu et al., 2019), which constitutes the latest baseline model for legal argument mining (Poudyal et al., 2020).

LEGAL-BERT Family
Legal-BERT$_{echr}$ and Legal-BERT$_{base}$ are two domain pre-trained transformers selected from the LEGAL-BERT family (Chalkidis et al., 2020). Instead of using general English corpora, the LEGAL-BERT family, a group of legal-specific BERT$_{base}$ variants, are pre-trained on a English legal text collection (see Table 1) with two different domain-adaptation methods: (a) further pre-train BERT$_{base}$ on legal text before fine-tuning, (b) pre-train BERT$_{base}$ from scratch on legal text before fine-tuning. More precisely, Legal-BERT$_{echr}$ is further pre-trained on the European Court of Human Rights (ECHR) legal case subset (see Table 2) from the BERT$_{base}$ checkpoint. Legal-BERT$_{base}$ is pre-trained from scratch on the whole collection of legal corpora.

Harvard Legal-BERT Variants
Distinct from the LEGAL-BERT family, two other domain-specific BERT$_{base}$ transformers from (Zheng et al., 2021) are pre-trained with the Harvard Law case corpus (see Table 1). To avoid name confusion, we use *Legal-BERT*$_{harv}$ and *Custom Legal-BERT*$_{harv}$ to represent Legal BERT and Custom Legal BERT in the original literature. Similar to Chalkidis et al. (2020), Zheng et al. (2021) also assess both *further pre-training* and *pre-training from scratch* domain-adaptation methods. Similar to Legal-BERT$_{echr}$, Legal-BERT$_{harv}$ is further pre-trained on the Harvard Law case corpus. Custom Legal-BERT$_{harv}$ is pre-trained from scratch using the same corpus with a custom vocabulary (see Section 4.4).

4.2 Pre-train Corpora Selection

In order to extract long contiguous sequences, both datasets in the BERT$_{base}$ original pre-train corpora are long documents (i.e., books and Wikipedia passages). Since Liu et al. (2019) suggest that BERT$_{base}$ is still under-trained, RoBERTa enlarged the scale of its pre-train corpora 10 fold (from 16 GB to 161 GB) by including news articles and online discussion web text. As for the domain pre-train corpora, Chalkidis et al. (2020) collected a wide range of English legal documents and cases with different functions, backgrounds, and text formats (i.e., legislation, case judgements, legal contracts). Zheng et al. (2021) focused on US legal decisions from the Harvard Law case corpus and gather a larger dataset (37 GB), which is three times larger than the LEGAL-BERT family (11.5 GB).

Model	Text Type	Size (GB)
RoBERTa	BooksCorpus Wikipedia	16
	CC-News	76
	OpenWebText	38
	STORIES	31
	total	**161**
Legal-BERT$_{base}$	EU legislation	1.9
	UK legislation	1.4
	ECJ case	0.6
	ECHR case	0.5
	US court case	3.2
	US contract	3.9
	total	**11.5**
Custom Legal-BERT$_{harv}$	Harvard Law case	**37**

Table 1: Different BERT Variant Pre-train Corpora

Model	P Corpus	FP Corpus
Legal-BERT$_{echr}$	BooksCorpus Wikipedia 16 GB	ECHR case 0.5 GB
Legal-BERT$_{harv}$		Harvard Law case 37 GB

Table 2: Different BERT Variant Pre-train (P) Corpora and Further Pre-train (FP) Corpora

Because Legal-BERT$_{echr}$ and Legal-BERT$_{harv}$ are first initialised from the BERT$_{base}$ checkpoint, which has been pre-trained on generic corpora, then further pre-trained on legal corpora, both variants have mixed pre-train corpora. Legal-BERT$_{harv}$ uses the same legal corpora as Custom Legal-BERT$_{harv}$ for further pre-training. The ECHR case sub-set used by Legal-BERT$_{echr}$ for further pre-training is only 0.5 GB, which is much smaller compared to other pre-train corpora.

4.3 Pre-train Procedure

The original BERT$_{base}$ pre-train procedure contains two objectives: masked language modelling (MLM), which aims to train the model for a deep bidirectional representation, and next sentence prediction (NSP) which aims to train the model for sentence relationship understanding (Devlin et al., 2019). In place of performing MLM once during data pre-processing (static masking), RoBERTa generates the masking pattern for each sequence input (dynamic masking) and removes the NSP loss (Liu et al., 2019). The LEGAL-BERT family use the same pre-train objectives as the original BERT$_{base}$ (Chalkidis et al., 2020). The Harvard Legal-BERT variants make adjustments based on the characteristics of legal corpora. In constrast with BERT$_{base}$, which selects and replaces the tokens in the input sequence, Zheng et al. (2021) use whole word masking in MLM and adds regular expressions to ensure legal citations are included as part of a segmented sentence in NSP.

Liu et al. (2019) suggest that training the BERT model longer with larger batches improves its performance. The original pre-training setup of BERT$_{base}$ is 1M steps and 256 sequences per batch. RoBERTa replaces this with 500K steps and a batch size of 8K. For the two domain pre-trained models from the LEGAL-BERT family, Legal-BERT$_{base}$ uses the same setup as BERT$_{base}$ when being pre-trained from scratch; while Legal-BERT$_{echr}$ is first initialised from BERT$_{base}$'s 1M checkpoint then further pre-trained with another 5K on legal text. Like RoBERTa, Zheng et al. (2021) train the model longer by using 2M total pre-train steps for both Harvard Legal-BERT variants. In particular, Custom Legal-BERT$_{harv}$ is pre-trained from scratch with 2M steps, and Legal-BERT$_{harv}$ is initialised from the 1M checkpoint of BERT$_{base}$ (same as Legal-BERT$_{echr}$) then further pre-trained with 1M steps on legal case documents (see Table 2 and 3).

4.4 Text Encoding

To encode text pieces into vectors, the BERT$_{base}$ transformer first splits the input raw text into words or sub-words through a tokenizer. These word pieces are then converted to ids by using pre-designed vocabularies. The original BERT$_{base}$ is implemented with the WordPiece tokenizer (Schuster and Nakajima, 2012) and a character-level Byte-Pair Encoding (BPE) (Sennrich et al., 2016) vocabulary (size 30K). Both Legal-BERT$_{echr}$ and Legal-BERT$_{harv}$ use the same tokenizer and BPE vocabulary from BERT$_{base}$ during the further pre-training. Rather than using character-level sub-word unit, RoBERTa's implementation uses the same tokenizer as (Radford et al., 2018) with a larger byte-level BPE vocabulary (size 50K). Using byte-level BPE makes it possible to encode any input text without introducing "unknown" tokens.

In order to adapt BERT$_{base}$ from generic English corpora to legal text, both Legal-BERT$_{base}$ and Custom Legal-BERT$_{harv}$, apply the SentencePiece (Kudo and Richardson, 2018) tokenizer with self-generated vocabularies. Legal-BERT$_{base}$ used a newly-created vocabulary of equal size to BERT$_{base}$ (30K), constructed on its complete pre-train legal corpus (see Table 1). Custom Legal-

Model	Training Objectives	Pre-train Setup			Further Pre-train Setup			Encoding
		Type	Step	Batch	Type	Step	Batch	
RoBERTa	Dynamic MLM	Generic	500K	8K	-	-	-	Byte BPE
Legal-BERT$_{base}$	MLM, NSP	Legal	1M	256	-	-	-	SP
Legal-BERT$_{echr}$		Generic	1M	256	Legal	5K	256	WP
Custom Legal-BERT$_{harv}$	Whole-Word MLM, Regexp NSP	Legal	2M	256	-	-	-	SP
Legal-BERT$_{harv}$		Generic	1M	256	Legal	1M	256	WP

Table 3: Different BERT Variant Pre-train Design (SP = SentencePiece, WP = WordPiece)

BERT$_{harv}$ also uses a legal domain-specific vocabulary (32K), which is constructed on a sub-sample of sentences of the Harvard Law case corpus.

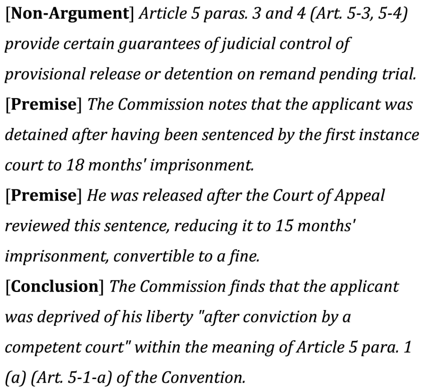
5 ECHR Dataset

The European Court of Human Rights (ECHR) case-law dataset, developed from the HUDOC database[1], is an open-source database of case documents, and has become one of the most commonly-used datasets for legal text processing research such as judicial decision prediction (Chalkidis et al., 2019; Medvedeva et al., 2020), court decision event extraction (Filtz et al., 2020), and legal argument mining (Mochales and Moens, 2011; Teruel et al., 2018; Poudyal et al., 2020).

The ECHR case-law dataset has been used for legal argument mining research from an early stage (Mochales and Moens, 2011). (Mochales and Moens, 2008) provides a detailed structural analysis of ECHR documents. Several legal argument mining corpora have been established based on the ECHR case-laws (Mochales and Moens, 2011; Teruel et al., 2018; Poudyal et al., 2020). Among them, we choose the recently released ECHR argument mining corpus (ECHR-AM) (Poudyal et al., 2020) for our experiments. The ECHR-AM corpus contains 42 cases, 20 decisions (the average word length is 3,500 words) and 22 judgements (the average word length is 10,000 words). The entire corpus is annotated at the sentence level using three labels according to the Walton Argumentation Model (see Section 2.2): premise, conclusion, and non-argument. The annotation focuses on internal argument relations which includes a total of 1,951 premises and 743 conclusions acting as argument components for individual arguments.

6 Legal Argument Mining Tasks

Our case study currently focuses on the argument extraction, which is the first stage within a typical argument mining framework (as mentioned in Section 2). Following the example of (Poudyal et al., 2020), we organise this as three tasks.



[**Non-Argument**] *Article 5 paras. 3 and 4 (Art. 5-3, 5-4) provide certain guarantees of judicial control of provisional release or detention on remand pending trial.*
[**Premise**] *The Commission notes that the applicant was detained after having been sentenced by the first instance court to 18 months' imprisonment.*
[**Premise**] *He was released after the Court of Appeal reviewed this sentence, reducing it to 15 months' imprisonment, convertible to a fine.*
[**Conclusion**] *The Commission finds that the applicant was deprived of his liberty "after conviction by a competent court" within the meaning of Article 5 para. 1 (a) (Art. 5-1-a) of the Convention.*

Figure 1: Annotation Example of the ECHR Argument Mining Corpus

6.1 Argument Clause Recognition

The first task is to filter those sentences that are argumentative from those that are not. We treat this task as a binary text classification, in which the segmented clauses from the case law documents are classified into two groups: argument clauses and non-argument clauses. The argument clauses are those sentences which functionally act as argument components in arguments (see Section 2.2).

6.2 Argument Relation Mining

This task focuses on identifying the argument relations that link argument components (i.e., argument clauses) within each argument. Here, the *argument relation* is the internal relation (i.e., inference) in the Walton argumentation model. The ultimate goal of this task is to label argument clauses that appear in the same argument as being in the same group. Since the same argument clause may appear in different arguments (for example, a single clause can be the conclusion of one argument and also the premise of another), this task is more difficult than a general text clustering problem. Previous studies imply this task is probably the bottleneck in the argument mining framework (Mochales and Moens, 2011; Poudyal et al., 2019, 2020).

[1] http://hudoc.echr.coe.int/

Instead of directly grouping argument clauses into individual arguments, we consider the solution in (Poudyal et al., 2020) and treat this task as an argument clause pair classification task. We analyse whether or not a pair of clauses are argument components from the same argument. This can help with the multi-correspondence issue between argument clauses and arguments. To get the input argument clause pairs, we order the argument clauses from the same case document into a sequence, then use a sliding window to generate input clause pairs. Next, we use transformers to predict whether the pair are related.

6.3 Argument Component Classification

The final task is to classify the argument clauses as premises or conclusions. Because an argument clause may belong to multiple arguments, and can be either a premise or a conclusion, we separated the argument component classification task into two individual binary classification sub-tasks (premise recognition and conclusion recognition). More specifically, if an argument clause is tagged as both premise and conclusion in the classification sub-tasks, it is included in multiple arguments. As a conclusion, the argument clause itself represents an individual argument connecting with other premise clauses. As a premise, this argument clause is also a part of an argument, whose conclusion is linked with this clause in the argument relation mining task.

7 Experiments

7.1 Experimental Setup

Baseline: Following the baseline setup given by (Poudyal et al., 2020), we use 5-fold cross validation during our experiment. We split 80% case law documents for training and the remaining 20% for testing. Of the training documents, we randomly select 20% for validation. The number of documents in each train-validation-test split is therefore 28-6-8. During each fold, we select the model with the best F-score on the validation set for testing. We performed five runs for each model and reported mean evaluation scores with standard deviations. For the baseline, we refer to the records in Poudyal et al. (2020) and use RoBERTa for extra tests (in argument clause recognition and argument relation mining). Moreover, to better understand the enhancement given by the BERT model, we also include an additional non-BERT baseline. A number of candidate approaches based on word embeddings (Wang and Lillis, 2020) were considered. We choose the one-layer BiLSTM architecture used by Zheng et al. (2021), tested with 300-dimension word embedding. For each task, we first encode the segmented clause with the transformer, then pool the final CLS token from the embedding vector and input it into the classifier head. For each selected BERT-based transformer in the experiment, we add the same classifier head containing a dropout layer (dropout rate = 0.1) and a liner layer (for the final classification task).

Hyper-parameters: Because Poudyal et al. (2020) do not provide full details of the hyper-parameters used in their experiments, we consult the hyper-parameter setups in (Devlin et al., 2019; Chalkidis et al., 2020; Zheng et al., 2021) for guidance on typical experiment configurations. Similar to Zheng et al. (2021), we perform the first round of grid search for learning rate in the range {2e-5, 3e-5, 4e-5, 5e-5} suggested by Devlin et al. (2019), then we expand this range with {5e-6, 1e-5, 6e-5} to test the boundary. Considering the small size of the ECHR-AM corpus, we search over batch size {8, 16, 32} as recommended by Chalkidis et al. (2020). Poudyal et al. (2020) set 15 fine-tune epochs for each baseline task, which we found redundant due to the fact that the BERT-base transformers are well trained and can adapt quickly on small corpora. We fine-tune each model with 4 epochs in each task.

7.2 Argument Clause Recognition Results

As is mentioned in Poudyal et al. (2020), over 99% of the argument clauses of the ECHR-AM dataset are present in a specific section ("AS TO THE LAW/ THE LAW") of each document. The baseline argument clause recognition task first uses text matching to detect target sections, then classifies segmented sentences (clauses outside that section are automatically predicted as non-argument). The upper part of Table 4 shows the results for the argument clause recognition after section detection. The baseline (from Poudyal et al. (2020); generated by RoBERTa) was outperformed by all the domain pre-trained transformers. Both models from the LEGAL-BERT family reached the highest F1 score (79.3%), which is probably because their pre-train text collection includes ECHR cases. Among all the domain pre-trained transformers, it is impressive that Legal-BERT$_{echr}$ only used 0.5 GB legal text (12K ECHR cases) in its further pre-train and gained a competitive performance on the argument clause recognition task. Further pre-trained Legal-BERT$_{harv}$ also reached a better precision (74.3% vs. 69.7%) than the RoBERTa baseline. Although BiLSTM achieves the highest recall, its relatively low precision leads to an F1 score that is below the BERT-based models.

Legal Sect	**P (%)**	**R (%)**	**F1 (%)**
baseline	69.7	84.8	76.5
BiLSTM	62.4±6.5	**91.8**±4.4	74.0±3.5
Legal-BERT$_{base}$	72.4±2.4	88.1±1.4	**79.3**±1.3
Legal-BERT$_{echr}$	73.5±2.1	86.5±2.2	**79.3**±1.3
C-Legal-BERT$_{harv}$	73.4±1.9	84.2±1.9	78.2±1.8
Legal-BERT$_{harv}$	**74.3**±1.8	84.0±0.8	78.7±1.0
Whole Doc	**P (%)**	**R (%)**	**F1 (%)**
RoBERTa	65.3±1.8	71.0±4.5	67.7±2.4
BiLSTM	61.0±7.1	51.4±7.5	55.5±6.1
Legal-BERT$_{base}$	66.1±1.8	**73.6**±3.9	69.3±2.2
Legal-BERT$_{echr}$	**67.5**±2.0	73.1±2.7	**69.9**±2.1
C-Legal-BERT$_{harv}$	65.3±2.8	69.8±3.9	67.1±2.3
Legal-BERT$_{harv}$	66.3±1.7	70.0±3.2	67.9±1.9

Table 4: Precision (P), recall (R), F1 measurement (± std. dev.) for the argument clause recognition task on the “AS TO THE LAW/ THE LAW” Section scope and the whole document scope (C = Custom).

The section detection in the baseline argument clause recognition task filters out a large group of non-argument clauses, and balances the candidate clause dataset. The number of input segmented clauses shrank from 10,456 to 4,683. To generalise this approach to practical applications, we expand the searching area to the complete document (10,456 clauses) and test again. The results are displayed in the lower part of Table 4. All domain pre-trained models exceeded the RoBERTa baseline, except the Custom Legal-BERT$_{harv}$ whose score is slightly lower (67.1% vs. 67.7%). Legal-BERT$_{echr}$ remained the best F1 score (69.9%). Among the four legal-specific transformers, the models pre-trained with further steps had slightly better scores than the other two models pre-trained from scratch. Considering the scale of each model's pre-train corpus: RoBERTa was pre-trained on 161 GB text, while other BERT models used much smaller pre-train corpora. The evaluation scores in this task suggest that domain-specific pre-training is effective when downstream tasks in argument mining (e.g., text classification) are focusing on text with a similar domain-specific background.

7.3 Argument Relation Mining Results

To be consistent with Poudyal et al. (2020), we assume that all the argument clauses have been successfully identified from previous task. As discussed in Section 6.2, we use a sliding window on the argument clause sequence to generate pairs of argument clauses. In order to compare the RoBERTa baseline and different domain pre-train variants, we first use the window size 5 mentioned in Poudyal et al. (2020). The upper part of Table 5 shows the results with domain pre-training again displaying its effectiveness when mining relations between clause pairs. All the domain pre-trained transformers substantially exceeded the baseline F1 score (8.4% on average). The Harvard Legal BERT variants slightly outperformed the LEGAL-BERT family in each corresponding pre-train type (pre-train from scratch, 59.9% vs. 58.1%; further pre-train, 60.6% vs. 59.4%). Among the four domain-specific pre-trained transformers, models using further pre-train strategy again displayed a further slight advantage. By using the further pre-train approach and adding special pre-training adjustments (whole word MLM, Regexp NSP, see Section 4.3), Legal-BERT$_{harv}$ reached the best scores for precision (59.7%), recall (62.1%) and F1 (60.6%).

window size = 5	**P (%)**	**R (%)**	**F1 (%)**
baseline	50.2	52.1	51.1
BiLSTM	45.5±5.5	52.7±14.6	47.8±6.1
Legal-BERT$_{base}$	57.3±2.4	59.7±3.5	58.1±2.6
Legal-BERT$_{echr}$	59.5±3.8	60.0±3.8	59.4±1.2
C-Legal-BERT$_{harv}$	58.7±2.4	61.2±1.6	59.9±1.8
Legal-BERT$_{harv}$	**59.7**±2.6	**62.1**±2.5	**60.6**±1.6
window size = 10	**P (%)**	**R (%)**	**F1 (%)**
RoBERTa	**47.2**±3.6	35.2±1.6	39.4±2.5
BiLSTM	26.3±3.8	**45.8**±20.7	31.3±7.3
Legal-BERT$_{base}$	45.8±4.9	36.4±4.3	39.6±1.8
Legal-BERT$_{echr}$	46.8±3.2	41.4±1.8	43.3±1.3
C-Legal-BERT$_{harv}$	45.8±2.3	41.8±2.4	43.2±2.1
Legal-BERT$_{harv}$	47.1±2.4	43.6±2.9	**44.5**±2.0

Table 5: Precision (P), recall (R), F1 measurement (± std. dev.) for the argument relation mining task with window size = 5 and window size = 10 (C = Custom).

After analysing the ECHR argument mining dataset, we decided to enlarge the window size to 10, in which almost all the actual argument relations are included, while the number of total pairs has not increased to an unmanageable degree (by doubling the window size, the total number of argument clause pairs increased from 10,356 to 22,329). Due to the imbalance of the argument clause pair dataset, the general performance of all transformers were lower as expected. Legal-BERT$_{harv}$ again reached the best F1 score (44.5%).

7.4 Argument Component Classification Results

As mentioned in Section 6.3, the argument component classification task consists of two sub-tasks: premise recognition and conclusion recognition. In the same way as (Poudyal et al., 2020), we assume that we have successfully identified argument clauses in the previous task. For the premise recognition sub-task, domain pre-training improves the BERT-base model's performance over all three evaluation scores, as shown in Table 6: Custom Legal-BERT$_{harv}$ reaches both highest recall (91.5% vs. 88.7%) and F1 (87.2% vs. 85.9%) among all the transformers; Legal-BERT$_{harv}$ also achieves a moderate improvement in precision (83.9% vs. 83.2%). Compared to the BERT-based models, BiLSTM has the highest recall value but a much lower precision, and is less robust across runs.

Premise	**P (%)**	**R (%)**	**F1 (%)**
baseline	83.2	88.7	85.9
BiLSTM	79.2±2.7	**94.6**±2.3	86.2±1.4
Legal-BERT$_{base}$	83.8±1.4	90.3±1.8	86.8±0.7
Legal-BERT$_{echr}$	83.5±2.5	90.4±1.7	86.7±0.9
C-Legal-BERT$_{harv}$	83.4±1.3	91.5±1.3	**87.2**±0.6
Legal-BERT$_{harv}$	**83.9**±1.6	90.3±1.8	86.9±0.9
Conclusion	**P (%)**	**R (%)**	**F1 (%)**
baseline	58.9	**67.2**	62.8
BiLSTM	54.9±11.3	54.0±11.2	52.6±4.7
Legal-BERT$_{base}$	65.2±2.8	61.2±5.2	61.9±3.0
Legal-BERT$_{echr}$	65.3±1.1	62.0±3.5	63.3±2.3
C-Legal-BERT$_{harv}$	**67.1**±0.9	60.1±3.7	62.9±2.0
Legal-BERT$_{harv}$	66.2±0.9	63.1±3.5	**64.2**±1.8

Table 6: Precision (P), recall (R), F1 measurement (± std. dev.) for the argument component (premise/conclusion) classification task (C = Custom).

For the conclusion recognition sub-task, Legal-BERT$_{harv}$ outperforms the RoBERTa baseline (64.2% vs. 62.8%) with the highest recall (63.1%) among all domain pre-trained transformers. Custom Legal-BERT$_{harv}$ also reaches the best precision (67.1%). Generally, pre-trained legal-BERT models show better precision than recall, in contrast to the baseline (58.9% precision, and 67.2% recall). Since (Poudyal et al., 2020) does not provide multiple cross-validation records, we suggest that this difference may be a result of randomness bias. Overall, the Harvard Legal BERT variants slightly outperformed the LEGAL-BERT family.

8 Discussion

Our case study on legal argument mining gives us insights on the potential of using domain pre-training to reduce the data annotation burden in interdisciplinary NLP research, as well as help us better understand the relationship between domain pre-training and domain-specific downstream tasks. To answer question (a) in Section 1, it is clear that domain pre-trained transformers work better than general pre-trained transformers in all three legal argument mining tasks, which also exceed the baseline from (Poudyal et al., 2020). This supports the idea that domain pre-training helps improve transformer's performance on downstream tasks where only small datasets are available. Combining the scope of pre-train corpora used by each transformer with its performance, we suggest that using small domain-specific pre-training corpora would be as effective as using a large general corpora. In Section 7, the LEGAL-BERT family use a much smaller legal text collection (11.5 GB) compared with RoBERTa's 161 GB general pre-train corpus, but also achieve competitive results despite less pre-training steps (see Table 3).

Both Harvard Legal-BERT variants present good performance on the ECHR-AM dataset. In contrast to the LEGAL-BERT family which includes ECHR cases as part of its pre-train legal text collection, the pre-train corpus used by Harvard Legal-BERT variants has no overlap with the ECHR-AM dataset. Therefore, for question (b) in Section 1, our case study indicates that domain pre-trained transformers can maintain good generalisability on downstream tasks focusing on different datasets. This "reusable" characteristic of domain pre-trained models is significant. Collecting relevant pre-train corpora and pre-training itself still require sufficient time and computing resources. If the domain pre-trained model is reusable in different domain-specific tasks, sharing domain pre-trained models will be superior to sharing corpora, especially for research groups who do not have the resources or capability for pre-training models on large corpora.

With reference to question (c) in Section 1 about merging domain pre-training with general pre-training, the experiment results indicate that the transformers using further pre-training work slightly better than those using pre-training from scratch. This indicates the potential of enhancing the transformer's generalisability on downstream domain-specific tasks by merging generic corpora with further pre-training, especially in tasks like argument relation mining that require the model to extract not only the features from the text but also the potential relations between different sequences.

Our work pays attention to a common issue in interdisciplinary NLP research that the background area (i.e., humanities, law) has considerable volumes of text material, but the annotation work is costly and impedes the process of adapting advanced NLP technologies to assist the research. We suggest that pre-train transformers can help this "data poverty" issue, and domain-specific pre-training will improve transformers' performance when adapting to interdisciplinary tasks with only small fine-annotated datasets.

We analyse state-of-the-art transformers in both pre-train categories, and present a comprehensive survey of available models for the legal domain. Our case study provides the first comparison between general pre-trained transformers and domain pre-trained transformers on legal argument mining tasks, and demonstrates that domain pre-trained transformers achieve better results on all three tasks than the baseline outlined by Poudyal et al. (2020). Our case study also compares the performance of two latest groups of transformers in legal domain, and offers an analysis of some key aspects when applying domain pre-training on interdisciplinary NLP tasks.

References

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

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, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical rnns. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Emad Elwany, Dave Moore, and Gaurav Oberoi. 2019. Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.

Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. 2020. Events matter: Extraction of events from court decisions. In *Legal Knowledge and Information Systems*, pages 33–42. IOS Press.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP (Demonstration)*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *International Workshop on Theory and Applications of Formal Argumentation*, pages 163–176. Springer.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Raquel Mochales and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *Legal Knowledge and Information Systems*, pages 11–20. IOS Press.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Ramesh Nallapati and Christopher D Manning. 2008. Legal docket-entry classification: Where machine learning stumbles. In *2008 Conference on Empirical Methods in Natural Language Processing*, page 438.

Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2019. Using clustering techniques to identify arguments in legal documents. In *ASAIL@ ICAIL*.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Gonçalves, and Paulo Quaresma. 2020. Echr: legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Douglas Walton. 2009. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer.

Congcong Wang and David Lillis. 2020. A Comparative Study on Word Embeddings in Deep Learning for Text Classification. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2020)*, Seoul, South Korea.

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. *arXiv preprint arXiv:2104.08671*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jecqa: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9701–9708.