Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

Enhancing Legal Argument Mining with Domain Pre-training and Neural Networks

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

[1]School of Computer Science, University College Dublin, Ireland
[2]Department of Computer Science, Birkbeck, University of London, UK

Corresponding author: David Lillis , david.lillis@ucd.ie

Abstract

The contextual word embedding model, BERT, has proved its ability on downstream tasks with limited quantities of annotated data. BERT and its variants help to reduce the burden of complex annotation work in many interdisciplinary research areas, for example, legal argument mining in digital humanities. Argument mining aims to develop text analysis tools that can automatically retrieve arguments and identify relationships between argumentation clauses. Since argumentation is one of the key aspects of case law, argument mining tools for legal texts are applicable to both academic and non-academic legal research. Domain-specific BERT variants (pre-trained with corpora from a particular background) have also achieved strong performance in many tasks. To our knowledge, previous machine learning studies of argument mining on judicial case law still heavily rely on statistical models. In this paper, we provide a broad study of both classic and contextual embedding models and their performance on practical case law from the European Court of Human Rights (ECHR). During our study, we also explore a number of neural networks when being combined with different embeddings. Our experiments provide a comprehensive overview of a variety of approaches to the legal argument mining task. We conclude that domain pre-trained transformer models have great potential in this area, although traditional embeddings can also achieve strong performance when combined with additional neural network layers.

I INTRODUCTION

Interdisciplinary research like digital humanities has been one of the most important trends in Natural Language Processing (NLP) research, where machine-based methods and tools are developed to automatically analyse texts from traditional humanities. Compared to general text analysis, in digital humanities many advanced approaches such as neural networks are still under-explored due to a lack of annotated datasets, which is a requirement of many supervised machine learning approaches. The complexity and labour cost required to produce new corpora constitute a major barrier in many areas of digital humanities research [Zhang et al., 2021]. Our specific research focus of legal argumentation mining is no exception in this regard.

The argument, a series of statements intended to determine the degree of truth for another statement, is one of the most important language structures used in the law. The ultimate goal of argument mining is to automatically identify arguments as well as their reasoning relations from texts [Mochales and Moens, 2011]. For legal argument mining, creating corpora that can be used to develop automated systems requires case law to be annotated in a way that identifies the argumentation components and the relationship between them. This implies that the annotators should have sufficient related knowledge, and so the annotation process usually requires expert legal professionals such as lawyers or law school students. Although this legal Artificial Intelligence (AI) field has attracted much attention, there has still been a lack of development and successful deployment of applications, in part due to the dearth of comprehensive corpora for research.

Meanwhile, large pre-trained transformer models like BERT [Devlin et al., 2019] have achieved outstanding performance on downstream tasks, even where only a limited amount of annotations are available. This has inspired a series of research studies on legal AI [Chalkidis et al., 2019, Reimers et al., 2019]. The pre-training plus fine-tuning strategy has improved both efficiency and performance when applying BERT. In particular, the model is first trained on a large dataset, then fine-tuned on the downstream tasks. The initial BERT model was trained on general corpus, which has inspired researchers to pre-train the model with different domain-specific corpora, for example legal texts [Chalkidis et al., 2020]. Recent studies of BERT-base transformers pre-trained with legal corpora have displayed better performance on several legal AI tasks [Xu et al., 2021b, Silveira et al., 2021]. To gain insights into the improvements acheived by BERT-based models on legal argument mining, in this paper we compare the performance from two groups of embedding models: four BERT-based transformers pre-trained with legal texts, and two non-BERT embedding models. We also explore the enhancement of classic NLP neural networks on argument mining tasks.

Section II discusses the general background of argument mining and the original BERT model as well as introducing the domain pre-trained BERT variants and neural networks used in our experiments. Section III provides details of the European Court of Human Rights (ECHR) case law corpus for argument mining. Section IV contains our experiment design and model implementation for argument extraction and relation prediction. We analyse the results in Section V and conclude our work in Section VI, along with a discussion of potential future work.

II RELATED WORK

2.1 Argument Mining

As mentioned in Section I, argument mining aims to automatically retrieve arguments and their related information from human language texts. Its interdisciplinary background makes argument mining a high-level research question in NLP, which is usually formalised as having two stages: *argument extraction* and *relation prediction* [Cabrio and Villata, 2018]. The first stage, argument extraction, aims to shrink the scope of argumentative texts (texts containing argument information) by filtering out unrelated parts and therefore focusing only on those sentences that are argumentative. During the second stage of relation prediction, the relations between identified argumentative texts are predicted. Predicting the inner relations between argument components, or the outer relations between individual arguments, or both, depends on the practical requirements and the annotation scheme. Here, we focus on the inner structure of arguments, which includes identifying argument components and the relations between them.

To facilitate reasoning about arguments, a computational model of argument is needed. These are generally divided into two major categories: structural argumentation models and abstract argumentation models. The abstract argumentation model, also known as argumentation frameworks in Dung [1995], regards arguments themselves to be the elementary units, without additional internal structures. Nevertheless, the complexity of legal texts requires such internal argument structures, which leads to structural argumentation models becoming the main approaches when annotating legal corpora. A structural argumentation model usually consists of components and relations. Argument components are the elementary units, which are usually defined and annotated with their logic roles (e.g., premise, conclusion). Argument relations are the reasoning connections (e.g., support, defeat) between argument components (internal) and between individual arguments (external). Our experiment corpus currently focuses on internal argument relations, in particular the support relation between premise and conclusion.

One of the representative annotation standards for legal texts is the Walton [2009] model. This is a tripartite structural argumentation model: a set of *premises* that contain the evidence or reasons for supporting an argument, a *conclusion* which is the stance and the central component of an argument, and the *inference* from the set of premises to the conclusion. The high generalisability of Walton model makes it suitable for various contexts, for example, ECHR case law.

2.2 BERT-based Models

In order to develop tools for text analysis, models like word embeddings are applied to express human language features as computational vectors. As an advanced word embedding model, BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019] has achieved leading performance on several NLP areas, including legal text processing [Chalkidis et al., 2019, Reimers et al., 2019, Poudyal et al., 2020]. BERT is a contextual word embedding model extracting text features with a deep transformer architecture [Vaswani et al., 2017]. The complete training procedure of BERT is a two-step process. First, the model is trained on a large roughly labelled corpus using self-supervised learning methods. Next, when being adapted to downstream tasks, the model is further trained to fine-tune its weight-matrix with a well-annotated corpus that is usually much smaller.

2.2.1 BERT Pre-train Strategy

A 16GB English corpus collected from online books [Zhu et al., 2015] and Wikipedia were used for the pre-training of the original BERT model, BERT$_{base}$. The self-supervised learning process during the pre-training procedure has two objectives, masked language modelling (MLM) and next sentence prediction (NSP). Together, the pre-train process aims to enhance the model for deep bidirectional representations as well as sentence relationship understanding [Devlin et al., 2019].

2.2.2 Legal Domain Pre-trained Models

Legal language is considered to be a unique writing system which differs from generic text materials. Several researchers have explored whether using domain-specific pre-training corpora can enhance the performance of the BERT-base transformer for downstream tasks from the same domain [Alsentzer et al., 2019, Beltagy et al., 2019, Lee et al., 2020]. In our case, we focus on the BERT-base models pre-trained with legal text materials. To simplify the description, we use *legal BERT models* as the BERT$_{base}$ variants which are pre-trained on text materials from the legal field. Previous legal text processing studies [Elwany et al., 2019, Chalkidis et al., 2020, Zhong et al., 2020a,b, Zheng et al., 2021] have demonstrated the improvements given by legal BERT models. In our experiment, we selected two groups of legal BERT models, the *LEGAL-BERT Family* and the *Harvard Legal-BERT Variants*.

LEGAL-BERT Family includes a series of BERT-based models using English legal texts for pre-training. Among them, we select two models: Legal-BERT$_{base}$ and Legal-BERT$_{echr}$. The total amount of pre-training text data collected for the LEGAL-BERT Family is 11.5GB, which covers several groups of legal documents, including EU and UK legislation, US court cases and contracts. Chalkidis et al. [2020] explored two different domain-adaptation methods: 1) pre-training from scratch, 2) further pre-training. Legal-BERT$_{base}$ is pre-trained from scratch on the 11.5GB legal text collection, while Legal-BERT$_{echr}$ is further pre-trained with 0.5GB of ECHR case documents collected from the HUDOC[1] dataset.

Harvard Legal-BERT Variants are two legal BERT models pre-trained with judicial texts from the US court. In particular, Zheng et al. [2021] collected 37GB of text from the Harvard Law case corpus[2]. We name the two variants as Legal-BERT$_{harv}$ and Custom Legal-BERT$_{harv}$. Similar to [Chalkidis et al., 2020], Zheng et al. [2021] also trained their legal BERT models with different adaptation methods. Custom Legal-BERT$_{harv}$ is pre-trained from scratch using a custom vocabulary. Legal-BERT$_{harv}$ is further pre-trained with the same Harvard Law case corpus. Based on the corpus characteristics, they made adjustments to their pre-training objectives by using whole-word MLM and added regular expressions to ensure complete legal citations during the NSP.

2.3 Other Approaches

Apart from the set of embedding models pre-trained from BERT$_{base}$, another typical contextual word embedding model used in legal text processing is ELMo. In addition to the contextual word embedding models, GloVe, the traditional word embedding model, is still considered as a powerful feature extractor on legal texts.

ELMo. Unlike BERT with its deep transformer architecture, ELMo (Embedding from Language Models) [Peters et al., 2018] gains contextual word representations from a bidirectional structure. It consists of a character-based convolutional neural network (CNN) and two bidirectional long short-term memory (BiLSTM) layers.

GloVe. Pennington et al. [2014] developed this unsupervised learning algorithm for obtaining word vector representations, pre-trained and stored as a dictionary-type matrix. The GloVe word vector representations exhibit linear substructures of the word vector space.

When designing legal text analysis models, after word embeddings, extra neural networks are typically applied to further extract the feature representation from embeddings to higher-level encoded vectors. Several legal text processing studies have combined word embedding models with layers such as classic BiLSTM, CNN, and ResNet.

BiLSTM. The long short-term memory (LSTM) is a recurrent neural network (RNN) using memory gates and hidden states to extract and calculate the text features sequentially. BiLSTM (bidirectional LSTM) is a variant LSTM network which summarises information from both directions. BiLSTM also works as a significant module in ELMo embeddings for generating contextual vector representations from input texts. The legal text processing experiment presented in [Zheng et al., 2021] used BiLSTM network as the baseline model.

CNN. The convolutional neural network (CNN) is one of the basic machine learning models, which utilises convolutional filters to extract the text features. This is a basic model in legal text processing. [Zhong et al., 2019] developed a pipeline for legal decision summarisation with a CNN classifier. Similarly, [Xu et al., 2021a] applied a CNN model for legal text classification, and also achieved good results through connecting CNN and BERT embeddings.

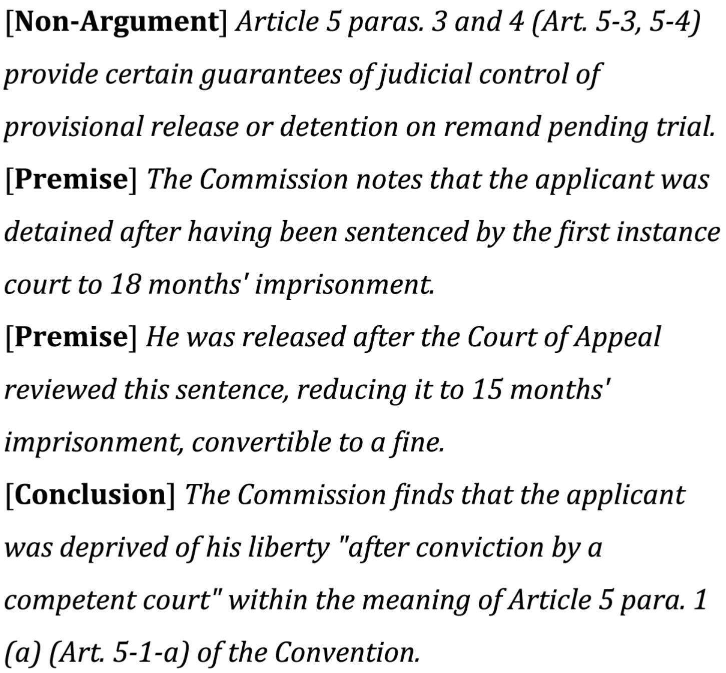
ResNet. The residual network [He et al., 2016] is a classic architecture with special shortcuts that connect neurons belonging to distant layers, which is different from the traditional feed-forward networks. The shortcut design provides ResNet with high efficiency when calculating deep

[1] https://hudoc.echr.coe.int/eng
[2] https://case.law/

networks with multiple layers. The study of argumentative link prediction in [Galassi et al., 2018] presented a combination model of ResNet and GloVe embeddings, which performed well.

III ECHR DATASET

For our experiments, we choose the argument mining corpus annotated on ECHR case-laws from the HUDOC database: an open-source database that has been commonly used for legal AI studies such as court decision event extraction [Filtz et al., 2020], judicial decision prediction [Chalkidis et al., 2019, Medvedeva et al., 2020], and legal argument mining [Mochales and Moens, 2011, Teruel et al., 2018, Poudyal et al., 2020]. Since the early stage of argument mining research [Mochales and Moens, 2011], ECHR case law has been used as a practical application scenario. Detailed information of the argumentation structure in ECHR case law has been provided in [Mochales and Moens, 2008]. In our experiment, we used the ECHR argument mining corpus (ECHR-AM), annotated and open-sourced in [Poudyal et al., 2020][3].



[**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

The ECHR-AM corpus contains text from 42 cases (20 decisions and 22 judgements) including approximately 290,000 words. The ECHR-AM annotation scheme is designed based on Walton's premise/conclusion model. The legal text is first segmented into sentence-level clauses then further annotated into three groups: premises, conclusions, and non-argument clauses. The premises and conclusions are argument components defined by the argumentation model. The amount of arguments is quite unbalanced between documents, ranging from a minimum of 4 arguments (8 premises and 4 conclusions) to 50 arguments (147 premises and 50 conclusions). As a result, we use document-level train-test splitting during our experiment, to verify the model's performance in practical situations.

IV EXPERIMENTS

The aim of our experiments is to explore which language models, word embeddings and machine learning techniques are most suited to the task of legal argument mining. The following sections

[3] http://www.di.uevora.pt/~pq/echr/

describe how the legal argument mining task is structured, the system architecture used for the experiments, and other specific details of the experiment setup.

4.1 Legal Argument Mining Tasks

Our implementation of the legal argument mining task primarily follows the previous experiments in [Poudyal et al., 2020], where the entire argument mining process is separated into three distinct tasks. To align with this approach, we have structured our work as a pipeline that consists of: argument clause recognition, argument relation mining, and argument component classification (including the classification of both premises and conclusions).

4.1.1 Argument Clause Recognition

In the first task of our argument mining pipeline, we aim to shrink the range of argument information in case documents by filtering out non-argument clauses. We formalise this task into a binary classification problem, and train the model to detect argument clauses, which are clauses containing argumentation information (either a premise or a conclusion).

4.1.2 Argument Relation Mining

Since the annotation scheme of the ECHR-AM corpus only includes the inner relations within arguments, we focus on identifying the support relations from the set of premises to the conclusion, which have been defined as the inference in Walton model (see Section 2.1). This task of relation mining is considered as the bottleneck not only with legal texts, but in the whole research field of argument mining [Mochales and Moens, 2011, Poudyal et al., 2019, 2020]. The goal of this task is to group clauses into different arguments before further identifying their labels or functions as specific argument components.

Due to the fact that an argument clause may belong to multiple arguments, we manage this task with the same implementation method as in [Poudyal et al., 2020], where they modelled this as a clause-pair relation prediction task. A classifier model is used to predict whether a pair of input clauses (already identified as argumentative from the previous step) are from the same argument or not. In particular, we assume all the argument clauses have been successfully detected from previous task. We first order the argument clauses within the same document into a sequence, then use a fixed-size sliding window (size = 5) to generate the clause-pair inputs, which are further classified into related or non-related groups.

4.1.3 Argument Component Classification

Following the previous design of argument component classification [Mochales and Moens, 2011, Poudyal et al., 2020], we treat this task as two separate text classification problems. Since the practical situation that an argument clause may act as premise in one argument and also be the conclusion in another, we apply two individual classifiers for premise recognition and conclusion recognition respectively.

4.2 Method Architecture

The general classifier in our experiment contains two parts: an embedding module for handling the text input, and an encoder module for compressing the feature vectors into categories output. By adapting different models into our classifier structure, this work employs a series of combinations which differ as to the way in which the text is embedded (BERT, ELMo, GloVe), and how the embedding vectors are encoded as output (BiLSTM, CNN, ResNet). Since our experiment design initially compares the performance of various embedding models, we choose the classification results from embeddings as the baselines in our research.

We include four domain-specific BERT models selected from two groups (see Section 2.2.2). Besides the transformers, we use ELMo as another contextual embedding model. The final ELMo embedding is obtained by three mixed representation layers from the pre-trained 5.5B ELMo model. To compare with contextual embedding models, we refer to the previous study [Galassi et al., 2021] of using GloVe word embeddings. We use the GloVe pre-trained vocabulary (840B) and turn input clauses into 300-dimensional embeddings. For each baseline model, we generate embeddings from the input segmented clauses, and add a simple classifier head containing a dropout layer (dropout rate = 0.1) and a liner layer (for the final output).

Apart from the baselines provided by both contextual and traditional word embeddings, we study the models' classification performance with extra encoder modules. First, we apply a BiLSTM (100-dimensional) layer to test the enhancement of a recurrent neural network. Second, we implement a four-layer CNN model with 100 convolution filters in each layer and 4 kernel sizes (3, 4, 5, 6). Each convolution layer in the model is activated by a ReLU function and stacked with one max pooling layer. Finally, by implementing the residual bottleneck block, we explore the use of ResNet in our experiment as well. We use a three-layer ResNet, composed by two layers with 64 filters and one layer with 256 filters.

4.3 Experimental Setup

Following the previous setup given by Poudyal et al. [2020], we use 80% of the legal documents (34 documents) for training and the remaining 20% (8 documents) for testing. We use k-fold (k = 5) cross validation for fine-tuning and selecting the model for testing, where each fold uses 20% of the training set for validation purposes. Before input into the embedding model, we pad the pre-processed token sequences with the same length (250 for single clause inputs, 500 for clause pair inputs, heuristically). Cross-entropy loss function and Adam optimiser (initial learning rate = 2e-5) are used for optimising the model. An early-stopping strategy (patience = 5) is used to stop the training procedure when the validation loss has not been decreased. For pre-trained embedding models with deep network layers, the tuning processes are shorter than randomly initialised networks [Devlin et al., 2019, Liu et al., 2019]. Considering the experimental setup in previous studies [Chalkidis et al., 2020, Bonifacio et al., 2020, Wang and Lillis, 2020], we train the BERT-base models with maximum 10 epochs, ELMo-base models with maximum 20 epochs, and GloVe-base models with maximum 150 epochs. In our experiment, we find most models stop early before reaching the maximum number of epochs. Due to the unbalanced distribution of the ECHR-AM dataset, we select the weighted scikit-learn evaluation metrics of precision, recall and F1 measure in our experiments, which account for label imbalance during calculation [Pedregosa et al., 2011]. We perform five runs for each model and report the mean evaluation scores.

V RESULTS AND DISCUSSION

This section presents the results[4] of our experiments for each of the three phases of the argument mining pipeline.

5.1 Argument Clause Recognition Results

Previous studies of the argumentation structures in ECHR case-law suggest that the majority of the argument information is concentrated within specific sections (e.g., "AS TO THE LAW/THE LAW" sections) [Mochales and Moens, 2008, Poudyal et al., 2020]. Besides, to align with previous experiments in [Poudyal et al., 2020], we employ two search scopes for argument

[4] https://github.com/LegalAM/JDMDH-2022-ECHR

clauses: a small one without contents before the “AS TO THE LAW/THE LAW” section; and the complete one that uses the entire document contents. In Table 1, the first three columns indicate how well the model performs on the specific-section search scope, and the rest are results of argument clause recognition on full-text ECHR cases.

Table 1: Weighted precision (P), recall (R), F1 measurement for the argument clause recognition task on the specific section scope and the whole document scope (sec = specific section, full = full text, C = Custom, underlines for baseline model scores, bold texts for best scores in each sub-task experiment, stars for best scores in model combination groups).

Model Combination	P-sec	R-sec	F1-sec	P-full	R-full	F1-full
GloVe	.519	.614	.488	.777	.789	.747
GloVe+bilstm	.423	.556	.424	.752	.770	.740
GloVe+cnn	.803*	.787*	.778*	**.910***	**.911***	**.908***
GloVe+resnet	.747	.738	.723	.831	.839	.826
ELMo	.748	.727	.698	.793	.802	.787
ELMo+bilstm	.710	.659	.604	.773	.784	.767
ELMo+cnn	.784*	.764*	.750	.855	.856	.849
ELMo+resnet	.772	.762	.752*	.860*	.864*	.856*
Legal-BERT$_{base}$	.788	.779	.771	.876	.885	.877
Legal-BERT$_{base}$+bilstm	.801*	.800*	.794*	.874	.871	.871
Legal-BERT$_{base}$+cnn	.796	.785	.776	.893*	.891*	.891*
Legal-BERT$_{base}$+resnet	.730	.736	.723	.868	.875	.865
Legal-BERT$_{echr}$	.814*	.806*	.800	.905*	.902*	.902*
Legal-BERT$_{echr}$+bilstm	.803	.793	.788	.895	.891	.891
Legal-BERT$_{echr}$+cnn	.776	.761	.750	.899	.899	.898
Legal-BERT$_{echr}$+resnet	.807	.804	.803*	.877	.875	.874
C-Legal-BERT$_{harv}$	.795	.792	.787	.860	.861	.858
C-Legal-BERT$_{harv}$+bilstm	.801	.798	.791	.889*	.886*	.887*
C-Legal-BERT$_{harv}$+cnn	**.825***	**.819***	**.817***	.873	.873	.872
C-Legal-BERT$_{harv}$+resnet	.781	.776	.774	.866	.865	.864
Legal-BERT$_{harv}$	.780	.778	.769	.862	.864	.860
Legal-BERT$_{harv}$+bilstm	.793	.788	.780	.876	.876	.875*
Legal-BERT$_{harv}$+cnn	.814*	.800*	.792*	.878*	.876*	.875*
Legal-BERT$_{harv}$+resnet	.770	.769	.763	.868	.870	.866

Here we start by presenting and analysing the results of argument clause recognition in specific sections. For the non-BERT embedding baselines given by GloVe and ELMo, the weighted F1 measure from ELMo is much higher (0.698 vs. 0.488). The additional CNN encoder, on the other hand, substantially improved the GloVe-based model’s performance. When applied together, GloVe+cnn reached outstanding evaluation scores (precision = 0.803, recall = 0.787, and F1 = 0.778), which are the highest among all the GloVe-based models. CNN also enhanced the performance of ELMo embeddings, by increasing the F1 score (0.698 vs. 0.750). The ELMo+cnn combination also obtained the ELMo group-wise greatest precision (0.784) and recall (0.764). Meanwhile, the ResNet encoder greatly increased both non-BERT models’ performance of identifying argumentative information, where GloVe+resnet reached better evaluation scores than its baseline (precision: 0.747 vs. 0.519, recall: 0.738 vs. 0.614, and F1: 0.723 vs. 0.488), and ELMo+resnet gained its group-wise maximum F1 value (0.752). It is obvious that multi-layered CNN and ResNet have the good ability to compress text feature vectors from embeddings. Besides, connecting GloVe embeddings with deeper networks like

CNN can substantially improve the classifier's ability.

The results of argument clause recognition generally aligned to previous research of applying domain-specific pre-trained BERT models on the legal argument mining downstream task [Xu et al., 2021b]. Among all four BERT embedding models, Legal-BERT$_{echr}$ reached the leading F1 score (0.800), compared to other baseline F1 results (Legal-BERT$_{base}$ F1 = 0.771, Legal-BERT$_{harv}$ F1 = 0.769, and Custom Legal-BERT$_{harv}$ F1 = 0.787), along with remarkable precision (0.814) and recall (0.806). When applied with the ResNet encoder, the performance of Legal-BERT$_{echr}$+resnet improved only very slightly when compared to its baseline F1 measure (0.803 vs. 0.800). Legal-BERT$_{base}$ was strengthened by adding extra BiLSTM layer as its encoder. All three evaluation scores of Legal-BERT$_{base}$+bilstm improved about 2%. For the two Harvard Legal-BERT variants, when using CNN as the extra encoder module, both models reached their group-wise best performance. All three scores approximately increased 3% in both evaluation tests of Custom Legal-BERT$_{harv}$ and Legal-BERT$_{harv}$ with the CNN encoder. The evaluation scores obtained by Custom Legal-BERT$_{harv}$+cnn are also the highest among all legal BERT models (precision = 0.825, recall = 0.819, and F1 = 0.817). Besides, like Legal-BERT$_{base}$, the BiLSTM layer also made both Harvard Legal-BERT models slightly better.

When enlarge the search scope to entire ECHR documents, the gap between the GloVe and ELMo baselines of weighted F1 shrank (0.747 vs. 0.787), where ELMo yet maintained its higher score. Similar to the previous argument clause recognition task within sections, both GloVe+cnn and ELMo+resnet reached their own group-wise best performance, among which the evaluation results of GloVe+cnn (precision = 0.910, recall = 0.911, and F1 = 0.908) were even slightly better than those of the legal BERT models. Apart from that, ResNet encoders reinforced the performance of both GloVe and ELMo (around 6%). Among all BERT pre-train models, Legal-BERT$_{echr}$ kept the best evaluation results (precision = 0.905, recall = 0.902, and F1 = 0.902). Apart from GloVe and ELMo, CNN remained useful for pre-trained transformers, the weighted F1 scores of Legal-BERT$_{base}$, Legal-BERT$_{harv}$, and Custom Legal-BERT$_{harv}$ all increased. Likewise, adding the extra BiLSTM also improved both Harvard Legal-BERT variants. In general, BERT variants outperform both GloVe and Elmo as embedding models, which shows the strong ability of pre-trained transformers to distinguish argument information form general text in legal documents. Adding extra convolutional and residual networks sightly improved the performance of some legal BERT models, but magnified the performance of GloVe embedding greatly. We suggest that the transformer model (i.e., BERT) already has complex network layers which is why adding the extra encoder layer has less improvement compared to simple embedding layers like GloVe.

5.2 Argument Relation Mining Results

Table 2 shows the results of the argument relation mining subtask. After adapting with deeper networks, the GloVe embedding model has better performance when processing paired long text and predicting the relations of clause pairs. Both GloVe+bilstm and GloVe+resnet reached higher F1 scores compared to their baseline (0.568 vs. 0.506, and 0.574 vs. 0.506). CNN has the greatest influence among all neural network encoders, which increased all three evaluations from the GloVe baseline remarkably (precision: 0.732 vs. 0.562, recall: 0.729 vs. 0.631, and F1: 0.703 vs. 0.506). When combined with additional BiLSTM and ResNet encoders, ELMo's performance was undermined, the F1 score decreased. Yet, ELMo+cnn continued its improved performance and reached the group-wise best F1 score (0.722). On the other hand, when applying different BERT models on the argument relation mining, Legal-BERT$_{echr}$ maintained its good ability on sentence-pair classification and obtained the best evaluation baseline among all pre-trained transformers (precision = 0.775, recall = 0.772, and F1 = 0.765). Besides, the BiLSTM layer slightly improved the whole LEGAL-BERT family, which increased the F1 score of Legal-BERT$_{base}$ and Legal-BERT$_{echr}$ (0.727 vs. 0.757, 0.765 vs. 0.771). Compared to GloVe and ELMo, the legal BERT models had conducted great results when mining argument relations even without an extra encoder.

Table 2: Weighted precision (P), recall (R), F1 measurement for the argument relation mining task (C = Custom, underlines for baseline model scores, bold texts for best scores in this task experiment, stars for best scores in model combination groups).

Model Combination	**P**	**R**	**F1**
GloVe	.562	.631	.506
GloVe+bilstm	.590	.644	.568
GloVe+cnn	.732*	.729*	.703*
GloVe+resnet	.639	.632	.574
ELMo	.709	.713	.683
ELMo+bilstm	.673	.658	.610
ELMo+cnn	.757*	.740*	.722*
ELMo+resnet	.711	.707	.680
Legal-BERT$_{base}$	.744	.738	.727
Legal-BERT$_{base}$+bilstm	.768*	.770*	.757*
Legal-BERT$_{base}$+cnn	.750	.750	.742
Legal-BERT$_{base}$+resnet	.737	.743	.730
Legal-BERT$_{echr}$	.775	.772	.765
Legal-BERT$_{echr}$+bilstm	**.779***	**.776***	**.771***
Legal-BERT$_{echr}$+cnn	.769	.766	.761
Legal-BERT$_{echr}$+resnet	.758	.757	.750
C-Legal-BERT$_{harv}$	.734	.740	.728
C-Legal-BERT$_{harv}$+bilstm	.731	.730	.720
C-Legal-BERT$_{harv}$+cnn	.764*	.768*	.757*
C-Legal-BERT$_{harv}$+resnet	.724	.738	.723
Legal-BERT$_{harv}$	.762*	.762*	.756*
Legal-BERT$_{harv}$+bilstm	.735	.746	.732
Legal-BERT$_{harv}$+cnn	.740	.741	.735
Legal-BERT$_{harv}$+resnet	.754	.755	.746

5.3 Argument Component Classification Results

As mentioned in Section 4.1, the practical argumentation in legal texts include overlaps between individual arguments where a clause can be a premise in one argument and the conclusion in another. Since each argument clause may require multiple component labels, like [Mochales and Moens, 2011, Poudyal et al., 2020], we maintain the separation between the two individual sub-tasks of classifying premises/non-premises and conclusions/non-conclusions. The results of the two argument component classification sub-tasks are recorded in Table 3. The first three columns display the weighted precision, recall, F1 measure for the classification of conclusions; the next three columns display the same evaluation scores for the classification of premises; the last column of the table is the average F1 score which stands for the average of class-wise (premise/conclusion) F1 scores.

When identifying conclusions, a similar pattern to the previous tasks emerges in terms of the performance of the two non-BERT embedding baselines. Again, the weighted F1 score of

Table 3: Weighted precision (P), recall (R), F1 measurement for the argument component (premise/conclusion) classification task (con = conclusion, pre = premise, C = Custom, underlines for baseline model scores, bold texts for best scores in each sub-task experiment, stars for best scores in model combination groups).

Model Combination	P-con	R-con	F1-con	P-pre	R-pre	F1-pre	avg-F1
GloVe	.514	.717	.598	.518	.720	.603	.601
GloVe+bilstm	.527	.726	.610	.511	.714	.595	.603
GloVe+cnn	.817*	.814*	.790*	.824*	.820*	.800*	.795*
GloVe+resnet	.774	.775	.727	.748	.757	.702	.714
ELMo	.770	.774	.738	.782	.788	.763	.750
ELMo+bilstm	.749	.766	.716	.761	.768	.709	.713
ELMo+cnn	.799*	.806	.791*	.833*	.832*	.818*	.805*
ELMo+resnet	.799*	.807*	.788	.808	.814	.797	.792
Legal-BERT$_{base}$	.843	.843	.837	.833	.838*	.829	.833
Legal-BERT$_{base}$+bilstm	.845	.847	.844*	.834*	.836	.832*	.838*
Legal-BERT$_{base}$+cnn	**.846***	**.849***	.842	.831	.833	.825	.834
Legal-BERT$_{base}$+resnet	.826	.830	.817	.827	.834	.821	.819
Legal-BERT$_{echr}$	.840	.841	.837	.828	.831	.825	.831
Legal-BERT$_{echr}$+bilstm	.842*	.843*	.840*	**.860***	**.862***	**.859***	**.850***
Legal-BERT$_{echr}$+cnn	.834	.833	.829	.836	.838	.835	.832
Legal-BERT$_{echr}$+resnet	.832	.836	.829	.831	.834	.828	.828
C-Legal-BERT$_{harv}$	.838*	.840*	.836*	.834	.837	.834	.835*
C-Legal-BERT$_{harv}$+bilstm	.826	.829	.822	.839*	.843*	.838*	.830
C-Legal-BERT$_{harv}$+cnn	.833	.837	.831	.830	.832	.822	.826
C-Legal-BERT$_{harv}$+resnet	.833	.836	.832	.823	.828	.820	.826
Legal-BERT$_{harv}$	.845*	**.849***	**.845***	.848	.850	.847	.846*
Legal-BERT$_{harv}$+bilstm	.830	.833	.824	.835	.838	.833	.829
Legal-BERT$_{harv}$+cnn	.829	.834	.827	.852*	.855*	.849*	.838
Legal-BERT$_{harv}$+resnet	.830	.835	.824	.849	.852	.847	.835

ELMo greatly outperforms the F1 score of GloVe (0.738 vs. 0.598). By adding the extra neural networks, the performance of the GloVe-base model was greatly improved, and the gap between these two models' evaluation scores was reduced. For the weighted F1 score, GloVe+cnn and ELMo+cnn were almost equal (0.790 vs. 0.791), each of which is also the best F1 in its group. The combination of GloVe+cnn also gained precision (0.817) and recall (0.814) scores which respectively exceed the best ELMo-based scores (from ELMo+resnet, precision = 0.799, recall = 0.807). When using BERT-base transformers, the general classification results of conclusion are better than GloVe and ELMo embeddings, as expected. Legal-BERT$_{harv}$ outperforms all the other pre-trained BERT variants with the top F1 measure (0.845). It is interesting that adding extra layers did not improve the ability of the Harvard Legal BERT models to identify conclusions. We suggest this is likely caused by the limited amount of positive data in the test set. On the other hand, BiLSTM improved the classification results of conclusions for both pre-trained models from the Legal-BERT family. Both Legal-BERT$_{base}$+bilstm and Legal-BERT$_{echr}$+bilstm obtained their group-wise highest F1 scores (0.844 and 0.840), in which the F1 of Legal-BERT$_{base}$ is almost the same as the best (0.845 from Legal-BERT$_{harv}$) within the conclusion classification task.

The performance of both GloVe and ELMo embeddings on the identification of premises are aligned to their results on the classification of conclusions, but generally better. The weighted

F1 score of GloVe increased (0.603 vs 0.598) when switching the classification target from conclusion to premise. Similarly, ELMo's performance slightly increased when identifying premises (0.763 vs 0.738). CNN and ResNet maintained their ability to enhance the performance of both non-BERT embedding models. For the CNN encoder, GloVe+cnn upgraded the evaluation F1 score almost 20% from the GloVe embedding baseline (0.800 vs. 0.603). Elmo+cnn also reached its group-wise best scores (precision = 0.833, recall = 0.832, and F1 = 0.818). For the ResNet encoder, the improvement for ELMo was not as significant as for GloVe, but still helped ELMo+resnet earned better results compared to the baseline (precision: 0.808 vs. 0.782, recall: 0.814 vs. 0.788, and F1: 0.797 vs. 0.763). The average performance of legal BERT models maintained higher scores than non-BERT embeddings. Legal-BERT$_{harv}$, which has the top F1 score (0.845) when identifying conclusions, reached the highest baseline F1 (0.847) again when detecting premises. By using the BiLSTM encoder, both models from the LEGAL-BERT family were augmented. The evaluation scores given by Legal-BERT$_{echr}$+bilstm are the highest in the premise classification task (precision = 0.860, recall = 0.862, and F1 = 0.859). Legal-BERT$_{base}$+bilstm also reached its group-wise best performance in both sub-tasks of argument component classification. It is noteworthy that the BiLSTM network raised the performance of Legal-BERT$_{echr}$ from the lowest BERT embedding baseline to the best classifier in the premise classification sub-task (0.825 vs 0.850). We use an extra evaluation score, which is the average value of both weighted F1 scores from the two sub-tasks (premise/conclusion) of argument component classification. When evaluating all the models using average F1, the CNN encoder again demonstrated its ability to enhance the performance of both GloVe and ELMo embeddings (0.795 vs 0.601, 0.805 vs 0.750). Although the Harvard BERT variants did not reach better performance with an additional encoder, the LEGAL-BERT family were enhanced with extra layer of BiLSTM, and Legal-BERT$_{echr}$+bilstm achieved the best average weighted F1 (0.850) among all transformer-based models.

VI CONCLUSION AND FUTURE WORK

In this paper, we selected multiple legal BERT variants as well as other classic pre-trained embedding models, and provided a broad study of word embedding models' performance on legal argument mining. Our study currently focuses on practical case law from the European Court of Human Rights (ECHR). During our experiments on word embedding models, we also adapted a number of classic NLP neural networks. A comprehensive evaluation of these combinations in the context of argument mining had not been conducted to date, to our knowledge. Thus, our experiments help to contribute to the set of baselines available to researchers going forward, as well as showing how these neural networks can enhance the performance of state-of-the-art transformer-base argument mining models.

Overall, legal BERT embeddings have better performance than ELMo and GloVe in most of the argument mining tasks. The strong performance of BERT models in the argument relation mining task suggests their great generalisability when handling long inputs of clause pairs. When applied with a BiLSTM network, the two models from the LEGAL-BERT family were enhanced for mining argument relations, which implies BiLSTM's improvement when processing long sequential inputs. Domain pre-training also improves the text classification performance when BERT is applied on a small dataset with similar sentence contents (e.g., argument component classification tasks). The Legal-BERT$_{echr}$ model is pre-trained with a limited domain-specific corpus, while exhibiting outstanding performance in both the relation prediction and component classification tasks. It indicates that the characteristics of legal language are quite distinct from that of general English texts, and also leads us to a conclusion that domain-specific pre-training can work effectively on this type of interdisciplinary downstream tasks, with a special language context. It is logical to believe that this approach will be beneficial in other digital humanities applications besides the legal domain also. Besides, from our experiments, both convolutional-base networks (CNN and ResNet) have displayed their ability to enhance GloVe and ELMo embedding models. In particular, the combination of GloVe+cnn showcases its classification ability when detecting small groups of argument clauses from the full text of legal documents.

This work still follows the classic pipeline structure for argument mining system design, which inevitable includes the error propagation issue between its serial tasks. This brings difficulties for evaluation as well for its practical application outside the laboratory environment.. To solve this significant problem, several novel methods and implementation strategies (e.g., joint learning [Niculae et al., 2017], multi-task learning strategy [Galassi et al., 2021], and graph neural network [Ye and Teufel, 2021]) have made breakthroughs on other argument mining corpora. We argue that those techniques are potential solutions for the error propagation issue in legal argument mining. Moreover, adjusting input token length may improve the embedding models' performance when dealing with long input texts from legal documents [Limsopatham, 2021]. Adapting these methods is part of our future work for updating the current traditional mining process.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72, 2019.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. 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 3615–3620, 2019.

Luiz Henrique Bonifacio, Paulo Arantes Vilela, Gustavo Rocha Lobato, and Eraldo Rezende Fernandes. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 648–662. Springer, 2020.

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

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

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.

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

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

Emad Elwany, Dave Moore, and Gaurav Oberoi. 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*, 2019.

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

Andrea Galassi, Marco Lippi, and Paolo Torroni. Argumentative link prediction using residual networks and multi-objective learning. *EMNLP 2018*, page 1, 2018.

Andrea Galassi, Marco Lippi, and Paolo Torroni. Multi-task attentive residual networks for argument mining. *arXiv preprint arXiv:2102.12227*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240, 2020.

Nut Limsopatham. Effectively leveraging bert for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 210–216, 2021.

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

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

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

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

Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. naacl hlt 2018-2018 conference of the north american chapter of the association for computational linguistics: Human language technologies-proceedings of the conference, 2018.

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

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

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1054. URL https://aclanthology.org/P19-1054.

Raquel Silveira, CG Fernandes, João A Monteiro Neto, Vasco Furtado, and José Ernesto Pimentel Filho. Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN*, 1613:0073, 2021.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *LREC 2018-11th International Conference on Language Resources and Evaluation*, pages 1–4, 2018.

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

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

Congcong Wang and David Lillis. 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, December 2020. ISBN 978-1-4503-7760-7. doi: 10.1145/3443279.3443304.

Huihui Xu, Jaromir Savelka, and Kevin Ashley. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *ICAIL'21: Proceedings of the Eighteenth International Conference on Artificial Intelligence and LawJune 2021*, 2021a.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In *Legal Knowledge and Information Systems*, pages 33–42. IOS Press, 2021b.

Yuxiao Ye and Simone Teufel. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages

669–678, 2021.

Gechuan Zhang, David Lillis, and Paul Nulty. Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, pages 121–130. Association for Computational Linguistics, 2021.

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

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 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, 2020a.

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

Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D Ashley, and Matthias Grabmair. 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, 2019.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.