Semantic Segmentation of Legal Documents via Rhetorical Roles

Vijit Malik[1*] **Rishabh Sanjay**[1*] **Shouvik Kumar Guha**[2]
Angshuman Hazarika[3] **Shubham Nigam**[1] **Arnab Bhattacharya**[1]
Ashutosh Modi[1†]
[1]Indian Institute of Technology Kanpur (IIT-K)
[2]West Bengal National University of Juridical Sciences (WBNUJS)
[3]Indian Institute of Management Ranchi (IIM-R)
{vijitvm21,rishabh.lfs}@gmail.com shouvikkumarguha@nujs.edu
angshuman.hazarika@iimranchi.ac.in sknigam@cse.iitk.ac.in
arnabb@cse.iitk.ac.in ashutoshm@cse.iitk.ac.in

Abstract

Legal documents are unstructured, use legal jargon, and have considerable length, making them difficult to process automatically via conventional text processing techniques. A legal document processing system would benefit substantially if the documents could be segmented into coherent information units. This paper proposes a new corpus of legal documents annotated (with the help of legal experts) with a set of 13 semantically coherent units labels (referred to as Rhetorical Roles), e.g., facts, arguments, statute, issue, precedent, ruling, and ratio. We perform a thorough analysis of the corpus and the annotations. For automatically segmenting the legal documents, we experiment with the task of rhetorical role prediction: given a document, predict the text segments corresponding to various roles. Using the created corpus, we experiment extensively with various deep learning-based baseline models for the task. Further, we develop a multitask learning (MTL) based deep model with document rhetorical role label shift as an auxiliary task for segmenting a legal document. The proposed model shows superior performance over the existing models. We also experiment with model performance in the case of domain transfer and model distillation techniques to see the model performance in limited data conditions.

1 Introduction

The number of legal cases has been growing almost exponentially in populous countries like India. For example, as per the India's National Judicial Data Grid, there are about 41 million cases pending in India (National Judicial Data Grid, 2021). As per some of recent estimates by a retired Supreme Court of India Judge, it will take about 450 years to clear the backlog of cases (Katju, 2019). Technology could come to the rescue in dealing with the backlog, for example, if there were a technology (based on NLP techniques) that could help a legal practitioner to extract relevant information from legal documents then it could make the legal process more streamlined and efficient. However, legal documents are quite different from conventional documents used to train NLP systems (e.g., newspaper texts). Legal documents are typically long (tens of pages) (Malik et al., 2021), unstructured (Skylaki et al., 2021; Leitner et al., 2019), noisy (e.g., grammatical and spelling mistakes due to manual typing in courts) (Malik et al., 2021; Kapoor et al., 2022), and use different lexicon (legal jargon). The use of a specialized lexicon and different semantics of words makes pre-trained neural models (e.g., transformer-based models) ineffective (Chalkidis et al., 2020). The legal domain has several sub-domains (corresponding to different laws, e.g., criminal law, income tax law) within it. Although some of the fundamental legal principles are common, the overlap between different sub-domains is low; hence systems developed on one law (e.g., income tax law) may not directly work for another law (e.g., criminal law), so there is the problem of a domain shift (Bhattacharya et al., 2019; Malik et al., 2021; Kalamkar et al., 2022a; Kapoor et al., 2022).

In this paper, we target legal case proceedings in the form of judgment documents. To aid the processing of long legal documents, we propose a method of segmenting a legal document into coherent information units referred to as *Rhetorical Roles* (Saravanan et al., 2008; Bhattacharya et al., 2019). We propose a corpus of legal documents annotated with Rhetorical Roles (RRs). RRs could be useful for various legal applications. Legal documents are fairly long, and dividing these into rhetor-

[*]Equal Contributions
[†]Corresponding Author

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 153 - 171

ical role units can help summarize documents effectively. In the task of legal judgment prediction, for example, using RRs, one could extract the relevant portions of the case that contributes towards the final decision. RRs could be useful for legal information extraction, e.g., it can help extract cases with similar facts. Similarly, prior cases similar to a given case could be retrieved by comparing different rhetorical role units. In this work, we make the following contributions:

1. We create a new corpus of legal documents annotated with rhetorical role labels. In contrast to previous work (8 RRs) (Bhattacharya et al., 2019), we create a more fine-grained set of 13 RRs. Further, we create the corpus on different legal domains (§3).

2. For automatically segmenting the legal documents, we experiment with the task of rhetorical role prediction: given a document, predict the text segments corresponding to various roles. Using the created corpus, we experiment with various deep text classification and baseline models for the task. We propose new multi-task learning (MTL) based deep model with document level rhetorical role shift as an auxiliary task for segmenting the document into rhetorical role units (§4). The proposed model performs better than the existing models for RR prediction. We further show that our method is robust against domain transfer to other legal sub-domains (§5). We release the corpus, model implementations and experiments code: https://github.com/Exploration-Lab/Rhetorical-Roles

3. Given that annotating legal documents with RR is a tedious process, we perform model distillation experiments with the proposed MTL model and attempt to leverage unlabeled data to enhance the performance (§5). We also show the use-case for RR prediction model.

2 Related Work

Legal text processing has been an active area of research in recent times. A number of datasets, applications, and tasks have been proposed. For example, Argument Mining (Wyner et al., 2010), Information Extraction and Retrieval (Tran et al., 2019), Event Extraction (Lagos et al., 2010), Prior Case Retrieval (Jackson et al., 2003), Summarization (Moens et al., 1999), and Case Prediction (Malik et al., 2021; Chalkidis et al., 2019; Strickson and De La Iglesia, 2020; Kapoor et al., 2022). Recently, there has been a rapid growth in the development of NLP and ML technologies for the Chinese legal system, inter alia, Chen et al. (2019); Hu et al. (2018); Jiang et al. (2018); Yang et al. (2019); Ye et al. (2018). Few works have focused on the creation of annotated corpora and the task of automatic rhetorical role labeling. Venturi (2012) developed a corpus, TEMIS of 504 sentences annotated both syntactically and semantically. The work of Wyner et al. (2013) focuses on the process of annotation and conducting inter-annotator studies. Savelka and Ashley (2018) conducted document segmentation of U.S. court documents using Conditional Random Fields (CRF) with handcrafted features to segment the documents into functional and issue-specific parts. Automatic labeling of rhetorical roles was first conducted in Saravanan et al. (2008), where CRFs were used to label seven rhetorical roles. Nejadgholi et al. (2017) developed a method for identification of factual and non-factual sentences using fastText. The automatic ML approaches and rule-based scripts for rhetorical role identification were compared in Walker et al. (2019). Kalamkar et al. (2022b) create a large corpus of RRs and propose transformer based baseline models for RR prediction. Our work comes close to work by Bhattacharya et al. (2019), where they use the BiLSTM-CRF model with sent2vec features to label rhetorical roles in Indian Supreme Court documents. In contrast, we develop a multi-task learning (MTL) based model for RR prediction that outperforms the system of Bhattacharya et al. (2019).

3 Rhetorical Roles Corpus

Corpus Acquisition: We focus on Indian legal documents in English; however, techniques we develop can be generalized to other legal systems. We consider legal judgments from the Supreme Court of India, High Courts, and Tribunal courts crawled from the website of IndianKanoon (https://indiankanoon.org/). We also scrape Competition Law documents from Indian Tribunal court cases (National Company Law Appellate Tribunal (NCLAT), COMPetition Appellate Tribunal (COMPAT), Competition Commission of India (CCI)). We focus on two domains of the Indian legal system: Competition Law (CL) (also called as Anti-Trust Law in the US and Anti-Monopoly law in China) and Income Tax (IT). CL deals with regulating the conduct of companies,

particularly concerning competition. With the help of legal experts, we narrowed down the cases pertinent to CL and IT from the crawled corpus (also see Ethical Considerations in App. A).

Choice of CL and IT domains: India has a common law system where a decision may not be exactly as per the statutes, but the judiciary may come up with its interpretation and overrule existing precedents. This introduces a bit of subjectivity. One of the biggest problems faced during the task of identifying the rhetorical roles in a judgment is that the element of subjectivity involved in the judicial perception and interpretation of different rhetorical roles, ranging from the factual matrix (i.e., perception about facts, relevant facts and facts in an issue may vary) to the statutory applicability and interpretation to determine the fitness of a particular judicial precedent to the case at hand. In order to overcome this particular obstacle, we focus on specific legal domains (CL and IT) that display a relatively greater degree of consistency and objectivity in terms of judicial reliance on statutory provisions to reach decisions (Taxmann, 2021).

Corpus Statistics: We randomly selected a set of 50 documents each for CL and IT from the set of acquired documents ($\approx$ 1.6k for IT and $\approx$ 0.8k for CL). These 100 documents were annotated with 13 fine-grained RR labels (vs. 8 by Bhattacharya et al. (2019)) by a team of legal experts. Our corpus is double the size of the RR corpus of Bhattacharya et al. (2019). The CL documents have 13,328 sentences (avg. of 266 per document), and IT has a total of 7856 sentences (avg. of 157 per document). Label-wise distribution for IT and CL documents are provided in Appendix B.3. Annotating legal documents with RRs is a tedious as well as challenging task. Nevertheless, this is a growing corpus, and we plan to add more annotated documents. However, given the complexity of annotations, the RR labeling task also points towards looking for model distillation (§5) and zero-shot learning-based methods.

Annotation Setup: The annotation team (legal team) consisted of two law professors from prestigious law schools and six graduate-level law student researchers. Annotating just 100 documents took almost three months. Based on detailed discussions with the legal team, we initially arrived at the eight main rhetorical roles (facts, arguments, statues, dissent, precedent, ruling by lower court, ratio and ruling by present court) plus one 'none' label. During the annotation, roles were further refined, and the documents were finally annotated with 13 fine-grained labels since some of the main roles could be sub-divided into more fine-grained classes. The list of RRs is as follows (example sentences for each role is in Table 15 in the Appendix B.3):

- **Fact (FAC):** These are the facts specific to the case based on which the arguments have been made and judgment has been issued. In addition to Fact, we also have the fine-grained label **Issues (ISS)**. The issues which have been framed/accepted by the present court for adjudication.
- **Argument (ARG)**: The arguments in the case were divided in two more fine-grained sub-labels: **Argument Petitioner (ARG-P):** Arguments which have been put forward by the petitioner/appellant in the case before the present court and by the same party in lower courts (where it may have been petitioner/respondent). Also, **Argument Respondent (ARG-R):** Arguments which have been put forward by the respondent in the case before the present court and by the same party in lower courts (where it may have been petitioner/respondent)
- **Statute (STA):** The laws referred in the case.
- **Dissent (DIS):** Any dissenting opinion expressed by a judge in the present judgment/decision.
- **Precedent (PRE):** The precedents in the documents were divided into 3 finer labels, **Precedent Relied Upon (PRE-R):** The precedents which have been relied upon by the present court for adjudication. These may or may not have been raised by the advocates of the parties and amicus curiae. **Precedent Not Relied Upon (PRE-NR):** The precedents which have not been relied upon by the present court for adjudication. These may have been raised by the advocates of the parties and amicus curiae. **Precedent Overruled (PRE-O):** Any precedents (past cases) on the same issue which have been overruled through the current judgment.
- **Ruling By Lower Court (RLC):** Decisions of the lower courts which dealt with the same case.
- **Ratio Of The Decision (ROD):** The principle which has been established by the current

judgment/decision which can be used in future cases. Does not include the obiter dicta which is based on observations applicable to the specific case only.

- **Ruling By Present Court (RPC):** The decision of the court on the issues which have been framed/accepted by the present court for adjudication.
- **None (NON):** any other matter in the judgment which does not fall in any of the above-mentioned categories.

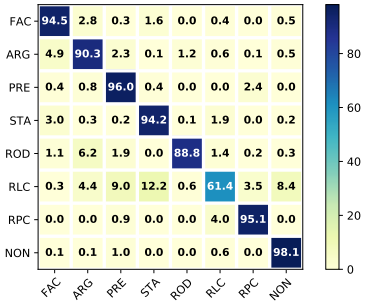
The dataset was annotated by six legal experts (graduate law student researchers), 3 annotated 50 CL documents, and the remaining 3 annotated 50 IT documents. We used Webanno (de Castilho et al., 2016) as the annotation framework. Each legal expert assigned one of the 13 Rhetorical roles to each document sentence. Note that we initially experimented with different levels of granularity (e.g., phrase level, paragraph level), and based on the pilot study, we decided to go for sentence-level annotations as it maintains the balance (from the perspective of topical coherence) between too short (having no labels) and too long (having too many labels) texts. Legal experts pointed out that a single sentence can sometimes represent multiple rhetorical roles (although this is not common). Each expert could also assign secondary and tertiary rhetorical roles to a single sentence to handle such scenarios (also App. B.4). As an example, suppose a sentence is a 'Fact' but could also be an 'Argument' according to the legal expert. In that case, the expert could assign the rhetorical roles 'Primary Fact' and 'Secondary Argument' to that sentence. We extended it to the tertiary level as well to handle rare cases.

Our corpus is different from the existing corpus (Bhattacharya et al., 2019). Firstly, we use 13 fine-grained RR labels and the size of the corpus is almost twice. Secondly, we focus on different legal sub-domains (IT and CL vs. Supreme Court Judgments). Lastly, we perform the primary, secondary, and tertiary levels of annotations since, according to legal experts, it is sometimes possible that a sentence might have multiple RR labels.

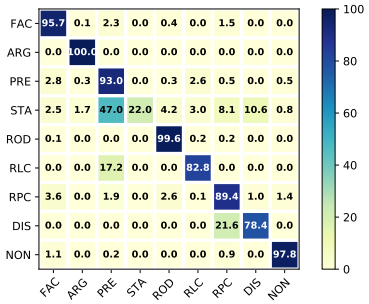
Adjudication and Data compilation: Annotating RR is not a trivial task, and annotators can have disagreements. We followed a majority voting strategy over primary labels to determine the gold labels. There were a few cases ($\approx 5\%$) where all the three legal experts assigned a different role to the same sentence. We asked the law professors to finalize the primary label in such cases. If the law professors decided to go with a label completely different from the three annotated labels, we went with their verdict. However, such cases were not frequent ($\approx$ 4% of adjudicated cases). In this paper, for RR prediction, we concentrate on the primary labels and leave explorations of secondary and tertiary labels for future work.

Label	IT	CL
AR	0.80	0.93
FAC	0.80	0.89
PR	0.70	0.86
STA	0.78	0.89
RLC	0.58	0.74
RPC	0.78	0.79
ROD	0.67	0.93
DIS	_	0.99
Macro F1	0.73	0.88

Table 1: Label-wise Inter-Annotator agreement (F1 Scores). Dissent label instance absent in IT.



(a) IT



(b) CL

Figure 1: Confusion matrix between Annotators A_1 and A_3. Numbers represent % agreement. Dissent label instance is absent in IT.

Inter-annotator Agreements: The Fleiss kappa (Fleiss et al., 2013) between the annotators is 0.65 for the IT domain and 0.87 for the CL domain, indicating a substantial agreement between annotators. Additionally, as done in Bhattacharya et al. (2019) and Malik et al. (2021), we calculate

the pair-wise inter-annotator F1 scores. To determine the agreement between the three annotators A_1, A_2, A_3 (each for IT and CL domain), we calculate the pairwise F1 scores (App. C) between annotators $(A_1, A_2), (A_2, A_3)$ and (A_3, A_1). We average these pairwise scores for each label and further average them out. We report the label-wise F1 and Macro F1 in Table 1. The table shows that the agreements between domains differ (0.73 for IT vs. 0.88 for CL). This is mainly due to (as pointed by law professors) the presence of more precedents and a greater number of statutory provisions in IT laws. These factors combine to produce more subjectivity (relative to CL) when it comes to interpreting and retracing judicial decisions. The confusion matrix between the annotators (A_1, A_3) is shown in Figure 1 (more details in App. B.5).

Analysis: Annotation of judgments to identify RR is a challenging task even for legal experts. Several factors contribute to this challenge. Annotators need to glean and combine information non-trivially (e.g., facts and arguments presented, the implicit setting, and the context under which the events described in the case happened) to arrive at the label. Moreover, the annotator only has access to the current document, which is a secondary account of what actually happened in the court. These limitations certainly make the task of the annotator more difficult and leave them with no choice other than to make certain educated guesses when it comes to understanding the various nuances, both ostensible and probable, of certain RR. It should, however, be noted that such variation need not occur for every RR since not all the roles are equally susceptible to it. A cumulative effect of the aforementioned factors can be observed in the results of the annotation. The analysis provided by the three annotators in the case of CL bears close resemblance with each other. On the other hand, in the case of IT, the analysis provided by Users 1 and 3 bears a greater resemblance with each other, compared to the resemblance between Users 1 and 2, or between Users 2 and 3. On a different note, it is also observed that the rhetorical role where the annotators have differed between themselves the most has been the point of Ruling made by the Lower Court, followed by the Ratio. This also ties in with the argument that all rhetorical roles are not equally susceptible to the variation caused by the varying levels of success achieved by the different annotators in retracing the judicial thought pattern (details and case studies in App. B.6).

Model	Dataset	F1
SBERT-Shift	IT	0.60
SBERT-Shift	CL	0.49
SBERT-Shift	IT+CL	0.47
BERT-SC	IT	0.66
BERT-SC	CL	0.64
BERT-SC	IT+CL	0.64

Table 2: Results for the auxillary task LSP

4 Rhetorical Roles Prediction

We would like to automate the process of segmenting a legal document, to develop ML models for the automation, we experiment with the task of Rhetorical Roles prediction.

Task Definition: Given a legal document, D, containing the sentences $[s_1, s_2, ...s_n]$, the task of rhetorical role prediction is to predict the label (or role) y_i for each sentence $s_i \in D$.

Baseline Models: For the first set of baseline models, the task is modeled as a single sentence prediction task, where given the sentence s, the model predicts the rhetorical role of the sentence. In this case, the context is ignored. We consider pre-trained BERT (Devlin et al., 2019) and LEGAL-BERT (Chalkidis et al., 2020) models for this. As another set of baseline models, we consider the task as a sequence labeling task, where the sequence of all the sentences in the document is given as input, and the model has to predict the RR label for each sentence. We used CRF with hand-crafted features (Bhattacharya et al., 2019) and BiLSTM network.

Label Shift Prediction: Rhetorical role labels do not change abruptly across sentences in a document, and the text tends to maintain topical coherence. Given the label y for a sentence s_i in the document, we hypothesize that the chances of shift (change) in the label for the next sentence s_{i+1} are low. We manually verified this using the training set and observed that on average in a document, if the label of sentence s_i is y, then 88% of the times the label of the next sentence s_{i+1} is same as y. Note that this is true only for consecutive sentences, but in general, label shift inertia fades as we try to predict beyond the second consecutive sentence. Since we are performing a sequence prediction task, this alone is not a good model for label prediction. Nevertheless, we think that this label shift inertia can provide a signal (via an auxiliary task) to the main sequence prediction model. Based on this observation, we define an auxiliary

binary classification task: Label Shift Prediction (LSP), that aims to model the relationship between two sentences s_i and s_{i+1} and predict whether the labels y_i for s_i and y_{i+1} for s_{i+1} are different (shift occurs) or not. In particular, for each sentence pair $S = \{s_i, s_{i+1}\} \in D$, we define the label of LSP task, $Y = 1$ if $y_i \neq y_{i+1}$, otherwise $Y = 0$, here y_i is the rhetorical role for sentence s_i. Note that for the full model at the inference time, the true label of a sentence is not provided; hence predicting a shift in label makes more sense than performing a binary prediction that the next sentence has the same label or not. We model the LSP task via two different models:

SBERT-Shift: We model the label shift via a Siamese network. In particular, we use the pre-trained SBERT model (Reimers and Gurevych, 2019) to encode sentences s_i and s_{i+1} to get representations e_i and e_{i+1}. The combination of these representations $(e_i \oplus e_{i+1} \oplus (e_i - e_{i+1}))$ is passed through a feed-forward network to predict the shift.

BERT-SC: We use the pre-trained BERT model and fine-tune it for the task of LSP. We model the input in the form of sentence semantic coherence task, $[CLS] \oplus s_i \oplus [SEP] \oplus s_{i+1} \oplus [SEP]$ to make the final prediction for shift. In general, the BERT-SC model performs better than SBERT-Shift (Table 2). Due to the superior performance of BERT-SC, we include it to provide label shift information to the final MTL model. The aim of our work is to predict RR, and we use label shift as auxiliary information even if it may not be predicted correctly at all times. As shown in results later, this limited information improves the performance.

Proposed Models: We propose two main models for the rhetorical role prediction: Label Shift Prediction based on BiLSTM-CRF and MTL models.

LSP-BiLSTM-CRF: Signal from label shift is used to aid the RR prediction in the LSP-BiLSTM-CRF model. The model consists of (Figure 2) a BiLSTM-CRF model with specialized input representation. Let the sentence embedding (from pre-trained BERT) corresponding to i^{th} sentence be b_i. Let, the representation of the label shift (the layer before the softmax layer in LSP model) between current sentence and previous sentence pair $\{s_{i-1}, s_i\}$ be $e_{i-1,i}$. Similarly for the next pair $(\{s_i, s_{i+1}\})$ we get $e_{i,i+1}$. The sentence representation for i^{th} sentence is given by $e_{i-1,i} \oplus b_i \oplus e_{i,i+1}$. This sentence representation goes as input to the BiLSTM-CRF model for RR prediction.

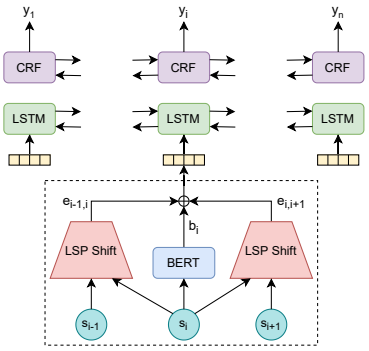


Figure 2: LSP-BiLSTM-CRF Model

MultiTask Learning (MTL): We use the framework of Multitask learning, where rhetorical role prediction is the main task and label shift prediction is the auxiliary task. Sharing representations between the main and related tasks helps in better generalization on the main task (Crawshaw, 2020). The intuition is that a label shift would help the rhetorical role component make the correct prediction based on the prospective shift. The MTL model (Figure 3) consists of two components: the shift detection component and the rhetorical role prediction component. The shift component predicts if a label shift occurs at i^{th} position. The output of the BiLSTM layer of shift component is concatenated with the BiLSTM output of the rhetorical role component. The concatenated output is passed to a CRF layer for the final prediction of the rhetorical role. The loss for the model is given by: $L = \lambda L_{shift} + (1 - \lambda) L_{RR}$, where, L_{shift} is the loss corresponding to label shift prediction and L_{RR} is the loss corresponding to rhetorical role prediction, and hyperparameter λ balances the importance of each of the task. If λ is set to zero, we are back with our baseline BiLSTM-CRF model. Since there are two components, we experimented with sending the same encodings of sentences to both the components $(E_1 = E_2)$, as well as sending different encodings of the same sentence to both components $(E_1 \neq E_2)$. The proposed model is very different from the previously proposed BiLSTM-CRF by Bhattacharya et al. (2019) that does not use any multitasking and label shift information.

5 Experiments, Results and Analysis

Due to the complexity of the task of RR prediction and to be comparable with the existing baseline systems, for experiments, we consider 7 main labels (FAC, ARG, PRE, ROD, RPC, RLC, and STA). We plan to explore all fine-grained RR label (13) pre-

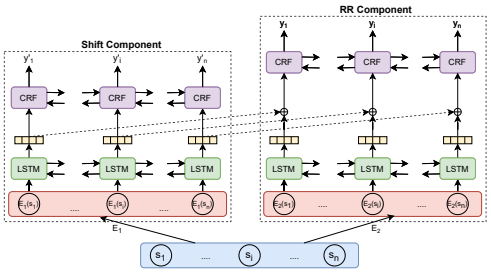


Figure 3: MTL architecture for Rhetorical Role Labelling and Shift Prediction.

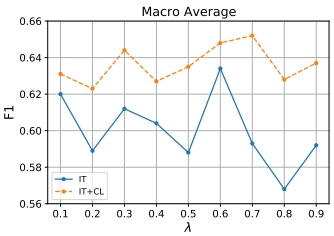


Figure 4: Variation of F1 score with λ on IT and IT+CL domain

Model	IT (F1)	CL (F1)
BERT	0.56	0.52
BERT-neighbor	0.53	0.51
LEGAL-BERT	0.55	0.53
CRF (Handcrafted)	0.55	0.52
BiLSTM (sent2vec)	0.55	0.54
BiLSTM-CRF (handcraft)	0.57	0.56
BiLSTM-CRF (sent2vec)	0.59	0.61
BiLSTM-CRF (BERT emb)	0.63	0.63
BiLSTM-CRF (MLM emb)	0.58	0.60
LSP (SBERT)	*0.64*	*0.63*
LSP (BERT-SC) •	*0.65*	*0.68*
MTL (MLM emb)	*0.67*	*0.67*
MTL (BERT-SC) ⋆ ◇	***0.70***$\pm$*0.02*	***0.69***$\pm$*0.01*

Table 3: Results of baseline and proposed models on IT and CL. LSP and MTL refer to the LSP-BiLSTM-CRF and MTL-BiLSTM-CRF models respectively. • LSP result is significant with $p \leq 0.05$ in comparison to baseline (BiLSTM-CRF(sent2vec)). Similarly, MTL (BERT-SC) has significant result in comparison to baseline ($\diamond$, $p \leq 0.05$). MTL (BERT-SC) is significant w.r.t. LSP ($\star$, $p \leq 0.05$).

dictions in the future. Based on recommendations by legal experts, we ignore sentences with NON (None) label (about 4% for IT and 0.5% for CL) (more details in App. D.1). Further, the IT domain did not have any instance of dissent (DIS) label, and CL has only three documents with very few DIS instances. Based on consultations with law experts, we discarded DIS sentences (more details in App. D.1). We randomly split (at document level) IT/CL into 80% train, 10% validation, and 10% test set. In contrast to Bhattacharya et al. (2019), we did not perform cross-validation for better comparison across different models. We also experiment with a combined dataset of IT and CL (IT+CL); the splits are made by combining individual train/val/test split of IT and CL. We experimented with a number of baseline models (Table 3, 4). In particular, we considered BiLSTM with sent2vec embeddings (Bhattacharya et al., 2019), non-contextual models (single sentence) like BERT (Devlin et al., 2019), LegalBERT (Chalkidis et al., 2020) and BERT-neighbour (we take both left and right neighboring sentences in addition to the sentence of interest). We also considered sentence-level sequence prediction models (contextual models): CRF model using handcrafted features provided by Bhattacharya et al. (2019), different variants of BiLSTM-CRF, one with handcrafted features, with sent2vec embeddings, with BERT embeddings, and with MLM embeddings. We finetuned BERT with Masked Language Modeling (MLM) objective on the train set to obtain MLM embeddings (CLS embedding) for each of the sentences (App. D has hyperparameters, training schedule, and compute settings). We use the Macro F1 metric for evaluation (App. C). We tuned the hyperparameter λ of the MTL loss function using the validation set. We trained the MTL model with $\lambda \in [0.1, 0.9]$ with strides of 0.1 (Figure 4). $\lambda = 0.6$ performs the best for the IT domain and performs competitively on the combined domains.

Results and Analysis: Among the baseline models (Table 3), we note that LEGAL-BERT performs slightly better on the CL domain but slightly worse on the IT domain when compared to pre-trained BERT. It might be attributed to that LEGAL-BERT (trained on EU legal documents, which also has European competition law) is not trained on Indian IT law documents. Using BERT embeddings with BiLSTM-CRF provides better results. Both the proposed approaches outperform the previous approaches by a substantial margin. The MTL approach (with $\lambda = 0.6$) provides the best results on both datasets with an average (over six runs) F1 score of 0.70 (standard deviation of 0.02) on the IT domain, an average F1 of $0.69(\pm 0.01)$ on CL domain, and an average F1 score of $0.71(\pm 0.01)$ for the combined domain. The MTL model shows variance across runs; hence we average the results. Other models were reasonably stable across runs.

We use the LSP shift component with BERT-SC as the encoder E_1 and the pre-trained BERT model as the encoder E_2 in our MTL architecture. We

Model	IT+CL (F1)
BiLSTM-CRF (sent2vec)	0.65
BiLSTM-CRF (BERT embs)	0.63
LSP-BiLSTM-CRF (BERT-SC)	0.67
MTL-BiLSTM-CRF (BERT-SC)	**0.70**±0.01

Table 4: Results of baseline and proposed models on combined dataset (IT+CL)

Label	IT	CL
AR	0.67±0.010	0.78±0.005
FAC	0.78±0.020	0.75±0.010
PR	0.69±0.005	0.62±0.005
STA	0.79±0.020	0.82±0.020
RLC	0.62±0.005	0.53±0.005
RPC	0.70±0.010	0.71±0.010
ROD	0.66±0.005	0.65±0.005
Macro F1	0.70±0.020	0.69±0.010

Table 5: Label-wise average (across 6 runs) F1 scores of MTL-BiLSTM-CRF (BERT-SC) model.

did not use SBERT since it was under-performing when compared to BERT-SC. We provide the label-wise F1 scores for the MTL model in Table 5. Note the high performance on the FAC label and low performance on the RLC label; this is similar to what we observe for annotators (Table 1). Also, the MTL model performs better on the AR label in the CL domain than the IT domain. An opposite trend can be observed for the RLC label. The contribution of the LSP task is evident from the superior performance. We conduct the ablation study of our MTL architecture from multiple aspects. Instead of using shift embeddings from BERT-SC as the encoder E_1, we use a BERT model fine-tuned upon the MLM task on the IT and CL domain. However, we obtain a comparatively lower score (see App. D). This observation yet again points towards the significance of the LSP in the task of rhetorical role prediction (results on other encoders in App. D). The results have two interesting observations: firstly, MTL model performance on IT cases comes close to the average inter-annotator agreement. In the case of CL, there is a gap. Secondly, for the model, the performance on the IT domain is better than the CL domain, but in the case of annotators opposite trend was observed. We do not know the exact reason for this, but the legal experts pointed out that this is possible because the selected documents might be restricted to specific sections of the IT law and model learned solely from these documents alone without any other external knowledge. However, annotators, having knowledge of the entire IT law, might have looked from a broader perspective.

Domain Transfer: In order to check the generalization capabilities of the MTL model compared to the baseline model, we conducted some domain transfer experiments. We experimented with a RR dataset of 50 documents (referred to as G) by Bhattacharya et al. (2019). G dataset comes from a different legal sub-domain (criminal and civil cases) with very less overlap with IT and CL. We tried different combinations of train and test datasets of IT, CL, and G. Note that G (criminal and civil cases) has very less overlap with IT and CL cases, so practically, it is a different domain. The results are in Table 6. We can observe that the MTL model generalizes better across the domains than the baseline model. Both the models perform better on the G_{test} when the combined (IT+CL)$_{\text{train}}$ set is used. This points towards better generalization.

Train Dataset	Test Dataset	BiLSTM-CRF (sent2vec)	MTL
G_{train}	G_{test}	0.55	**0.59**
G_{train}	CL_{test}	0.48 (12.78%)	**0.50** (15.25%)
G_{train}	IT_{test}	0.41 (25.45%)	**0.46** (22.03%)
G_{train}	(IT+CL)$_{\text{test}}$	0.42 (23.64%)	**0.48** (18.64%)
(IT+CL)$_{\text{train}}$	G_{test}	0.60	**0.63**

Table 6: Domain transfer experiments to compare the performance of MTL-BiLSTM-CRF with the baseline BiLSTM-CRF. The number in parenthesis denotes Δ_G : the % difference between the performance on G_{test} and the new domain.

Model distillation: RR annotation is a tedious process, however, there is an abundance of unlabelled legal documents. We experimented with semi-supervised techniques to leverage the unlabelled data. In particular, we tried a self-training based approach (Xie et al., 2020). The idea is to learn a teacher model θ_{tea} on the labelled data D_L. The teacher model is then used to generate hard labels on unlabeled sentences $s_u \in d_i$: $\hat{y}_i = f_{\theta_{tea}}(\hat{d}_i)\ \forall \hat{d}_i \in D_U$. Next, a student model θ_{stu} is learned on labeled and unlabeled sentences, with the loss function for student training given by: $L_{ST} = \frac{1}{|D_L|}\sum_{d_j \in D_L} L(f_{\theta_{stu}}(d_j), y_j) + \frac{\alpha_U}{D_U}\sum_{\hat{d}_i \in D_U} L(f_{\theta_{stu}}(\hat{d}_i), \hat{y}_i)$. Here, α_U is a weighing hyperparameter between the labelled and unlabelled data (details in App. D). The process can be iterated and the final distilled model is used for prediction. The results of model distillation are shown in Table 8 for two iterations (initializing the teacher model of the current iteration as the learned student model of the previous iteration; further iterations do not improve results). MTL model was

run just once, due to variance it shows F1 of 0.68. The results improve for majority of labels with an increment of 0.11 F1 score for the RLC label in the first iteration. Also, the variance of F1 scores across labels decreases.

5.1 Application of Rhetorical Role to Judgment Prediction

To check the applicability of RR in downstream applications, as a use-case, we experimented with how RR could contribute towards judgment prediction (ethical concerns discussed later). We use the legal judgment corpus (ILDC) provided by Malik et al. (2021) and fine-tune a pre-trained BERT model on the train set of ILDC for the task of judgment prediction on the last 512 tokens of the documents. Malik et al. (2021) observed that training on the last 512 (also the max size of the input to BERT) tokens of a legal document give the best results; we use the same setting. We use this trained model directly for predicting the outcome on 84 IT/CL cases. We removed text corresponding to the final decisions (and extracted gold decisions) from these documents with the help of legal experts. In the first experiment, we use the last 512 tokens of IT/CL cases for prediction. To study the effect of RRs, in another experiment, we extract the sentences corresponding to gold ratio (ROD) and ruling (RPC) RR labels in IT/CL documents and use this as input to the BERT model. We consider these two RR only since, by definition, these sentences denote the principles and the decision of the court related to the issues in the proceedings. There were no ROD or RPC labels for some documents (16 out of 100 for both IT and CL); we removed these in both experiments. The results are shown in Table 7. Using the gold RR gives a boost to the F1 score. We also experimented with using predicted RR, and the performance was comparable to that of the BERT model.

To explore how predicted rhetorical roles would perform on judgment prediction task, we perform the following experiment. We use our best performing model MTL (BERT-SC), trained on the combined IT+CL domain to check the applicability of rhetorical roles for the task of Judgment Prediction. In the first step, we obtain the predicted rhetorical roles for each sentence in the documents. Next, we select the sentences labeled as ROD or RPC[1]. Third, we use a BERT base model fine-tuned on the last 512 tokens of each document in the ILDC corpus (Malik et al., 2021) and use it to predict the judgment of the test set documents, given only the predicted ROD and RPC sentences. We compare the results by the MTL model and BiLSTM-CRF baseline on performing judgment prediction with predicted rhetorical roles. Refer to Appendix Table 14 for the results. Since RR prediction for ROD and RPC is not perfect, improving it would greatly enhance the results as shown in Table 7.

[1]We select only these two labels since by definition, these sentences provide the necessary cues towards the judgment.

Model	IT+CL docs	F1
BERT-ILDC	last 512 tokens	0.55
BERT-ILDC	Gold ROD & RPC	**0.58**

Table 7: Judgment prediction using RR. The model using gold ROD and RPC is found to be statistically significant ($p \leq 0.05$).

Label	Base MTL	Dist. Iter 1	Dist. Iter 2
AR	0.62	0.70	**0.70**
FAC	0.74	**0.75**	0.73
PR	0.68	0.72	**0.74**
STA	0.76	**0.77**	0.75
RLC	0.59	**0.70**	0.70
RPC	0.67	0.63	**0.73**
ROD	0.68	0.66	**0.68**
Macro F1	0.68	0.71	**0.72**

Table 8: Model Distillation: F1 scores of MTL-BiLSTM-CRF (BERT-SC) model after two distillation iterations on the IT domain.

6 Conclusion

We introduce a new corpus annotated with rhetorical roles. We proposed a new MTL model that uses label shift information for predicting labels. We further showed via domain transfer experiments the generalizability of the model. Since RR are tedious to annotate, we showed the possibility of using model distillation techniques to improve the system. In the future, we plan to explore cross-domain transfer techniques to perform RR identification in legal documents in other Indian languages. Nevertheless, we plan to grow the corpus. We also plan to apply RR models for other legal tasks such as summarization and information extraction.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments. We would like to thank student research assistants Tridib Mandal, Chirag Mittal, Shefali Deshmukh, Shailja Beria, and Syamantak Sinha from West Bengal National University of Juridical Sciences (WBNUJS) for annotating the documents. This work would not have been possible without their help.

References

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. *CoRR*, abs/1911.05405.

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, Florence, Italy. Association for Computational Linguistics.

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.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. 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 6362–6367, Hong Kong, China. Association for Computational Linguistics.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *CoRR*, abs/2009.09796.

Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. 2003. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150(1-2):239–290.

Xin Jiang, Hai Ye, Zhunchen Luo, WenHan Chao, and Wenjia Ma. 2018. Interpretable rationale augmented charge prediction system. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 146–151, Santa Fe, New Mexico. Association for Computational Linguistics.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022a. Corpus for automatic structuring of legal documents. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022b. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536, Dublin, Ireland. Association for Computational Linguistics.

Justice Markandey Katju. 2019. Backlog of cases crippling judiciary. `https://perma.cc/D8V4-L566`.

Nikolaos Lagos, Frederique Segond, Stefania Castellani, and Jacki O'Neill. 2010. Event extraction for legal case building and reasoning. In *International Conference on Intelligent Information Processing*, pages 92–101. Springer.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham. Springer International Publishing.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. 1999. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50(2):151–161.

National Judicial Data Grid. 2021. National judicial data grid statistics. https://www.njdg.ecourts.gov.in/njdgnew/index.php.

Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In *JURIX*, pages 125–134.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

M Saravanan, Balaraman Ravindran, and S Raman. 2008. Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Jaromir Savelka and Kevin D Ashley. 2018. Segmenting us court decisions into functional and issue specific parts. In *JURIX*, pages 111–120.

Stavroula Skylaki, Ali Oskooei, Omar Bari, Nadja Herger, and Zac Kriegman. 2021. Legal entity extraction using a pointer generator network. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 653–658.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, pages 204–209.

Taxmann. 2021. Interpretation of statutes: Strict versus liberal construction. https://tinyurl.com/2p85h3xd.

Vu Tran, Minh Le Nguyen, and Ken Satoh. 2019. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 275–282.

Giulia Venturi. 2012. Design and development of temis: a syntactically and semantically annotated corpus of italian legislative texts. In *Proceedings of the Workshop on Semantic Processing of Legal Texts (SPLeT 2012)*, pages 1–12.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. In *ASAIL@ ICAIL*.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In *Semantic processing of legal texts*, pages 60–79. Springer.

Adam Z Wyner, Wim Peters, and Daniel Katz. 2013. A case study on legal case annotation. In *JURIX*, pages 165–174.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4085–4091. International Joint Conferences on Artificial Intelligence Organization.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Appendix

A Ethical Considerations

The proposed corpus and methods do not have direct ethical consequences to the best of our knowledge. The corpus is created from publicly available data from a public resource: www.indiankanoon.org. The website allows free downloads, and no copyrights were violated. With the help of law professors, we designed a course project centered around RR annotations for the student annotators. The students **voluntarily** participated in the annotations as a part of the course project. Moreover, annotators were curious about learning about AI technologies and further contributing towards its progress. There was no compulsion to take part in the annotation activity.

The cases were selected randomly to avoid bias towards any entity, situation, or laws. Any meta-information related to individuals, organizations, and judges was removed so as to avoid any introduction of bias. For the application of corpus to judgment prediction task, we are not the first ones to do the task of judgment prediction. For the task, we took all the steps (names anonymization and removal of meta-information) as outlined in the already published work of Malik et al. (2021). The focus of this paper is rhetorical role prediction, and the task of judgment prediction is only a use-case. Moreover, in this paper we focus mainly on IT and CL cases where facts and scenarios are more objective and there are less biases compared to other types of cases (e.g., criminal and civil cases). As also described by Malik et al. (2021), we do not believe that the task could be fully automated, but rather it could augment the work of a judge or legal practitioner to expedite the legal process in highly populated countries.

Legal-NLP is a relatively new area; we have taken all the steps to avoid any direct and foreseeable ethical implications; however, a lot more exploration is required by the research community to understand implicit ethical implications. For this to happen, resources need to be created, and we are making initial steps and efforts towards it.

B Dataset and Annotations

B.1 Data Collection and Preprocessing

The IT and CL cases come from the Supreme Court of India, Bombay and Kolkata High Courts. For CL cases, we use the cases from the tribunals of NCLAT (National Company Law Appellate Tribunal)[2], CCI (Competition Commission of India)[3], COMPAT (Competition Appellate Tribunal)[4]. Since the IT laws are 50 years old and relatively dynamic, we stick to certain sections of IT domain only, whereas we use all the sections for CL domain. We restrict ourselves to the IT cases that are based on Section 147, Section 92C and Section 14A only to limit the subjectivity in cases. We randomly select 50 cases from IT and CL domain each to be annotated. We used regular expressions in Python to remove the auxillary information in the documents (For example: date, appellant and respondent names, judge names etc.) and filter out the main judgment of the document. We use the NLTK[5] sentence tokenizer to split the document into sentences. The annotators were asked to annotate these sentences with the rhetorical roles.

B.2 Annotators Details

With the help of law professors, we designed a course project centered around RR annotations for the student annotators. The students **voluntarily** participated in the annotations as a part of the course project. Moreover, annotators were curious about learning about AI technologies and further contributing towards its progress. There was no compulsion to take part in the annotation activity.

The 6 annotators come from an Indian Law University. Three of them specialize in Income Tax domain and the other three specialize in Competition Law domain.

B.3 Rhetorical Roles

We provide the definition of each of the Rhetorical Role in the main paper. Examples for each of the RR are given in Table 15. Figure 5 provides the number of sentences for each label in the IT and CL dataset. Note that representation of both the domains is similar with the exception of DIS label.

B.4 Secondary and Tertiary Annotation Labels

Legal experts pointed out that a single sentence can sometimes represent multiple rhetorical roles (although this is not common). Each expert could also assign secondary and tertiary rhetorical roles to a single sentence to handle such scenarios and

[2] https://nclat.nic.in/
[3] https://www.cci.gov.in/
[4] http://compatarchives.nclat.nic.in
[5] http://www.nltk.org/

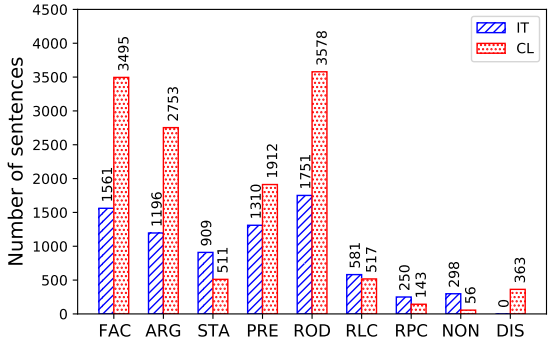


Figure 5: Distribution of RR labels in IT and CL documents.

motivate future research. On an average annotators assigned secondary role in 5-7% cases and assigned tertiary roles in 0.5-1% cases.

B.5 Inter-annotator Agreement

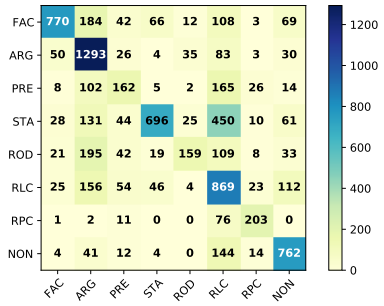
Fleiss Kappa between all (fine-grained) labels is 0.59 for IT and 0.87 for CL, indicating substantial agreement. We provide the inter-annotator agreement (averaged pairwise macro F1 between annotators) upon 13 fine-grained labels in Table 9. Also, we provide the pairwise confusion matrices of annotators (A_1, A_2) and (A_2, A_3) for both IT and CL domain in Figure 6.

Label	IT	CL
ARG-P	0.74	0.90
ARG-R	0.73	0.97
FAC	0.77	0.88
ISS	0.75	0.75
PRE-RU	0.67	0.86
PRE-NR	0.58	0.80
PRE-O	0.43	_
STA	0.78	0.89
RLC	0.58	0.74
RPC	0.75	0.74
ROD	0.64	0.93
DIS	_	0.98
NON	0.45	0.52
F1	*0.73*	*0.88*

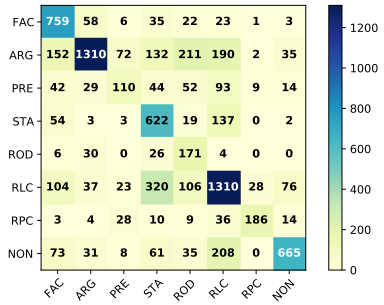
Table 9: Label-wise inter-annotator agreement for all 13 fine-grained labels.

B.6 Annotation Analysis

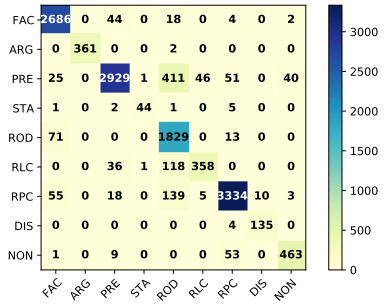
Annotation of judgments in order to identify and distinguish between the rhetorical roles played by



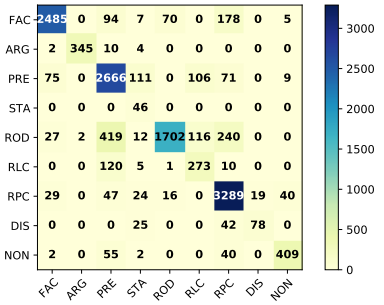
(a) Between annotators A_1 and A_2 for IT domain



(b) Between annotators A_2 and A_3 for IT domain



(c) Between annotators A_1 and A_2 for CL domain



(d) Between annotators A_2 and A_3 for IT domain

Figure 6: Confusion matrix between Annotators for IT and CL domains.

its various parts is in itself a challenging task even for legal experts. We provide some qualitative examples of sentences and their corresponding rhetorical roles in Table 15 There are several factors involved in the exercise that requires the annotator to retrace the judicial decision making and recreate the impact left by the inputs available to the judge such as certain specific facts of the case, a particular piece of argument advanced by the lawyer representing one of the parties, or a judicial precedent from a higher court deemed applicable in the current case by the lawyer(s) or by the judge or by both. Moreover, the annotator only has access to the current document which is secondary account of what actually happened in the court. These limitations certainly makes the task of the annotator further difficult, and leaves them with no choice other than to make certain educated guesses when it comes to understanding the various nuances, both ostensible and probable, of certain rhetorical roles. It should, however, be noted that such variation need not occur for every rhetorical role, since not all the roles are equally susceptible to it –for instance, the facts of the case as laid down by the judge are more readily and objectively ascertainable by more than one annotator, whereas the boundaries between the issues framed by the judge and those deemed relevant as per the arguments advanced by the lawyers may blur more, especially because if the judge happens to agree with one of the lawyers and adopts their argument as part of the judicial reasoning itself. Similarly, it should also be noted that despite differing in their views of the nature and extent of rhetorical role played by a certain part of the judgment, the annotators may still agree with each other when it comes to identifying and segregating the final ruling made by the judge in that case –this phenomenon of having used two different routes to arrive at the same destination is not uncommon in the reenactment or ex-post-facto analysis of a judicial hearing and decision making process. A cumulative effect of the aforementioned factors can be observed in the results of the annotation. The analysis provided by the three annotators in case of competition law bear close resemblance with each other. On the other hand, in case of income tax law, the analysis provided by Users 1 and 3 bear greater resemblance with each other, compared to the resemblance between Users 1 and 2, or between Users 2 and 3. On a different note, it is also observed that the rhetorical role where the annotators have differed between themselves the most has been the point of Ruling made by the Lower Court, followed by the Ratio. This also ties in with the aforesaid argument that all rhetorical roles are not equally susceptible to the variation caused by the varying levels of success achieved by the different annotators in retracing the judicial thought pattern.

B.7 Annotation Case Studies

Along with law professors, we analyzed some of the case documents. Please refer to data files for the actual judgment.

In the case of CL cases, the best resemblance that has been achieved is in the case of SC_Competition Commission of India vs Fast Way Transmission Pvt Ltd and Ors 24012018 SC.txt, one would find that the judgment has been written in a manner as to provide specific indicators before every rhetorical role. For instance, before the Ruling by Lower Court starts, reference has been made that this is the opinion given the Competition Commission of India (the lower court in the relevant domain). Similarly, before Arguments made by Petitioner/Respondent, reference has been made that this is the argument made by the lawyer representing the petitioner/respondent. This judgment also provides a nice, consistent flow following the arrangement of the rhetorical roles in order. The relatively smaller size of the judgment also indicates a lower level of complexity (although there need not always be a consistent correlation between the two). On the other hand, if one considers the least resemblance achieved in the competition law domain, in the case of SC_Excel Crop Care Limited vs Competition Commission of India and Ors 08052017 SC(1).txt, one would find that such specific indicators are usually absent, thus leaving scope for individual discretion and interpretation, the judgment goes back and forth between certain rhetorical roles (Issue, Ruling by Lower Court, Ratio by Present Court, Argument by Petitioner/Respondent, Precedent Relied Upon), and the relatively bigger size also involves additional complexity and analysis, which make room for further nuances as described above.

Similarly, if one considers the best resemblance that has been achieved in the income tax domain, in the case of SC_2014_17.txt, one would find the case has involved fewer rhetorical roles, cut down on facts (mainly dealing with procedural issues on an appellate stage), and even among the

rhetorical roles, it has focused on statutes and provisions thereof and the ratio and ruling. This has significantly reduced the possibility of the aforementioned richer jurisprudence, greater range of precedents, and resulting greater degree of subjective interpretation being at play. On the other hand, if one considers the least resemblance that has been achieved in the income tax domain, in the case of SC_2008_1597.txt, discusses Precedents to a greater detail including facts thereof, goes back and forth between certain rhetorical roles instead of maintaining a consistent order, and is not very clear about whether the judge is at times merely reiterating the arguments made by the lawyers, or is demonstrating their own view of such arguments. Collectively, these leave the scope for greater involvement of subjective interpretation of the aforesaid nuances.

Yet on an overall basis, the elements of subjectivity, personal discretionary interpretation, and arbitrariness have been minimized by the selection of the chosen domains, along with the methodology adopted for annotation, thus leading to the present success attained in identification of rhetorical roles and using the same for prior relevant case identification and prediction.

C Evaluation Metrics

We use the Macro F1 metric to evaluate the performance of models upon the task of Rhetorical Role labelling. Macro F1 is the mean of the label-wise F1 scores for each label. Given the true positives (TP), false positives (FP) and false negatives (FN), the F1 score for a single label is calculated as:

$$F1 = \frac{TP}{TP + \left(\frac{FP+FN}{2}\right)} \quad (1)$$

The pairwise inter-annotator agreement F1 between two annotators A and B is calculated by considering the annotations by annotator A as the true labels and the annotations by annotator B as the predicted labels.

We also calculate Fleiss Kappa[6] to measure the inter-annotator agreement.

D Model Training Details

All of our baseline experiments and training of Label shift prediction models (SBERT and BERT-SC) were conducted on Google Colab[7] and used the default single GPU Tesla P100-PCIE-16GB, provided by Colab. Our models were trained upon a single 11GB GeForce RTX 2080 TI. We used the SBERT model provided in the sentence-transformers library[8]. We use the Huggingface[9] implementations of BERT-base and LEGAL-BERT models. Refer to Table 10, 11 and 12 for dataset-wise results and hyperparameters for each model. We also provide the training time and number of parameters of each model in Table 13.

For SBERT-Shift, we kept the SBERT model as fixed and tuned the 3 linear layers on top. We used the Binary Crossentropy loss function with Adam Optimizer to tune the model upon the LSP task.

For BERT-SC, we fine-tuned the pre-trained BERT-base model upon the LSP task. We used the maximum sequence length of 256 tokens, a learning rate of $2e-5$ and kept the number of epochs as 5 during training. We used the same loss function and optimizer as the SBERT-Shift model.

D.1 Reduced Label Set

Due to the complexity of the task of RR prediction, we consider seven main labels (FAC, ARG, PRE, ROD, RPC, RLC, and STA) only. We plan to explore developing predictive models using fine-grained labels.

NON Label: We ignore sentences with NON (None) labels (about 4% for IT and 0.5% for CL). We believe that this was necessary since the inter-annotator agreement for the NON label in both IT and CL domains, has an F1 score as low as 0.45, implying that even the legal experts themselves do not agree whether a particular sentence has a NON label.

Dissent Label: Analysis of the annotated dataset reveals that the IT domain does not have any instance of dissent (DIS) label. There were only three documents (out of 50) in the CL domain having few instances of dissent label. Moreover, the instances of dissent label were present as a contiguous chunk of sentences at the end of the document. Hence, we discarded the sentences with dissent labels. Furthermore, law experts told us that the dissent phenomenon is rare; from a practical (application) point of view, these labels can be discarded.

[6] https://en.wikipedia.org/wiki/Fleiss%27_kappa

[7] https://colab.research.google.com/

[8] https://pypi.org/project/sentence-transformers/

[9] https://huggingface.co/

D.2 Single Sentence Classification Baselines

We train single sentence classification models for the task of rhetorical role labelling. We use BERT-base-uncased and Legal-BERT models and fine-tune them upon the sentence classification task. We also try a variant of using context sentences (left sentence and the right sentence) along with the current sentence to make classification, we call this method BERT-neighbor. We use CrossEntropyLoss as the criterion and Adam as the optimizer. We use a batch size of 32 with a learning rate of 2e-5 and fine-tune for 5 epochs for all our experiments. Refer to Tables 10 , 12 and 11 and for results and more information about the hyperparameters.

D.3 Sequence Classification Baselines

We experiment with Sequence Classification Baselines like CRF with handcrafted features, BiLSTM with sent2vec embeddings and different versions of BiLSTM-CRF in which we varied the input embeddings. We experimented with sent2vec embeddings fine-tuned on Supreme Court Cases of India (same as in (Bhattacharya et al., 2019)). We also tried with sentence embeddings obtained from the BERT-base model. In another experiment, we fine-tuned a pre-trained BERT model upon the task of Masked Language Modelling (MLM) on the unlabelled documents of IT and CL domain, and used this model to extract the sentence embeddings for the BiLSTM-CRF model.

We used the same implementation of BiLSTM-CRF from (Bhattacharya et al., 2019), with Adam optimizer and NLL loss function. Refer to Tables 10 , 12 and 11 for experiment-wise hyperparameters.

D.4 LSP-BiLSTM-CRF and MTL-BiLSTM-CRF models

In our proposed approach of LSP-BiLSTM-CRF, we experiment with two methods of generating shift embeddings, namely BERT-SC and SBERT-Shift. These embeddings were then used as input to train a BiLSTM-CRF with similar training schedules. Refer to Tables 10 , 12 and 11 for other hyperparameters.

For MTL models, we experimented with different encoders E_1 and E_2. We experimented with using Shift embeddings (or BERT embeddings of sentences obtained from pre-trained BERT model) from BERT-SC in both the components. However, the best performing model was the one in which we used shift embeddings for the shift component and BERT embeddings for the RR component. We used the NLL loss in both components of the MTL model weighted by the hyperparameter λ. We use the Adam Optimizer for training. We provide dataset-wise hyperparameters and results in Tables 10 , 12 and 11.

D.5 Hyperparameter λ

We tuned the hyperparameter λ of the MTL loss function upon the validation set. We trained the MTL model with $\lambda \in [0.1, 0.9]$ with strides of 0.1 and show the performance of our method on IT and IT+CL datasets in Figure 4. $\lambda = 0.6$ performs the best for the IT domain and also performs competitively on the combined domains.

D.6 Model Distillation

For model distillation experiments we trained the teacher model with same hyperparameters in Table 10 on the IT dataset. For the next two iteration of learning a student model, we used 48 unlabelled cases in each iteration. The weighing hyperparameter, α_U was kept as 0.3. In each iteration, the student model was trained with a batch size 16, a learning rate of 0.005 and for 300 epochs.

Model	**Hyperparameters(E=Epochs), (LR=Learning rate), (BS=Batch Size), (Dim=Embedding dimension), (E1=Embedding dimension Shift), (E2=Embedding dimension RR), (H=Hidden dimension),**	**IT (Macro F1)**
BERT	LR=2e-5, BS=32, E=5	0.56
BERT-neighbor	LR=2e-5, BS=32, E=5	0.53
Legal-BERT	LR=2e-5, BS=32, E=5	0.55
CRF(handcrafted)	LR=0.01, BS=40, Dim=172, E=300	0.55
BiLSTM(sent2vec)	LR=0.01, BS=40, Dim=200, H=100, E=300	0.55
BiLSTM-CRF(handcrafted)	LR=0.01, BS=40, Dim=172, H=86, E=300	0.57
BiLSTM-CRF(sent2vec)	LR=0.01, BS=40, Dim=200, H=100, E=300	0.59
BiLSTM-CRF(BERT emb)	LR=0.01, BS=40, Dim=768, H=384, E=300	0.63
BiLSTM-CRF(MLM emb)	LR=0.01, BS=40, Dim=768, H=384, E=300	0.58
LSP(SBERT)	LR=0.005, BS=40, Dim=2304, H=1152, E=300	0.64
LSP(BERT-SC)	LR=0.005, BS=40, Dim=2304, H=1152, E=300	0.65
MTL(MLM emb)	LR=0.005, BS=40, E1=2304, E2=768 , H=1152(Shift), H=384(RR), E=300	0.67
MTL(BERT-SC)	LR=0.005, BS=40, E1=2304, E2=768, H=1152(Shift), H=384(RR), E=300	0.70
MTL(BERT-SC)	LR=0.005, BS=40, E1=2304, E2=2304, H=1152(Shift), H=384(RR), E=300	0.68
MTL(BERT-SC)	LR=0.005, BS=40, E1=768, E2=768, H=1152(Shift), H=384(RR), E=300	0.64

Table 10: Hyperparameters and results on the IT dataset

Model	**Hyperparameters(E=Epochs), (LR=Learning rate), (BS=Batch Size), (Dim=Embedding dimension), (E1=Embedding dimension Shift), (E2=Embedding dimension RR), (H=Hidden dimension),**	**IT+CL (Macro F1)**
BiLSTM-CRF(sent2vec)	LR=0.01, BS=40, Dim=200, H=100, E=300	0.65
BiLSTM-CRF(BERT)	LR=0.01, BS=40, Dim=768, H=384, E=300	0.63
LSP-BiLSTM-CRF(BERT-SC)	LR=0.005, BS=20, Dim=2304, H=1152, E=300	0.67
MTL-BiLSTM-CRF(BERT-SC)	LR=0.005, BS=20, E1=2304, E2=768, H=1152(Shift), H=384(RR), E=300	0.70
MTL-BiLSTM-CRF(BERT-SC)	LR=0.005, BS=20, E1=2304, E2=2304, H=1152(Shift), H=384(RR), E=300	0.68
MTL-BiLSTM-CRF(BERT-SC)	LR=0.005, BS=20, E1=768, E2=768, H=1152(Shift), H=384(RR), E=300	0.65

Table 11: Hyperparameters and results on the combined (IT+CL) dataset

Model	**Hyperparameters(E=Epochs), (LR=Learning rate), (BS=Batch Size), (Dim=Embedding dimension), (E1=Embedding dimension Shift), (E2=Embedding dimension RR), (H=Hidden dimension),**	**CL (Macro F1)**
BERT	LR=2e-5, BS=32, E=5	0.52
BERT-neighbor	LR=2e-5, BS=32, E=5	0.51
Legal-BERT	LR=2e-5, BS=32, E=5	0.53
CRF(handcrafted)	LR=0.01, BS=40, Dim=172, E=300	0.52
BiLSTM(sent2vec)	LR=0.01, BS=40, Dim=200, H=100, E=300	0.54
BiLSTM-CRF(handcrafted)	LR=0.01, BS=40, Dim=172, H=86, E=300	0.56
BiLSTM-CRF(sent2vec)	LR=0.01, BS=40, Dim=200, H=100, E=300	0.61
BiLSTM-CRF(BERT emb)	LR=0.01, BS=40, Dim=768, H=384, E=300	0.63
BiLSTM-CRF(MLM emb)	LR=0.01, BS=40, Dim=768, H=384, E=300	0.60
LSP(SBERT)	LR=0.005, BS=40, Dim=2304, H=1152, E=300	0.63
LSP(BERT-SC)	LR=0.005, BS=40, Dim=2304, H=1152, E=300	0.68
MTL(MLM emb)	LR=0.005, BS=20, E1=2304, E2=768 , H=1152(Shift), H=384(RR), E=300	0.67
MTL(BERT-SC)	LR=0.005, BS=20, E1=2304, E2=768, H=1152(Shift), H=384(RR), E=300	0.69
MTL(BERT-SC)	LR=0.005, BS=20, E1=2304, E2=2304, H=1152(Shift), H=384(RR), E=300	0.67
MTL(BERT-SC)	LR=0.005, BS=20, E1=768, E2=768, H=1152(Shift), H=384(RR), E=300	0.64

Table 12: Hyperparameters and results on the CL dataset

Model	No of Parameters		Training Time(min)	
	IT	CL	IT	CL
BiLSTM(sent2vec)	240000	240000	15	30
BiLSTM-CRF(sent2vec)	240000	240000	15	30
BiLSTM-CRF(BERT emb)	3538944	3538944	30	50
BiLSTM-CRF(MLM emb)	3538944	3538944	30	50
LSP(SBERT)	31850496	31850496	90	250
LSP(BERT-SC)	31850496	31850496	90	250
MTL(MLM emb)	35411060	35411060	300	1200
MTL(BERT-SC)	35411060	35411060	300	1200

Table 13: Approx. number of parameters and computational budget of models.

Model	**IT+CL docs**	**F1**
BERT-ILDC	Predicted ROD & RPC using BiLSTM-CRF(sent2vec)	0.55
BERT-ILDC	Predicted ROD & RPC using MTL(BERT-SC)	0.56

Table 14: Judgment Prediction results using predicted ROD & RPC

Label	Sentence
Fact	It has also been alleged that the copies of the notices were also sent, inter alia, to the principal officer of the said company and also to the ladies as mentioned herein before, who has sold the immovable property in question.
Fact	For executing this contract, the assessee entered into various contracts -Offshore Supply contract and Offshore Service Contracts.
Ruling By Lower Court	But the words inland container depot were introduced in Section 2(12) of the Customs Act, 1962, which defines customs port.
Ruling By Lower Court	We may also mention here that the cost of superstructure was Rs. 2,22,000 as per the letter of the assessee dated 28-11-66 addressed to the ITO during the course of assessment proceedings.
Argument	Such opportunity can only be had by the disclosure of the materials to the court as also to the aggrieved party when a challenge is thrown to the very existence of the conditions precedent for initiation of the action.
Argument	In this connection, it was urged on behalf of the assessee(s) that, for the relevant assessment years in question, the Assessing Officer was required to obtain prior approval of the Joint Commissioner of Income Tax before issuance of notice under Section 148 of the Act.
Statute	In the meantime, applicant has to pay the additional amount of tax with interest without which the application for settlement would not be maintainable.
Statute	On the other hand, interest for defaults in payment of advance tax falls under section 234B, apart from sections 234A and 234C, in section F of Chapter XVII.
Ratio of the Decision	The State having received the money without right, and having retained and used it, is bound to make the party good, just as an individual would be under like circumstances.
Ratio of the Decision	Therefore, the Department is right in its contention that under the above situation there exists a Service PE in India (MSAS).
Ruling by Present Court	For these reasons, we hold that the Tribunal was wrong in reducing the penalty imposed on the assessee below the minimum prescribed under Section 271(1)(iii) of the Income-tax Act, 1961.
Ruling by Present Court	Hence, in the cases arising before 1.4.2002, losses pertaining to exempted income cannot be disallowed.
Precedent	Yet he none the less remains the owner of the thing, while all the others own nothing more than rights over it.
Precedent	I understand the Division Bench decision in Commissioner of Income-tax v. Anwar Ali, only in that context.
None	Leave granted.
None	There is one more way of answering this point.
Dissent	Therefore a constructive solution has to be found out.
Dissent	In the light of the Supreme Court decision in the case of CCI vs SAIL (supra) t his issue has to be examined.

Table 15: Example sentences for each label.