Corpus for Automatic Structuring of Legal Documents

Prathamesh Kalamkar[1,2,*], **Aman Tiwari**[1,2,*], **Astha Agarwal**[1,2,*], **Saurabh Karn**[3,*],
Smita Gupta[3], **Vivek Raghavan**[1], **Ashutosh Modi**[4]

[1]EkStep Foundation, [2]Thoughtworks Technologies India Pvt Ltd.,
[3]Agami , [4] Indian Institute of Technology Kanpur (IIT-K)
{prathamk, aman.tiwari, astha.agarwal}@thoughtworks.com,
{saurabh, smita}@agami.in, vivek@ekstep.org, ashutoshm@cse.iitk.ac.in

Abstract

In populous countries, pending legal cases have been growing exponentially. There is a need for developing techniques for processing and organizing legal documents. In this paper, we introduce a new corpus for structuring legal documents. In particular, we introduce a corpus of legal judgment documents in English that are segmented into topical and coherent parts. Each of these parts is annotated with a label coming from a list of pre-defined *Rhetorical Roles*. We develop baseline models for automatically predicting rhetorical roles in a legal document based on the annotated corpus. Further, we show the application of rhetorical roles to improve performance on the tasks of summarization and legal judgment prediction. We release the corpus and baseline model code along with the paper.

Keywords: Legal NLP, Rhetorical Roles, Legal Document Segmentation

1. Introduction

In populous countries (e.g., India), pending legal cases have been growing exponentially. For example, according to India's National Judicial Data Grid, as of December 2021, there are approximately 40 million cases pending in various courts of the country (National Judicial Data Grid, 2021). India follows a common-law system; consequently, due to subjectivity involved in the legal process, it may not be possible to automate the entire judicial pipeline completely; nevertheless, many intermediate tasks can be automated to augment legal practitioners, and hence expedite the system. For example, legal documents can be processed with the help of Natural Language Processing (NLP) techniques to organize and structure the data to be amenable to automatic search and retrieval. However, legal texts are different from commonly occurring texts typically used to train NLP models. Legal documents are quite long, running into tens (sometimes hundreds) of pages. Long documents make automatic processing challenging as information is spread throughout the document (Malik et al., 2021b). Another challenge with legal documents is the use of different lexicons. Though legal documents use natural language (e.g., English), many commonly occurring words/terms have different legal connotations. The use of different lexicons makes it challenging to adapt existing NLP models to legal texts (Malik et al., 2021b). Moreover, in countries like India, legal documents are manually typed and are highly unstructured and noisy (e.g., spelling and grammatical mistakes). Above mentioned challenges make it difficult to apply existing NLP models and techniques directly, which calls for the development of legal domain-specific techniques.
Existing state-of-the-art models in NLP are data-driven and are trained on annotated corpora. However, the legal domain suffers from the deficiency of availability of annotated corpora. It has hindered the growth of the Legal NLP domain. For example, much of the recent success in the computer vision community can be owed to the creation and availability of annotated vision corpora such as ImageNet (Deng et al., 2009; Russakovsky et al., 2013; Russakovsky et al., 2015). In this paper, we contribute to creating annotated legal text corpora. In particular, we create a new corpus of Indian legal judgments in English that are structured and annotated with topically coherent semantic units. Since legal documents are long and unstructured, these can be divided into topically coherent parts (e.g., facts, arguments) referred to as *Rhetorical Roles* (Saravanan et al., 2008; Bhattacharya et al., 2019b; Malik et al., 2021a). In this paper, with the help of legal experts, we annotate legal documents with 12 different Rhetorical Roles (RRs) (details in §3). An example text annotated with some of the RRs is shown in Figure 1. As shown in the figure, an unstructured legal judgment document is segmented into semantically coherent parts, and each part is annotated with a rhetorical role label such as preamble, fact, ratio, etc. We experimented with different levels of granularity (phrase level, sentence level, paragraph level) for annotating RRs and decided to go for sentence-level RR annotations based on initial experiments. Each sentence in a legal document is annotated with a rhetorical role label in the proposed corpus. Typically, consecutive sentences can have a similar role in a judgment document. The rhetorical role corpus is part of a general open-source effort of creating various legal corpora for promoting the development and bench-marking of legal NLP systems. This project is called BUILDNyAI.[1] We make the following contribu-

---

* Authors contributed equally

[1]The word BUILDNyAI is a code-mixed (English+Hindi)

Figure 1: Example of document segmentation via Rhetorical Roles labels. On the left is excerpt from a legal document and on the right is document segmented and labelled with rhetorical role labels.

tions in this paper:

- We create a corpus of 354 Indian legal documents annotated with rhetorical roles. The corpus has 40,305 sentences annotated with 12 different RRs. To the best of our knowledge, this is the largest corpus of legal documents annotated with RRs.

- In order to be of practical value, using the corpus, we develop a transformer-based baseline model for automatically annotating legal documents with sentence-level RR.

- We show two use-cases for RRs. In particular, we show applications of RRs to the task of legal case summarization and legal judgment prediction.

- We release the corpus and the model implementations: `https://legal-nlp-ekstep.github.io/Competitions/Rhetorical-Role/`

2. Related Work

In recent times, there has been lot of work in the area of legal text processing. Different tasks and techniques have been proposed. For example, Prior Case Retrieval (Jackson et al., 2003), Summarization (Moens et al., 1999; Saravanan et al., 2007), Case Prediction (Malik et al., 2021b; Chalkidis et al., 2019; Strickson and De La Iglesia, 2020a; Şulea et al., 2017; Kapoor et al., 2022), Argument Mining (Wyner et al., 2010; Moens et al., 2007), Information Extraction and Retrieval (Tran

et al., 2019; Grabmair et al., 2011; Tran et al., 2019), and Event Extraction (Lagos et al., 2010; Maxwell et al., 2009; Lagos et al., 2010).

Recently, efforts have been made to develop corpora that could aid various legal NLP tasks; for example, Malik et al. (2021b) have released a corpus of 35K Indian Supreme Court documents for the task of judgment prediction and explanation. Chalkidis et al. (2019) have released 11,478 legal documents corresponding to the European Court of Human Rights (ECHR). Strickson and De La Iglesia (2020b) have proposed a corpus of 4,959 UK Supreme Court documents. Xiao et al. (2018) have created a large-scale corpus of 2.68 million criminal case documents and released CAIL (Chinese AI and Law Challenge) dataset for judgment prediction. A new multilingual dataset of European Union (EU) legal documents has been recently released by Chalkidis et al. (2021).

Research in rhetorical roles for legal text processing has been active in the past few years. Farzindar and Lapalme (2004; Hachey and Grover (2006) have leveraged rhetorical roles to create summaries of legal texts. Saravanan et al. (2008) proposed a CRF-based model using hand-crafted features for segmenting documents using seven different roles. Bhatia (2014) created Genre Analysis of Legal Texts to create seven rhetorical categories. Bhattacharya et al. (2019b) have proposed CRF-BiLSTM model for automatically assigning rhetorical roles to sentences in Indian legal documents. (Malik et al., 2021a) have created a RR corpus and annotated with 13 fine-grained roles and further they have developed a multi-task learning based model

---

term having English word BUILD and Hindi word nyAI (short for nyayi, which means justice). The project is hosted at `https://legal-nlp-ekstep.github.io/Competitions/Rhetorical-Role/`

for predicting RR. In this paper, we also propose a corpus of English Indian legal judgment documents annotated with Rhetorical Roles; however, we annotate the documents with a more extensive set of 12 rhetorical role labels and a NONE label (in the case none of the 12 labels are applicable). Moreover, to the best of our knowledge, we create the largest corpus of 354 documents (vs. 100 documents in previous RR corpus by Malik et al. (2021a)), with 40,315 sentences annotated with 13 (12 + NONE) different types of rhetorical role labels. We propose state-of-the-art transformer models for RR prediction and show the use case of RRs for case summarization and legal judgment prediction.

Recent success in almost every area in NLP has been due to transformer-based neural architectures (Wang et al., 2018). We do not discuss the details of transformer architectures here and refer the reader to the survey on transformers by Tay et al. (2020). We develop transformer-based baseline models for automatically segmenting legal documents into RRs units.

3. Rhetorical Roles Corpus

As outlined earlier, legal documents are typically long, and information is spread throughout the document. In order to make the automatic processing of documents easier, documents are divided into topically coherent segments referred to as Rhetorical Roles (Malik et al., 2021a). In this paper, we propose the use of 12 RRs and a NONE label. We started with the list of RR labels proposed by Bhattacharya et al. (2019b); however, we found some of the RR to be ambiguous, hence after having elaborate discussions with law professors, we split some of the RRs (arguments and precedents) to arrive at the list of 12 main roles. Details and definitions for each of the RR are as follows:

- **Preamble (PREAMBLE):** This covers the metadata related to the legal judgment document. A typical judgment would start with the court name, the details of parties, lawyers and judges' names, headnote (summary). This section typically would end with a keyword like (JUDGMENT or ORDER). Some documents also have HEADNOTES, ACTS sections in the beginning. These are also part of the Preamble.
- **Facts (FAC):** This corresponds to the facts of the case. It refers to the chronology of events that led to filing the case and how it evolved (e.g., First Information Report (FIR) at a police station, filing an appeal to the Magistrate, etc.) Depositions and proceedings of the current court, and summary of lower court proceedings.
- **Ruling by Lower Court (RLC):** Cases are not directly filed in the higher courts but are appealed from lower courts. Consequently, the documents contain judgments given by the lower courts (Trial Court, High Court) based on the present appeal (to the Supreme Court or high court). The lower court's verdict, analysis, and the ratio behind the judgment by the lower court is annotated with this label.
- **Issues (ISSUE):** Some judgments mention the key points on which the verdict needs to be delivered. Such Legal Questions Framed by the Court are ISSUES.
- **Argument by Petitioner (ARG_PETITIONER):** Arguments by petitioners' lawyers. Precedent cases argued by petitioner lawyers fall under this category, but when the court discusses them later, then they belong to either the relied / not relied upon category.
- **Argument by Respondent (ARG_RESPONDENT):** Arguments by respondents' lawyers. Precedent cases argued by respondent lawyers fall under this, but when the court discusses them later, they belong to either the relied / not relied category.
- **Analysis (ANALYSIS):** These are views of the court. This includes courts' discussion on the evidence, facts presented, prior cases, and statutes. Discussions on how the law is applicable or not applicable to the current case. Observations (non-binding) from the court. It is the parent tag for three tags: PRE_RLEIED, PRE_NOT_RELIED, and STATUTE i.e., every statement which belongs to these three tags should also be marked as ANALYSIS.
- **Statute (STA):** This includes texts in which the court discusses established laws, that can come from a mixture of sources: Acts , Sections, Articles, Rules, Order, Notices, Notifications, and Quotations directly from the bare act. The statute will have both the tags Analysis + Statute.
- **Precedent Relied (PRE_RELIED):** Texts in which the court discusses prior case documents, discussions and decisions which were relied upon by the court for final decisions. Precedent will have both the tags Analysis + Precedent.
- **Precedent Not Relied (PRE_NOT_RELIED):** Texts in which the court discusses prior case documents, discussions and decisions which were not relied upon by the court for final decisions. It could be due to the fact that the situation, in that case, is not relevant to the current case.
- **Ratio of the decision (Ratio):** This includes the main reason given for the application of any legal principle to the legal issue. It is the result of the analysis by the court. It typically appears right before the final decision. It is not the same as "Ratio Decidendi" taught in the legal academic curriculum.
- **Ruling by Present Court (RPC):** Final decision + conclusion + order of the Court following from the natural/logical outcome of the rationale.
- **NONE:** If a sentence does not belong to any of the above categories, it is labeled as NONE.

Dataset	Docs	Sen-tences	Tokens	Avg To-kens
Train	247	28986	938K	3797
Validation	30	2879	88K	2947
Test (in-domain)	50	4158	134K	2681
Test (out-domain)	27	4292	127K	4722
Total	**354**	**40315**	**1.3M**	**3638**

Table 1: Corpus Statistics: The corpus is split into train, val and test. The table shows number of documents, sentences, tokens and average number of tokens per document.

3.1. Corpus Documents

The corpus consists of legal judgment documents from the Supreme Court of India, High Courts in different Indian states, and some district-level courts. Raw judgment text files were scraped from Indian Court websites.[2] Data has a mix of Supreme Court judgments (40%), High Courts judgments (40%) and district court judgments (20%). To develop baseline models, we divided the dataset into train, and validation. Test set was further divided into in-domain and out of domain. Train, validation and test (in-domain) datasets contain annotated judgments belonging to tax and criminal cases. Test (out-domains) contains annotated judgements from 3 domains : Motor Vehicles Act (9), Industrial and Labour law (8) and Land and Property law (10). The statistics of the corpus are shown in Table 1. Table 2 gives number of sentences for each role in the entire corpus. Qualified law experts annotated test data with cross checks.

Rhetorical Role	Sentences
ANALYSIS	14300
ARG PETITIONER	1771
ARG RESPONDENT	1068
FAC	8045
ISSUE	535
NONE	2037
PREAMBLE	6116
PRE NOT RELIED	217
PRE RELIED	1934
RATIO	1014
RLC	1081
RPC	1562
STA	625
Overall	40305

Table 2: Role-wise sentence count in the entire corpus

3.2. Annotation Process

The annotation process was designed in consultation with legal experts (law professors and legal practitioners). Given the nature of the task, the RR annotations

---

require a deep understanding of the law and the legal process. Consequently, we involved law students and legal practitioners in annotating the documents. The process involved annotating each sentence in a given document with one of the 12 RR + None labels described earlier. We experimented with different levels of granularity (phrase level, sentence level, paragraph level, etc.) for annotating the documents with RR. Pilot experiments indicated sentence level RR annotation to be appropriate as it maintains the balance (with regard to semantic coherence) between too short and too long texts. The legal documents were split using spaCy library (spaCy, 2021). Rhetorical role annotation is not a trivial task; we faced two main challenges in the annotation activity: availability for a large group of legal experts and, secondly, motivating the legal experts to perform annotation consistently while maintaining quality. We performed the annotation activity via crowdsourcing as described next.

3.3. Data Annotation Pipeline

Corpus documents were annotated via a crowdsourcing activity. We invited law students from various law schools across the country to volunteer for the data annotation exercise. We created processes to onboard student volunteers and introduced them to the entire activity and its goal. Filtering was carried out at multiple stages to retain the most motivated and consistent (from the perspective of quality of the annotations) students. The entire pipeline is shown in Figure 2. We describe each stage of the pipeline in the next sections.
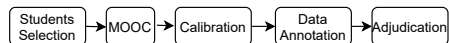
Figure 2: Data Annotation Pipeline

3.3.1. Student Selection

We did a nationwide call for volunteers through a network of law students. The application required students to describe their motivation. A basic screening was done to eliminate applications that were partially filled. Finally, after filtering, we selected an initial group of 50 students. The selected students were then on-boarded and were motivated by explaining the big picture of the impact of their contribution. The data annotations were done voluntarily by law students from multiple Indian law universities. Interaction with the law students revealed that they were motivated to learn more about AI and contribute towards the development of the AI field, and hence they volunteered for the activity. In order to smoothly conduct the annotation activity via crowdsourcing, we organized the volunteers in a hierarchical structure based on their experience and performance during a pilot study. The organizational structure for this exercise is shown in Figure 3.

Project Administrators: They designed data collection and communication processes, built tools for data

Figure 3: Organization Structure

Figure 4: Ground Truth Score Histogram

collection, and supervised the overall activity. This group included law experts and authors of the paper.

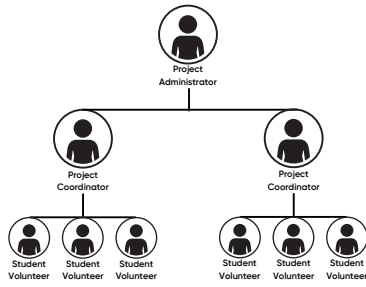**Project Coordinators:** They mentored and resolved the doubts of the students. They were responsible for assuring the quality of the data. Coordinators identified and rectified conceptual errors among the students. Further, the coordinators assisted the administrators during the adjudication process.

Student Volunteers: They annotated the data and also provided feedback on the entire process. Volunteers were in constant communication with the coordinators. At later stages of annotations, some of the best-performing students assisted in the adjudication process (§3.3.5). Best performing students were selected based on two criteria: timely submissions and ground truth agreement score. Students were assessed if they completed the task within a stipulated time at each annotation stage. Furthermore, each batch of annotation document consisted of sentences for which true (gold) RR labels were known apriori (also §3.3.4). Students were assessed for their performance on the ground truth (sentences with gold RR labels), and students who were correct on at least 90% of ground truth sentences were considered for the best performing category.

Before beginning the entire activity, we conducted a small pilot to assess the feasibility of crowdsourcing with student volunteers. Volunteers who completed MOOC, calibration and annotation exercises with satisfactory performance were then invited to become project coordinators for the subsequent data collection phase. The chance to become coordinator further provided positive reinforcement for the efforts, thus keeping the students well motivated. In the end, we selected eight students as project coordinators.

3.3.2. MOOC

Law students do not have an understanding of the workings of AI. We designed a MOOC (Massive Open Online Course)[3] for the annotators. The MOOC explained the AI technologies to the law students, described the process of building datasets for AI algorithms, and explained the concept of the rhetorical role. Students were expected to complete the MOOC in a stipulated amount of time and complete the associated
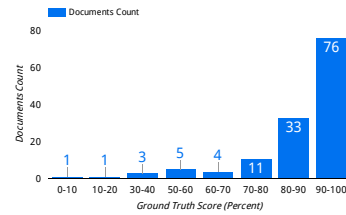
quiz, which checked for a basic understanding of the rhetorical role definitions.

3.3.3. Calibration

Since in the initial stages, students can differ in understanding RRs. We calibrated the students to bring them to a common ground. Calibration focused on shaping a common understanding of definitions among students. Students were asked to annotate three judgments that experts had already annotated. The sentences that differed from expert (gold) annotations were highlighted, and students were asked to calibrate their annotations. Calibration was an iterative process, and it was carried out till students came at the level of expert annotations.

3.3.4. Data Annotation

In the end, 35 out of 50 selected students qualified for the calibration stage, and this was the final pool that annotated the entire corpus. Each student annotated 24 documents, and three students annotated each document. We did not observe any student dropout after the calibration stage. On average, it took about 40 minutes to annotate a single document. The entire annotation activity took around six weeks. Students annotated train and validation documents (= 277), and experts annotated 77 test documents. As described earlier, during the annotation process, each student was also randomly assigned four documents (chosen randomly with replacement from the test set) for which gold (ground truth) annotations were known to coordinators and administrators but not to the students. The performance of students (referred to as Ground Truth Score) on these gold documents was assessed. Ground truth score is the percentage of sentences in gold documents that are correctly annotated. The average ground truth score for all students was 85%. Figure 4 shows histogram of ground truth scores for a judgment. It shows that the majority of documents are in the 90 to 100 percent range, indicative of consistent annotations with ground truth docs. Note that documents shown in Figure 4 (y-axis) are chosen randomly (with replacement) from the test set and hence there is overlap between documents across different batches. Furthermore, coordinators provided feedback to students with lower scores to improve their overall annotation quality.

3.3.5. Adjudication

A majority voting scheme was used to decide the final RR label. However, in some instances, annotators as-

---

[3] https://www.youtube.com/playlist?list=PL1z52lLL6eWnDnc3Wgfcu6neczrU3fFw0

signed three different labels; such documents were further sent for adjudication. The adjudication was done by experts, project coordinators, and some of the best-performing students (§3.3.1).

3.3.6. Annotation Quality Assessment

Final annotation quality was evaluated using Fleiss Kappa (Fleiss et al., 2013). Overall, Fleiss Kappa score was 0.59, pointing towards moderate agreement. We saw high agreement amongst annotators on PREAMBLE, RPC, NONE, and ISSUE. There were medium agreements on FACTS, RLC, ANALYSIS, PRECEDENT, and ARGUMENTS. RATIO was the most ambiguous role. ANALYSIS was very often confused with FACTS and ARGUMENTS. In a judgment, a judge emphasizes some of the facts, which as per definition, are considered as analysis role; however, annotators often confuse them as facts role. Moreover, sometimes the judge may mention arguments and give their opinion on it; this, as per definition, is the analysis role, but annotators sometimes confuse it with the argument role. FACTS was sometimes confused with RLC (Ruling by Lower Court).

4. RR Prediction Baseline Models

The end goal behind this work has been to encourage the development of systems that can segment a new legal document automatically in terms of rhetorical roles. Towards this goal, we experimented with some baseline models. Since transformer-based models (Wolf et al., 2020) have shown state-of-the-art (SOTA) performance on most of the NLP tasks, including the tasks in legal NLP domain (Malik et al., 2021b), we mainly experimented with them. In the RR prediction task, given a legal document, the task is to predict the RR label for each sentence in the document. We pose this as a multi-class sequence prediction problem. We initially experimented with variants of the model by Bhattacharya et al. (2019b). In particular, we use a CRF (Conditional Random Field) model for RR prediction. The features for this CRF model come from a transformer, i.e., the BERT-BASE (Devlin et al., 2018) model is used to get sentence embeddings corresponding to the CLS token. These sentence embeddings are then passed through the CRF layer to get final predictions. We call this model BERT_CRF. We also tried the architecture proposed by Cohan et al. (2019) which captures contextual dependencies using only BERT without the need for hierarchical encoding using a CRF. We call this model BERT_only. After experiments with vanilla transformer models, we finally created the baseline system using the SciBERT-HSLN architecture (Brack et al., 2021). Figure 5 shows the overall architecture of the proposed model. In the proposed model, each sentence is passed through BERT BASE model to get word embeddings, these embeddings are further processed by Bi-LSTM layer followed by attention-based pooling layer to get sentence representations $\{s_1, s_2, ..s_n\}$.
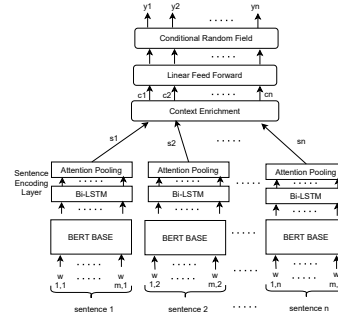

Figure 5: RR Prediction Baseline model inspired by Brack et al. (2021)

Model	Precision	Recall	F1
BERT_CRF	0.24	0.24	0.23
BERT_only	0.67	0.68	0.67
SciBERT-HSLN	0.79	0.80	0.79

Table 3: Performance of models on test (in-domain) data

Context Enrichment layer encodes the contextual information, by taking sequence of sentence representations, resulting in contextualized sentence representations: $\{c_1, c_2, .., c_n\}$. This is followed by MLP layers and CRF that leverage the distributed representation features to predict the RR label for each sentence via softmax activation.

Results: The performance of different models was tested on test(in-domain) data and results are given in Table 3. We use standard weighted F1 score metric for evaluation. As can be observed, the BERT_CRF model performs the worst, and the BERT_only model performs worse than the proposed model SciBERT-HSLN, which achieved a weighted F1 score of 78%. It is perhaps because SciBERT-HSLN, being a sequential model, can capture longer range dependencies between sentences in a document. The results of the model on the test set for each of the RR labels are shown in Table 4. Figure 6 shows the confusion matrix for the SciBERT-HSLN model. As can be observed from Table 4 and Figure 6, ARGUMENTS based roles are miss-classified very often and confused among the two types of ARGUMENTS and also sometimes confused with FACTS and ANALYSIS. PREAMBLE is almost perfectly classified. As can be seen, PRECEDENT NOT RELIED is completely miss-classified and confused with PRECEDENT RELIED and ANALYSIS. RATIO is often confused with ANALYSIS, and this trend is similar to what was observed for annotators as well. Similar to what was observed for annotators, RPC, PREAMBLE, NONE and ISSUE are classified with decent F1 scores. STATUES are also not well classified as many times a judge mentions some laws in their opinion and model tends to learn these spurious patterns as analysis and miss-classifies actual stat-

Rhetorical Role	Precision	Recall	F1
ANALYSIS	0.77	0.89	0.83
ARG_PETITIONER	0.60	0.64	0.62
ARG_RESPONDENT	0.84	0.41	0.55
FAC	0.80	0.84	0.82
ISSUE	0.93	0.87	0.90
NONE	0.85	0.84	0.85
PREAMBLE	0.96	0.98	0.97
PRE_NOT_RELIED	0.00	0.00	0.00
PRE_RELIED	0.79	0.60	0.68
RATIO	0.53	0.56	0.54
RLC	0.75	0.45	0.57
RPC	0.78	0.87	0.82
STA	0.77	0.54	0.64
Overall	0.79	0.80	0.79

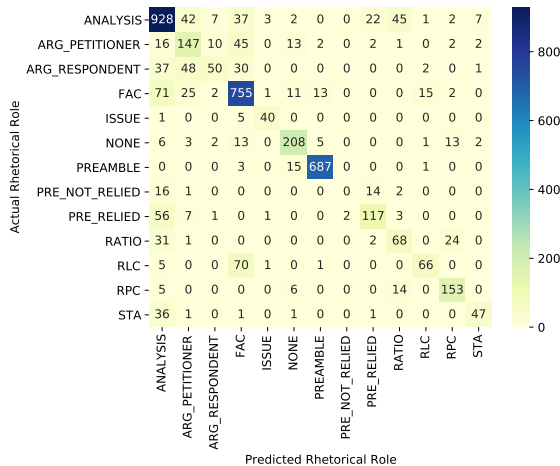Table 4: F1 scores of RR baseline model for each of the rhetorical role on test data

Figure 6: Confusion Matrix for SciBERT-HSLN model predictions on the test data

ues as analysis. We have also created a leaderboard[4] for the task of RR prediction where other researchers can experiment with various approaches.

Results on test (out-domain) data: In order to check if the baseline model trained on Criminal and Tax cases generalized to other domains, we tested the baseline model on 27 judgments from Motor Vehicles, Industrial and Labour and Land and Property cases. Weighted F1 reduced to 0.70. This degradation in performance is mainly due to different style of writing in the judgments.

5. Applications of Rhetorical Roles Prediction Task

The purpose of creating a rhetorical role corpus is to enable automated understanding of legal documents by segmenting them into topically coherent units. This can be helpful in various applications such legal document

---

[4]https://legal-nlp-ekstep.github.io/Competitions/Rhetorical-Role/

Model	ROUGE-1	ROUGE-2	ROUGE-L
BERTSUM	0.60	0.42	0.59
BERTSUM RR	0.62	0.46	0.61

Table 5: Extractive Summarization Results

summarization (Bhattacharya et al., 2019a), and legal judgment prediction (Malik et al., 2021b). In this paper, we explore both the use-cases. We experimented with how rhetorical roles prediction could help create abstractive, extractive summaries of Indian court judgments and predict the judgment outcome based on the judgment text.

5.1. Extractive Summarization of Court Judgments using Rhetorical Roles

We explored the task of extractive summarization. For a given legal document, the task requires extracting the salient sentences that would summarize the document. We experimented with the LawBriefs corpus consisting of 285 extractive summaries of Indian court judgments prepared by law students from a National Law University in India. The corpus was created by providing judgment documents to law students, followed by a questionnaire that required them to pick salient sentences that would answer the questions and, in the process, create the summaries. The questions pertained to facts, arguments, issues, ratio, and decisions. We wanted to experiment with how rhetorical roles could be helpful in extracting summaries.

We finetuned BERTSUM (Liu and Lapata, 2019) model on the Lawbriefs data to pick up the top 20% of the sentences as summaries. Since the judgments are much longer than 512 token limits of BERTSUM, we created non-overlapping chunks of 512 tokens and created 3151 chunks in training data from 235 judgments and 827 chunks from 50 judgments as test data. We then trained another model, which also takes as input a rhetorical role for each sentence. We concatenated 768-dimensional sentence vector from CLS token to one-hot encoded sentence rhetorical roles. The idea is that if certain rhetorical roles are more important than others while creating summaries, then the model will learn those. We call this model BERTSUM RR. Discussion with Legal Experts revealed that ISSUE, RATIO, and RPC are important in summary and must always be selected without the need of summarizing. So we copied all the sentences with predicted rhetorical roles ISSUE, RATIO and RPC regardless of whether they are present in the top 20% sentences. Model performance evaluated using ROUGE scores (Lin, 2004) are compared in Table 5. Results indicate that rhetorical roles are useful in selecting better summary sentences.

5.2. Abstractive Summarization of Court Judgments using Rhetorical Roles

The task of abstractive summarization requires generating concise text summaries of legal documents. For

our experiments, we considered 50 randomly selected documents from the Law Briefs dataset (as described in 5.1) as test data. For this task we used pre-trained Legal Pegasus model.[5] Legal Pegasus is fine-tuned version of Pegasus (Zhang et al., 2020) on US securities litigation dataset.[6] We used the pre-trained Legal Pegasus model for generating abstractive summaries for the baseline. In particular, we split the document into non-overlapping chunks of 1024 tokens, and each chunk was passed through the model to generate summaries. The final summary was obtained by concatenating summaries of each chunk. It constituted the baseline model. We wanted to see how RR could help generate better summaries. Towards this goal, we segmented the document in terms of rhetorical roles, and each of the segments was passed separately through the Legal Pegasus model to generate summaries. The final summary was obtained by concatenating the summaries corresponding to each of the rhetorical roles in the order they appear in the document. This corresponds to the Legal Pegasus RR model. Both models are compared on the test set and ROUGE scores for both the model are shown in Table 6. As can be observed in Table 6 use of rhetorical roles helps to improve the performance on the task of abstractive summarizing.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Legal Pegasus	0.55	0.34	0.47
Legal Pegasus RR	0.56	0.36	0.48

Table 6: Abstractive Summarization Results

5.3. Court Judgment Prediction using Rhetorical Roles

Malik et al. (2021b) created the corpus (ILDC: Indian Legal Documents Corpus) and the task (CJPE: Court Judgment Prediction and Explanation) for predicting and explaining the court judgments based on legal judgment texts. It is essential for the judgment prediction task to identify which sentences provide hints about the final decision and use that filtered data as input for prediction. We predicted rhetorical role for each sentence of the train, test data using the baseline rhetorical role model. In the ILDC dataset, we removed the sentences with RPC and RATIO tags making the task more challenging. We also removed the judgments for which no ANALYSIS was predicted. Note that the ILDC dataset is already anonymized and takes care of the biases and ethical concerns associated with the task of judgment prediction. Moreover, we use judgment prediction only as a use case and do not believe that an automated system could remove a human judge; rather,

such a system could augment a human and expedite legal processes, especially in highly populated countries like India.

For the task of judgment prediction, training data had 5044 judgments, and test data had 977 judgments. The idea is to filter the training data using rhetorical roles to check the impact on model performance, keeping the model architecture the same. We used XLNet on the ILDC single model proposed in Malik et al. (2021b) to predict the judgment outcome on the last 512 tokens of the judgment text. We call this approach XLNet_last512. The model ran for 13 epochs, and then it was early stopped. In another experiment, we trained the same architecture to predict judgment outcome on the last 512 tokens of ANALYSIS role sentences. We call this model as XLNet_last512_Analysis. The model ran for 12 epochs, and then it was early stopped. The model performance comparison are given in Table 7. As observed from the results, filtering the input text for the ANALYSIS role improves the prediction.

Model	Precision	Recall	F1
XLNet_last512	0.76	0.49	0.59
XLNet_last512_Analysis	0.71	0.55	0.62

Table 7: Judgment prediction Results

6. Conclusion and Future Directions

In this paper, we proposed a new corpus of legal judgment documents annotated with 13 different Rhetorical Roles. The corpus was created via crowdsourcing involving law students. We also proposed baseline models for automatic rhetorical role prediction in a legal document. For some of the roles, the model shows similar trends in predicting the roles as human annotators. Nevertheless, there is scope for further improvement and we have created a leaderboard for the task, so that researchers from community can contribute towards improving the RR prediction system. We also showed two applications of rhetorical roles: summarization and judgment prediction. For both the usecases use of rhetorical role helps to improve results. We have released the corpus and the baseline models and encourage the community to use these to develop other legal applications as well.

Acknowledgements

7. Bibliographical References

Bhatia, V. K. (2014). *Analysing genre: Language use in professional settings*. Routledge.

Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., and Ghosh, S. (2019a). A comparative study of summarization algorithms applied to legal case judgments. In *European Conference on Information Retrieval*, pages 413–428. Springer.

Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., and Wyner, A. (2019b). Identification of rhetorical roles of sentences in indian legal judgments.

Brack, A., Hoppe, A., Buschermöhle, P., and Ewerth, R. (2021). Sequential sentence classification in research papers using cross-domain multi-task learning.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. (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, July. Association for Computational Linguistics.

Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. (2021). MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Cohan, A., Beltagy, I., King, D., Dalvi, B., and Weld, D. (2019). Pretrained language models for sequential sentence classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Farzindar, A. and Lapalme, G. (2004). Letsum, an automatic legal text summarizing system.

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Grabmair, M., Ashley, K. D., Hwa, R., and Sweeney, P. M. (2011). Toward extracting information from public health statutes using text classification machine learning. In *Legal Knowledge and Information Systems*, pages 73–82. IOS Press.

Hachey, B. and Grover, C. (2006). Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

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

Kapoor, A., Dhawan, M., Goel, A., Arjun, T., Agrawal, V., Agrawal, A., Bhattacharya, A., Kumaraguru, P., and Modi, A. (2022). HLDC: Hindi Legal Documents Corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*. Association for Computational Linguistics.

Lagos, N., Segond, F., Castellani, S., and O?Neill, J. (2010). Event extraction for legal case building and reasoning. In *International Conference on Intelligent Information Processing*, pages 92–101. Springer.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders.

Malik, V., Sanjay, R., Guha, S. K., Nigam, S. K., Hazarika, A., Bhattacharya, A., and Modi, A. (2021a). Semantic Segmentation of Legal Documents via Rhetorical Roles. *CoRR*, abs/2112.01836.

Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., and Modi, A. (2021b). 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, August. Association for Computational Linguistics.

Maxwell, K. T., Oberlander, J., and Lavrenko, V. (2009). Evaluation of semantic events for legal case retrieval. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 39–41.

Moens, M.-F., Uyttendaele, C., and Dumortier, J. (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.

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

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

Russakovsky, O., Deng, J., Huang, Z., Berg, A. C., and Fei-Fei, L. (2013). Detecting avocados to zucchinis: what have we done, and where are we going? In *International Conference on Computer Vision (ICCV)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Saravanan, M., Ravindran, B., and Raman, S. (2007). Using legal ontology for query enhancement in generating a document summary. *Frontiers In Artificial Intelligence and Applications*, 165:171.

Saravanan, M., Ravindran, B., and Raman, S. (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*.

spaCy. (2021). spaCy Toolkit. `https://spacy.io/`.

Strickson, B. and De La Iglesia, B. (2020a). Legal Judgement Prediction for UK Courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, pages 204–209, Cambridge United Kingdom, March. ACM.

Strickson, B. and De La Iglesia, B. (2020b). Legal Judgement Prediction for UK Courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, pages 204–209, Cambridge United Kingdom, March. ACM.

Şulea, O.-M., Zampieri, M., Vela, M., and van Genabith, J. (2017). Predicting the law area and decisions of French Supreme Court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria, September. INCOMA Ltd.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2020). Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.

Tran, V., Nguyen, M. L., and Satoh, K. (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.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wyner, A., Mochales-Palau, R., Moens, M.-F., and Milward, D. (2010). Approaches to text mining arguments from legal cases. In *Semantic processing of legal texts*, pages 60–79. Springer.

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M.,

Feng, Y., Han, X., Hu, Z., Wang, H., et al. (2018). Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.