Simulating the Court: Legal Judgment Prediction through Relational Learning

Stanford CS224N Custom Project
Mentor:
(Optional) Sharing project: No

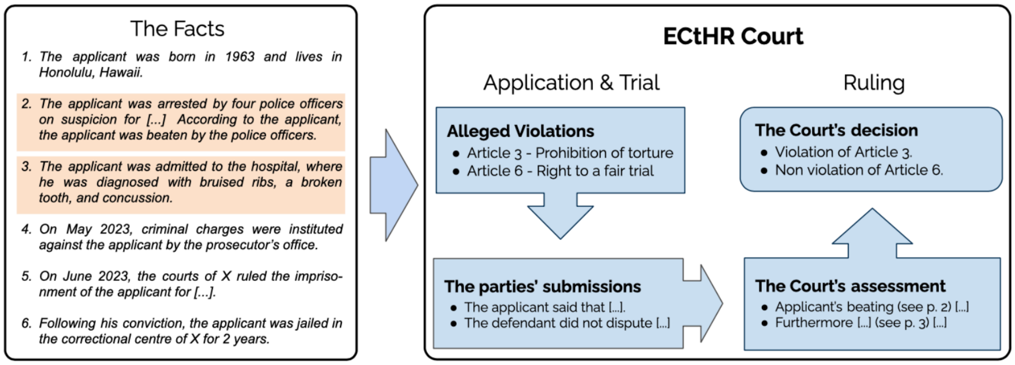
Ein Jun
Department of Computer Science
Stanford University
ejun26@stanford.edu

Abstract

While after the publishing of Chalkidis et al. (2019), there has been a significant development in models which address legal applications, there has been evident stagnation in the resolving of particular challenges in modeling problems in the legal domain which have not been fully explored or resolved. Specifically, in Chalkidis et al. (2019), it was mentioned that while neural methods outperform feature-based models, such models provide no justification for their predictions, or "process" of reaching a conclusion, which is as significant as the prediction itself within the legal domain. Furthermore, following studies employing neural methods have also been limited to elucidating which parts of the text affect predictions the most based upon the attention scores. Therefore, this paper implements a hybrid approach, which both takes advantage of the representational capabilities of transformer models and the reasoning capabilities of human judges by introducing a model which closely mimics the actual court judgment process. This method of procedural learning not only supplements the lack of explanation that current model results suffer from, but also integrates understanding of semantic and logical relations within text to existing models, stepping beyond the currently existing methods in the discourse, and achieves state-of-the-art performance.

1 Background

Introduction and Related Work As previously mentioned, the papers which address the specific problem of legal judgement prediction generally do not directly address the task of reasoning or exploring the semantic and logical relations between the article and the facts corpus, despite the object being itself determining whether the case is a violation of a certain article, given the facts corpus which describes the case. These naturally give birth to challenges in further improving the performance for existing models; specifically, in Chalkidis et al. (2019), it was mentioned that while neural methods outperform feature-based models, they provide no justification for their predictions, and following studies employing neural methods have also been limited to elucidating which parts of the text affect predictions the most based upon the attention scores. Applications of transformers or attention-based frameworks in LJP have been only employed to survey the relations between cases, modeling the cases or law articles as nodes of a graph, and predicting their relations through edges ((Huang et al., 2023), (Khatri et al., 2023))- which suffers from the bottleneck problem, given that these implementations rely on chunking off text which exceeds the maximum number of tokens or undermining the dependencies or semantic relations between different parts of a corpus. In this paper, I present a model which diverges from the common approach, which formulates the problem as a multi-label classification problem, independent of the content of the articles, and instead attempts to closely simulate the actual legal judgment process. Given the importance and contingency that patterns of hierarchical relations within the text hold for legal documents and the scarcity of models



The Facts

1. The applicant was born in 1963 and lives in Honolulu, Hawaii.
2. The applicant was arrested by four police officers on suspicion for [...] According to the applicant, the applicant was beaten by the police officers.
3. The applicant was admitted to the hospital, where he was diagnosed with bruised ribs, a broken tooth, and concussion.
4. On May 2023, criminal charges were instituted against the applicant by the prosecutor's office.
5. On June 2023, the courts of X ruled the imprisonment of the applicant for [...].
6. Following his conviction, the applicant was jailed in the correctional centre of X for 2 years.

ECtHR Court
Application & Trial
Ruling
Alleged Violations
• Article 3 - Prohibition of torture
• Article 6 - Right to a fair trial
The Court's decision
• Violation of Article 3.
• Non violation of Article 6.
The parties' submissions
• The applicant said that [...].
• The defendant did not dispute [...]
The Court's assessment
• Applicant's beating (see p. 2) [...]
• Furthermore [...] (see p. 3) [...]

Figure 1: ECHR process

which deviate from the formulation as a multi-label classification objective, applying this method of procedural learning to the LJP task is a promising endeavor.

Approach The design of the model implemented in this project is motivated by the process of a verdict being determined on a case within the European Court of Human Rights (ECtHR), described in (Chalkidis et al., 2021). Unlike the judges in a court case, the computational formulation of the LJP problem entails that the model *also* performs the actions that an applicant of a trial would; that is, the judge of a court case is met with the task of surveying a given case and a article, as well as past precedents, regarding whether a case would be a violation/non-violation of the 'alleged violations' submitted by the applicant. However, the formulation of the problem as pertains to this model is not as such; the model must i) survey the case to determine which articles are *relevant* to a given court case; ii) survey the case again to find key evidence which determines that there is indeed a violation of the article(s) submitted by the applicant as alleged violation(s). Thus, the model is designed as such:

1.0.1 Relevance Scoring Module

The Relevance Scoring Module (RCM) is designed to determine the relevance between of each article (there are 29 possible articles) and each segment of text within the facts corpus. The facts corpus is originally split into numbered paragraphs, however, given the token limit of the BERT model, paragraphs which exceeded this limit were further split into segments to fit this limitation. The relevance scoring model adapts the design of models used in extractive text summarization, which given a sentence, gauges the semantic significance/relevance of the sentence with respect to the document by utilizing their respective embeddings of a pretrained transformer model, and feeding their concatenated representation into a dense layer with sigmoid activation, which outputs the probability that the sentence is relevant to the document. In this case, we model the relations of each article a_i for $i \in 1, ..., 29$ and each segment of a corpus in the sample n_j for j = fact corpus number, n = segment number, with the corresponding label set as $y_1 * y_2$, where $y_1[a_i, j] \in 0, 1$ = 1 if article a_i is determined as violated for facts corpus j, 0 otherwise; and $y_2[j, n_j] \in 0, 1$ = 1 if segment n_j is in the set of "Silver allegation rationales" for case n, and the cross entropy loss across each sample (a_i, n_j) corresponding to the pairs of each segment in each facts corpus and each article, is the first part of the loss function, L_1.

1.0.2 Judgment Prediction Module

The judgment prediction module is split into mainly two parts: i) the multi-headed self attention across all segments n_i for a given facts corpus n; and ii) the multi-headed cross-attention between each BERT-embedded article text that is determined as an "alleged violation" (an article is an "alleged violation" if the sum of the binarized relevance score $a_i \geq 1$). Feeding this final layer ii) into a dense layer with a sigmoid activation, with outputs a value corresponding to the judgment prediction on an allegedly violated article. Here, as we are no longer dealing with token embeddings, but instead sequence-level embeddings, both positional embeddings as well as the module for multi-headed self attention and multi-headed cross attention were implemented from scratch (referencing work done on

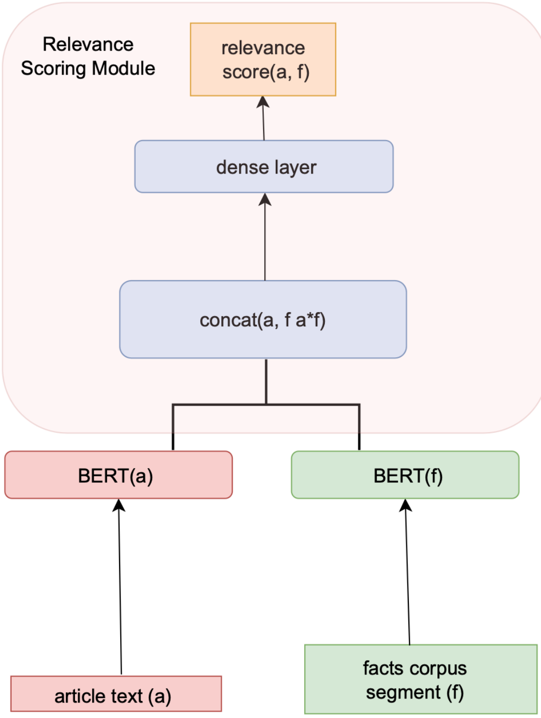


Figure 2: Relevance Scoring Module

Assignment 4), with the corresponding equations: For the sequence positional embeddings, I used segment positional embedding $\Phi \in \mathbb{R}^{T \times d}$:

$$\Phi_{(t,2i)} = \sin\left(t/10000^{2i/d}\right)$$

$$\Phi_{(t,2i+1)} = \cos\left(t/10000^{2i/d}\right)$$

for $t \in \{0, 1, \ldots T-1\}$ and $i \in \{0, 1, \ldots d/2 - 1\}$, for input segments $\mathbf{X} \in \mathbb{R}^{T \times d}$, T the number of sequences and d the dimension of each sequence embedding (they are all padded to be the max BERT embedding length).

Again, treating each facts corpus and article pair separately as a sample pair, the binary cross-entropy loss was calculated across all pairs, constituting the second part of the loss equation, L2, with the total loss/objective function set as L1 + L2.

2 Experiments

Data. The data that is be used is the ECTHR dataset (Chalkidis et al., 2021), which is a set of 11.5k cases from the ECHR's public database. This is an augmented version of the ECHR dataset originally presented in 2019 (Chalkidis et al., 2019). In the original dataset, there is a list of facts that have been extracted from the case description, which are mapped to article(s) of the Convention that have been violated, in case of any. The training and development sets are balanced (contain equal number of cases with and without violations), and the test set contains 66% more cases with violations (which is the approximate rate of cases with violations in the database). The augmentation added the following labels to the original dataset:

Allegedly violated articles: The articles to be discussed (and ruled on) put forward (as alleged article violations) by the applicant

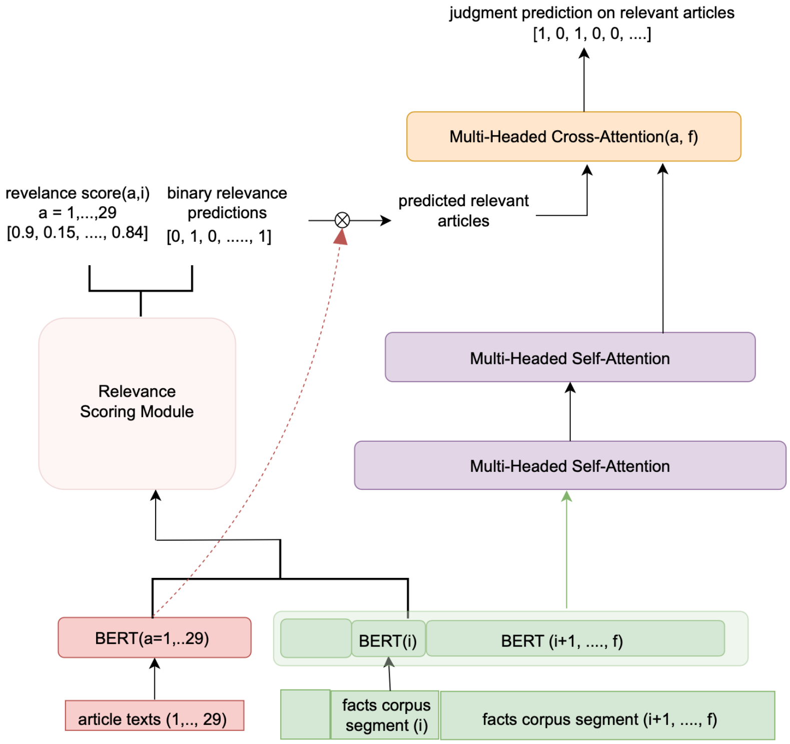


Figure 3: Full Model

Silver allegation rationales: the passage numbers of the passages of the facts corpus that are directly cited/referenced in the judgment as evidence of violation

Gold allegation rationales: A legal expert annotation of facts passage numbers relevant to the ruling.

Of these labels, the allegedly violated articles and silver allegation rationales are used in the RCM layer, and the violated articles (in the original dataset also) are used in the final judgment prediction module, as true labels.

Evaluation Method As the shortcoming in performance for existing models is in the multi-label classification task, evaluation is performed on the multi-label classification task.

In line with the evaluation method in Chalkidis et al. (2019), in which the dataset was presented, evaluation is performed on the micro-averaged precision (P), recall (P) and F1 for the multi-label classification task. For the multi-label classification task, LWAN (multi-label violation prediction), BOW-SVM, BIGRU-ATT, HAN, and HIER-BERT (the proposed model) are used in the aforementioned paper, of which HIER-BERT and HAN perform best overall. As of these models, HIER-BERT is the only neural model, yet it suffers from wrongly assigned attention scores due to fact-level attention.

Experimental Details. Firstly, adapting the code and approach presented in the papers by Chalkidis et al. (2019) and Haas and Skreta (2022), I trained the baselines on the first 512 tokens given the word token limit. Furthermore, as the experimental results were consistent with the observation that a high learning rate (of $\alpha = 1\dot{1}0^{-3}$) resulted in a divergence of the training loss (Haas and Skreta, 2022), I also used a learning rate of $\alpha = 2\dot{1}0^{-5}$, and specifically created a baseline with respect to the LEGAL-XLM-RoBERTA$_{LARGE}$, which is a multilingual model pretrained on legal data with the purpose of fine-tuning on downstream tasks.

	P	R	F1
(Haas and Skreta, 2022)			
LEGAL-BERT	64.8	**59.7**	**62.1**
(Chalkidis et al., 2019)			
HAN	65.0	55.5	59.9
HIER-BERT	65.9	55.1	60.0
Jun (2024)			
RoBERTA	63.5	57.0	60.1
XLM-RoBERTA$_{LARGE}$	65.9	43.3	52.2
Jun-BERT$_{CASED}$	**68.5**	57.2	60.2
Jun- ROBERTA$_{BASE}$	67.8	58.7	59.0

Figure 4: Results on Experiments for Multi-Classification

I used the same hyperparameters on my own model, as I also observed that a higher learning rate led to the training loss diverging and therefore these hyperparameters gave the most optimal results; I believe that this would be more likely attributed to the judgment prediction module, which uses multi-headed attention blocks, as the previously mentioned models also did.

Results and Discussion. Using ROBERTA-BASE, our model was able to obtain results that are only 0.1-0.2 percent lower than that of state-of-the-art models that are pretrained on large amounts of legal corpora in recall and F1 scores, and excelled performance in precision using BERT relative to all existing benchmarks. This is likely attributed to both how the model is able to capture the entire sequence of the corpus and its long-term dependencies, unlike other models which truncate the text and utilize additional information such as the text of the articles, and the allegedly violated articles, as well as relevant parts of the facts corpus, in a way that the previous implementations overlooked. However, the resource limitations, which entailed that we were only able to process a single case embedding at once, entailed a highly volatile gradient, which was likely a detractor for performance.

With respect to the baseline models tested, as shown in the results below, the performance of XLM-RoBERTA$_{LARGE}$ was equivalent to that of HIER-BERT in terms of precision. This is reasonable, given that both are pre-trained on large amounts of legal documents, therefore show similar performance. However, they clearly demonstrate performance inferior to LEGAL-BERT in terms of recall and F1 scores.

Conclusion and Future Work. Based on the observations from this research, it was evident that contriving a way to utilize the relations between the different features available, as well as of existing text, was a direct arbiter of performance- yet there was clearly a lack of papers which addressed such issues in the domain relevant to this paper. Consequently, we project that an end-to-end model which also jointly trains on the relations between cases, as "precedents", which is expected to boost performance.

Given that legal judgment prediction is a domain which is yet to be fully explored through natural language processing, given that its particularities are directly inter-related to the current limitations and shortcomings in natural language processing methods, we anticipate the application of models which attempt to address logical reasoning into the domain will be a fruitful endeavor for both realms.

References

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Annual Meeting of the Association for Computational Linguistics*.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases.

Lukas Haas and Michal Skreta. 2022. From roberta to alexa: Automated legal expert arbitrator for neural legal judgment prediction.

Yinya Huang, Lemao Liu, Kun Xu, Meng Fang, Liang Lin, and Xiaodan Liang. 2023. Discourse-aware graph networks for textual logical reasoning.

Mann Khatri, Mirza Yusuf, Yaman Kumar, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2023. Exploring graph neural networks for indian legal judgment prediction.