Artificial Intelligence and Law (2023) 31:53–90
https://doi.org/10.1007/s10506-021-09304-5

ORIGINAL RESEARCH

DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents

Paheli Bhattacharya, et al. ***[full author details at the end of the article]***

Accepted: 8 October 2021 / Published online: 13 November 2021

Abstract

The task of rhetorical role labeling is to assign labels (such as Fact, Argument, Final Judgement, etc.) to sentences of a court case document. Rhetorical role labeling is an important problem in the field of Legal Analytics, since it can aid in various downstream tasks as well as enhances the readability of lengthy case documents. The task is challenging as case documents are highly various in structure and the rhetorical labels are often subjective. Previous works for automatic rhetorical role identification (i) mainly used Conditional Random Fields over manually handcrafted features, and (ii) focused on certain law domains only (e.g., Immigration cases, Rent law), and a particular jurisdiction/country (e.g., US, Canada, India). In this work, we improve upon the prior works on rhetorical role identification by proposing novel Deep Learning models for automatically identifying rhetorical roles, which substantially outperform the prior methods. Additionally, we show the effectiveness of the proposed models over documents from five different law domains, and from two different jurisdictions—the Supreme Court of India and the Supreme Court of the UK. Through extensive experiments over different variations of the Deep Learning models, including Transformer models based on BERT and LegalBERT, we show the robustness of the methods for the task. We also perform an extensive inter-annotator study and analyse the agreement of the predictions of the proposed model with the annotations by domain experts. We find that some rhetorical labels are inherently hard/subjective and both law experts and neural models frequently get confused in predicting them correctly.

Keywords Rhetorical role labeling · Legal document segmentation · Court case documents · Hierarchical BiLSTM · Hierarchical BiLSTM CRF · BERT · LegalBERT

This manuscript is an extended version of our prior work: Bhattacharya *et al.*, *"Identification of Rhetorical Roles of Sentences in Indian Legal Judgments"*, International Conference on Legal Knowledge and Information Systems (JURIX), 2019.

1 Introduction

Rhetorical role labelling of sentences in a legal document refers to understanding what semantic function each sentence is associated with. Examples of these roles/semantic labels/themes are – facts of the case, arguments of the contending parties, the final judgement of the Court, and so on. Identifying the rhetorical roles of sentences in a legal case document can help in a variety of downstream tasks like semantic search (Nejadgholi et al. 2017), summarization (Saravanan et al. 2008; Farzindar and Lapalme 2004), case law analysis (Savelka and Ashley 2018) and so on.

However, legal case documents are highly varying in structure (Shulayeva et al. 2017; Bhattacharya et al. 2019a), and various themes often interleave with each other. For instance, the reason behind the judgment (Ratio of the decision) often interleaves with Precedents and Statutes. Hence it sometimes becomes difficult even for human experts to understand the intricate differences between the rhetorical roles. Hence, *automating* the identification of these rhetorical roles is a challenging task.

For supervised machine learning of the roles, it is important to develop a high quality gold standard corpus, capturing the rhetorical roles of sentences as accurately as possible. Different prior works attempting the task have constructed their own set of annotated documents (Nejadgholi et al. 2017; Saravanan et al. 2008; Savelka and Ashley 2018), but none of them report an extensive analysis of the annotation process. Apart from Inter-Annotator Agreement (IAA) scores, it is useful to do a qualitative analysis for understanding issues such as the amount of subjectivity associated to the rhetorical labels. In this paper, we perform a systematic annotation study and an extensive inter-annotator agreement analysis. We show that even legal experts find it difficult to distinguish some specific pairs of rhetorical labels, thus showing that some subjectivity is inherent in these labels.

Prior attempts to automate the identification of rhetorical roles of sentences in legal documents relied on hand-crafted features such as linguistic cue phrases indicative of a particular rhetorical role (Saravanan et al. 2008; Farzindar and Lapalme 2004; Savelka and Ashley 2018), the sequential arrangement of labels (Saravanan et al. 2008), and so on (see Sect. 2 for details). Some of these features, e.g., indicator cue phrases, are *largely dependent on legal-expert knowledge* which is expensive to obtain. Moreover, these features have always been developed for legal documents of a particular jurisdiction (e.g., High Court of Kerala, an Indian state (Saravanan et al. 2008), Canada (Nejadgholi et al. 2017; Farzindar and Lapalme 2004), US (Savelka and Ashley 2018), etc.), and features developed for documents of one legal jurisdiction do not always scale to other jurisdictions (Bhattacharya et al. 2019a). Also, the hand-crafted features developed in the prior works are often specific to one or a few domains/categories (e.g., Rent Control, Income Tax and Sales Tax in (Saravanan et al. 2008), Cyber crime and Trade secrets in (Savelka and Ashley 2018)). It has not been explored whether one can devise a set of features that works for documents across various domains and/or across various jurisdictions. Hence we can conclude that existing methods for this

task may not be robust. Additionally, the reliance on hand-crafted features may be very expensive, if it has to be conducted separately for different domains and jurisdictions.

Recently developed deep learning, neural network models do not require hand-engineering features, but are able to automatically learn the features, given sufficient amounts of training data. Such models have been shown to perform better in tasks like classification than methods using hand-crafted features. Also, since these models can automatically learn features from data, they have the ability of generalizing to any domain of law or jurisdiction, without the necessity of curating features by hand for each domain/jurisdiction. We attempt to prove this claim in our work.

In this paper, we explore four neural network models to automatically identify the rhetorical roles of sentences in legal documents–(1) a Hierarchical BiLSTM model (Hier-BiLSTM) (2) a Hierarchical BiLSTM/BiGRU model combined with Conditional Random Fields (Hier-BiLSTM/BiGRU-CRF), (3) Hierarchical BiLSTM / Hierarchical BiLSTM-CRF model with Attention (Hier-BiLSTM+Attn, Hier-BiLSTM-CRF+Attn), and (4) Transformer-based models Tf-BiLSTM/BiGRU-CRF. Furthermore, we experiment with several pre-trained language models including Bert (Devlin et al. 2019) and LegalBert (Chalkidis et al. 2020), in combination with the neural models stated above.

These models learn the sentence and tag sequences automatically, and there is no need for hand-crafting of features. We use these models for supervised classification across *seven rhetorical labels* (classes) and over legal case documents from two different jurisdictions—the Indian Supreme Court and the U.K. Supreme Court. The best performing neural model achieves a very good performance (Macro F-score of 0.821 for Indian documents and 0.600 for U.K. documents), substantially out-performing baseline methods that use hand-crafted features. Furthermore, we analyse the rhetorical roles predicted by our model, and correlate the observations from the predicted results with those on the inter-annotator study. We find that the subjectivity between certain pairs of labels (e.g., Ratio vs. Precedent) that is present among the human annotators is also reflected in the predictions by the algorithm for documents from both the jurisdictions. We also perform a cross-domain study wherein we investigate how well a model trained on the India dataset can label sentences from the UK dataset, and vice-versa.

Although there have been a few prior attempts towards using neural models for rhetorical role identification (Nejadgholi et al. 2017; Yamada et al. 2019), these prior works have considered only a few labels (e.g., binary classification between factual and non-factual sentences) and only one jurisdiction (see Sect. 2 for details). To the best of our knowledge, this is the first work on identifying rhetorical roles in legal documents that experiments on case documents from two different jurisdictions (Indian Supreme Court and U.K. Supreme Court). As such, this study is the first one (to our knowledge) that brings together (i) an extensive annotation study, (ii) several deep learning models and their variations for automating the task, and (iii) in-depth analyses of how the model predictions correlate with the labels assigned by domain experts.

In summary, the contributions of this work are as follows:

1. We develop two datasets (annotated by senior law students) for the task of automatic rhetorical role identification, using case documents from the Supreme Courts of India and UK. The datasets will be available upon request to the first author.
2. We conduct detailed studies of inter-annotator agreement over the datasets, which lead to insights about the subjectivity of some of the rhetorical roles.
3. We propose several neural models for automatic rhetorical role identification, which substantially outperform prior methods using hand-crafted features.
4. We also study the agreement of the labels predicted by the neural models with the labels assigned by domain experts. These experiments lead to a better understanding of human-human agreement and human-model agreement.
5. We analyse the performance of rhetorical labeling models trained on data from one jurisdiction over documents from some other jurisdiction. This analysis leads to important insights on cross-jurisdictional performance of such models.

It is to be noted that this work is an extension of our prior work (Bhattacharya et al. 2019b). The differences between the prior work and the present work are elaborated in Sect. 2.4.

2 Related work

In this section, we briefly discuss prior work about annotation studies on legal case documents, automatic identification of rhetorical roles, and applications of deep learning in the legal domain.

2.1 Annotation studies on legal case documents

Automatic labelling of the rhetorical role of sentences relies heavily on manual annotation. While papers that aim to automate the task of semantic labelling also perform an annotation analysis (Savelka and Ashley 2018; Shulayeva et al. 2017), other works focus on the process of annotation – developing a manual/set of rules for annotation, inter-annotator studies, curation of a gold standard corpus, and so on.

For instance, TEMIS, a corpus of 504 sentences, that were annotated both syntactically and semantically, was developed in Venturi (2012). This is a collection of heterogeneous legislative documents including legal acts such as national and regional laws, European directives, legislative decrees, etc., as well as administrative acts, such as ministerial circulars, decision, etc. The syntactic level of annotation includes dependency annotation and sentence splitting. For semantic level annotation, they use the FrameNet standards.

An in-depth annotation study and curation of a gold standard corpus for the task of sentence labelling can be found in Wyner et al. (2013), where assessor agreement was low for labels like Facts and Reasoning Outcomes. They annotated 20

documents by 3 law school students for the labels–'citation indices', 'legal facts', 'rationale', 'judgement', 'cause of action', and 'others'. They show that the annotators achieve a high IAA on those labels that are fairly straightforward, such as JudgeName, HearingDate, DecisionDate, CaseCitation. However, labels like Facts and Reasoning Outcomes show considerably low agreement.

LegalRuleML is emerging as an standard for representing the semantic contents of legal texts in XML. Wyner et al. (2017) studies how a corpus of legal instruments can be transformed into LegalRuleML, for the purpose of information retrieval. They use a partial set of rules of LegalRuleML (Permission, Obligation, Prohibition etc.) which associate with text annotations.

Towards automating the annotation task, Wyner (2010) discusses an initial methodology using NLP tools on 47 criminal cases drawn from the California Supreme Court and State Court of Appeals. They annotate legal case elements such as case name, case role, judge name, appellant counsel, attack term, weapon, cause of action, decision statement etc. GAZETTEER lists (to capture simple, unsystematic patterns) and JAPE rules (capture systematic, complex patterns for higher order annotation) are framed for the purpose. They follow an iterative process where rules are refined based on the desired level of annotation.

2.2 Automatic identification of rhetorical roles of sentences

There have been several prior attempts towards automatically identifying rhetorical roles of sentences in legal documents. Initial experiments for understanding the rhetorical/thematic roles in court case documents/judgements/case laws were developed as a part of achieving the broader goal of summarizing these documents (Farzindar and Lapalme 2004; Hachey and Grover 2006; Saravanan et al. 2008). For instance, LetSum Farzindar and Lapalme (2004) divides the text structure into five themes–Introduction, Context, Juridical Analysis and Conclusion. They work on Canadian case documents and leverage 'section titles' that the documents contain. They curate linguistic phrases based on these section titles indicating their inclusion in one of the five themes. Again, Saravanan et al. (2008) used Conditional Random Fields (CRF)(Lafferty et al. 2001) for the same task, considering seven rhetorical roles. They experimented on 200 Kerala High Court documents (Kerala is an Indian state) from 3 domains – rent control, income tax and sales tax. Labelling sentences as *facts* or *principles* using handcrafted features and the Multinomial Bayes Classifier was explored in Shulayeva et al. (2017).

Segmenting a document into functional parts (Introduction, Background, Analysis and Footnotes) and issue-specific parts (Analysis and Conclusion) was looked into by Savelka and Ashley (2018) on U.S. court documents using CRF with handcrafted features. In this approach, a first-level segmenter identifies the functional parts – they iteratively predict each label and remove it from the set for the next label prediction. A second-level segmenter which identifies the issue-specific parts, further breaks down the Analysis (from the functional part) into Analysis and Conclusion. They annotate 50 U.S. court documents from the domains of 'cyber crime' and 'trade secrets'. For automating, CRF with handcrafted features was used.

A method for identification of factual and non-factual sentences was developed in Nejadgholi et al. (2017) using the FastText classifier. They annotated 150 Canadian immigration case documents with sentences labelled as facts and non-fact. They trained word embeddings on a large legal corpus and used the the FastText classifier for classification. The end goal was to retrieve documents whose fact asserting sentences were similar to a given query. Yamada et al. (2019) use Bi-LSTM-CRF along with heading encoders for the task for rhetorical labeling of sentences in Japanese documents. Distinguishing "facts" and "legal principles" of cited cases in a legal document was studied in Shulayeva et al. (2017). They considered only those paragraphs of the document that contain at least one citation. For the task they annotate 50 common law reports available in the British and Irish Legal Institute (BAILII) website. Towards automating the task, they use Multinomial Bayes Classifier with hand-engineered linguistic features.

In another line of work, Walker et al. (2019) compared use of rule-based scripts (that require much less training data) with Machine Learning approaches for the rhetorical role identification task.

2.3 Application of deep learning in the legal domain

Deep Learning (DL) methods are increasingly being applied for several tasks in the legal domain (Zhong et al. 2020). For instance, DL has been applied to contract element extraction in Chalkidis and Androutsopoulos (2017). DL models have also been used for other applications such as crime classification in Wang et al. (2018) and Wang et al. (2019a), summarization in Liu and Chen (2019) and Bhattacharya et al. (2019a), judgement prediction in Chalkidis et al. (2019), estimating legal document similarity in Bhattacharya et al. (2020), and several other legal AI tasks.

As stated above, some prior works have also applied DL methods for the task of rhetorical role identification. For instance, word embeddings along with the FastText classifier were applied in Nejadgholi et al. (2017) for a binary classification of factual and non-factual sentences in a legal document. (Nejadgholi et al. 2017) show the benefit of using deep learning where they model the task as a binary classification problem (identifying factual and non-factual sentences) on a specialised domain of Immigration documents only. Similarly (Yamada et al. 2019) employed neural architectures for rhetorical role labeling of sentences.

Several transformer architectures like Bert (Devlin et al. 2019) and LegalBert (Chalkidis et al. 2020) have been shown to prove beneficial in many tasks such as legal judgement prediction (Chalkidis et al. 2019), prior-case retrieval (Shao et al. 2020), crime classification (Wang et al. 2019b) and many others. In this work, we explore the use of Bert and LegalBert for the task of rhetorical role segmentation.

It can be noted that none of these prior works have attempted the task of rhetorical role labels across multiple domains and multiple jurisdictions. In the present work, we demonstrate the effectiveness of our proposed models over court case documents from several domains, and two different jurisdictions (UK and India).

2.4 Present work as an extension of our prior work

This work is an extension of our previous work (Bhattacharya et al. 2019b). In our prior work, we had experimented on only the Indian Supreme Court case documents with two neural models, Hier-BiLSTM and Hier-BiLSTM-CRF. Also each of these two neural models had two variations–using randomly initializing word embeddings and using pretrained sentence embeddings. In this work, we have performed the following substantial extensions:

- *A new model based on attention mechanism* We propose a third neural model (Hier-BiLSTM+Attn / Hier-BiLSTM-CRF+Attn) by adding an Attention layer to the existing Hier-BiLSTM and Hier-BiLSTM-CRF architectures. We also experiment with Hier-BiGRU-CRF.
- *Experiments with pretrained word embeddings* Apart from randomly intializing the word embeddings (which was done in our prior work (Bhattacharya et al. 2019b)), we also use pretrained word embeddings for initializing the neural models. Specifically, we use (i) the legal domain-specific pretrained word embbedings Law2Vec, and (ii) the general Google News-based word embeddings. We show that it is almost always better to use Law2Vec word embeddings than to use random initialization or initialization with Google News word vectors.
- *Experiments with pretrained transformer embeddings* We explore the performance of pretrained Bert and LegalBert embeddings for the task of rhetorical role labeling.
- *Training transformer models for the task* We also propose transformer-based models that fine-tune Bert and LegalBert for the task.
- *Added a new dataset from the U.K. Supreme Court* Our prior work only considered documents from the Indian Supreme Court. In this work, we have added a completely new dataset from the U.K. Supreme Court. Similar to the Indian dataset, we report extensive Inter-Annotator Agreement studies on the U.K. dataset as well. Thus, in this work, we validate the robustness of the neural models to automate the task of rhetorical role labeling of sentences across two different jurisdictions—the Indian Supreme Court and the U.K. Supreme Court.

3 Datasets

In this paper, we consider legal judgments from (i) the Supreme Court of India, and (ii) the Supreme Court of the United Kingdom (U.K.). We describe each of the datasets in this section, as well as a comparison of the two datasets.

India Dataset We crawled 53, 210 documents in total from the website of Thomson Reuters Westlaw India (http://www.westlawindia.com)[1]. Westlaw assigns each document a legal domain, such as ‘Criminal’, ‘Constitutional’, etc.

[1] We use only the publicly available full text judgement. All other proprietary information had been removed before performing the experiments.

Table 1 Comparison of the two datasets used in this work

Property	India dataset	UK dataset
Number of documents	50	50
Total number of sentences (in all documents)	9,308	18,155
Average number of sentences per document	188	363
Average number of words per document	5,153	11,486
Average number of words per sentence	27	32

One of our objectives was to check if the methods perform well across case documents in various law domains or law categories. To this end, we calculated the frequency of these domains, chose the top 5 domains and randomly sampled 50 documents from these 5 domains in proportion to their frequencies.

Thus we have the following set of 50 documents from 5 law domains/categories—(i) Criminal—16 documents (ii) Land and property—10 documents (iii) Constitutional—9 documents (iv) Labour and Industrial—8 documents (v) Intellectual Property Rights—7 documents. All experiments reported in this paper are performed on these 50 case documents.

U.K. Dataset We crawled 742 documents in total, from the official website of the U.K. Supreme Court (https://www.supremecourt.uk/decided-cases/). No category-specific information is given on this website. Hence, we randomly sampled 50 documents for the experiments.

Splitting the documents into sentences We split each document into sentences using the SpaCy tool (https://spacy.io/). It can be noted that splitting a legal document into sentences is challenging due to frequent presence of abbreviations (Sanchez 2019). We observed SpaCy to do a reasonably good splitting (accuracy close to 90%, as judged by human annotators for a small subset of the documents), which agrees with observations in prior works (Nejadgholi et al. 2017).

Comparing the two datasets Table 1 shows a comparison of the two datasets (each containing 50 documents). We find that the UK Supreme Court documents are substantially longer than the Indian Supreme Court documents, having almost double the number of sentences/words per document on average. The average number of words per sentence is comparable for both the datasets. Some other differences in the style of writing UK Supreme Court documents and Indian Supreme Court documents will be stated in later sections.

4 Annotation process

This section details our annotation study – we describe the rhetorical roles/semantic labels, the annotation procedure, and analysis of inter-annotator agreement.

4.1 Annotation labels / rhetorical roles

Our annotators were three senior Law students from the Rajiv Gandhi School of Intellectual Property Law[2] which is one of the most reputed Law schools in India. The annotators are very familiar with not only Indian court case documents, but also with court case documents from UK courts of law.

Based on discussions with the annotators, we consider the following seven (7) rhetorical roles for sentences in a case document. Both the India and U.K. datasets were labeled with the same rhetorical roles.

1. *Facts (abbreviated as FAC)* This refers to the chronology of events that led to filing the case, and how the case evolved over time in the legal system (e.g., First Information Report at a police station, filing an appeal to the Magistrate, etc.)
2. *Ruling by Lower Court (RLC)* Since we are considering Supreme Court case documents, there were some judgements given by the lower courts (Trial Court, High Court, Tribunal, etc.) based on which the present appeal was made (to the Supreme Court). The verdict of the lower Court and the ratio behind the judgement by the lower Court was annotated with this label.
3. *Argument (ARG)* The present Court's discussion on the law that is applicable to the set of proven facts by weighing the arguments of the contending parties.
4. *Statute (STA)* Established laws referred to by the present court, which can come from a mixture of sources – Acts , Sections, Articles, Rules, Order, Notices, Notifications, Quotations directly from the bare act, and so on.
5. *Precedent (PRE)* Prior cases referred to by the present court.
6. *Ratio of the decision (Ratio)* Application of the law along with reasoning/rationale on the points argued in the case; Reason given for the application of any legal principle to the legal issue.
7. *Ruling by Present Court (RPC)* Ultimate decision / conclusion of the present Court following from the natural / logical outcome of the rationale

4.2 Annotation process

Now we describe the process of annotating the sentences in case documents with the rhetorical labels. As stated earlier in Sect. 3, each document was first split into sentences using the SpaCy tool (https://spacy.io/). Each such sentence was considered a unit and was marked with any one of the seven rhetorical role described above.

We followed a systematic way of annotation similar to Shulayeva et al. (2017) and Wyner et al. (2013). An annotation manual was developed in discussion with the annotators, containing descriptions and example sentences for each rhetorical role, along with other instructions (e.g., a label should be assigned to a full sentence and not a part of it, a sentence should have only one label, etc.). Initially, each annotator was asked to annotate 5 documents independently, i.e., without consulting each

[2] http://www.iitkgp.ac.in/department/IP.

other. Then we had a joint discussion with all the annotators to resolve any issues, and refined the manual if necessary. This process was followed iteratively for annotation of the 50 documents in each dataset.

Note that, we uniformly annotate sentences from both the India and UK datasets with the seven (07) rhetorical labels stated above. These rhetorical labels are different from the five *law domains/categories* from which documents of the India dataset are chosen (as was stated in Sect. 3).

4.3 Analysis of inter-annotator agreement

Since the rhetorical role of a sentence is subjective, such annotation tasks are usually performed with multiple annotators. As stated earlier, our annotators were three senior Law students. So it is important to study the level of agreement between the annotators.

To understand the Inter-Annotator Agreements (IAA), we perform two types of agreement computations—(i) Average IAA measure, where we calculate document-wise average F-Score and rhetorical-label wise F-Score, and (ii) Sentence-level agreement, where we compare the labels given by the three annotators for each individual sentence. We do these analyses to specifically understand for which labels annotators frequently disagree. The rest of this section describes these studies in detail.

4.3.1 Average IAA measure

There are different ways of measuring inter-annotator agreement. While some studies report Kappa measures (Artstein and Poesio 2008), other studies report Precision, Recall and F-Score (Wyner et al. 2013). In this paper, we report both Cohen's Kappa, F-Scores and Fleiss Kappa. Since we have three annotators (A_1, A_2 and A_3), we compute three sets of pairwise IAA (A_1, A_2), (A_2, A_3), (A_1, A_3), and then take the average of the three sets.

To find the average IAA across 50 documents for an annotator pair (A_i, A_j) according to F-measure, we consider the labels annotated by A_i as the *key annotations* and the labels annotated by A_j as the *response annotations*. We then compute the F-Score using the following formulae:

$$\text{Precision} = \frac{|\{\textit{labels given by } A_i\} \cap \{\textit{labels given by } A_j\}|}{|\{\textit{labels given by } A_j\}|} \tag{1}$$

$$\text{Recall} = \frac{|\{\textit{labels given by } A_i\} \cap \{\textit{labels given by } A_j\}|}{|\{\textit{labels given by } A_i\}|} \tag{2}$$

$$\text{F-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Table 2 Document-wise average inter-annotator agreement in terms of FScore and Cohen's Kappa of the 3 annotators A_1, A_2 and A_3 The values reported are averaged over all documents in a certain dataset (details in text)

Dataset	A_1 and A_2		A_2 and A_3		A_1 and A_3		Average among all 3 annotators	
	FScore	Kappa	FScore	Kappa	FScore	Kappa	FScore	Kappa
India	0.972	0.972	0.936	0.947	0.950	0.961	0.953	0.980
UK	0.886	0.909	0.915	0.927	0.916	0.940	0.915	0.929

Table 3 Label-wise average inter-annotator agreement of the 3 annotators in terms of F-score (described in text)

Label	ARG	RLC	STA	PRE	RPC	FAC	RATIO
India	0.884	0.809	0.877	0.881	0.978	0.978	0.973
UK	0.761	0.774	0.782	0.874	0.936	0.962	0.969

Cohen's Kappa between a certain annotator pair is calculated using the following formula, where p_o is the relative observed agreement among the annotators and p_e is the hypothetical probability of chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

We compute this value for each pair of annotators.

The results of the document-wise average IAA (averaged over the 50 documents in each dataset) are given in Table 2. We find that the IAA values are mostly above 0.9 for both F-score and Cohen's Kappa, thus denoting a high overall agreement among the three annotators. The average agreement is slightly higher for the India dataset than that for the U.K. dataset; this difference is possibly because our annotators are Indian law students, and they are likely to be more well-versed with Indian documents in comparison to U.K. documents. The Fleiss Kappa[3] for the India dataset is 0.872 and for the UK dataset is 0.927, which also denote excellent agreement.

It is also interesting to study the IAA for each individual rhetorical label. To calculate the average IAA for one of the labels L (out of the seven rhetorical labels) between two annotators (A_i, A_j) across the 50 documents, we use the following formulae:

$$\text{Precision}_L = \frac{|\{\textit{sentences labeled } L \textit{ by } A_i\} \cap \{\textit{sentences labeled } L \textit{ by } A_j\}|}{|\{\textit{sentences labeled } L \textit{ by } A_j\}|} \tag{5}$$

[3] https://en.wikipedia.org/wiki/Fleiss_kappa.

Table 4 Table showing the the sentence level agreement in the *India dataset* between the two annotators (A_2, A_3) who have the lowest IAA (0.936 and 0.947 as measured by the document-wise average F-Score and Cohen's Kappa respectively). Each value shows the percentage of sentences, out of all sentences in the dataset. High values in the diagonals are in bold. The relatively high non-diagonal values are underlined

	FAC	ARG	PRE	STA	RATIO	RLC	RPC
FAC	**23.14**	0.05	0	0.03	**0.43**	0.09	0
ARG	0.18	**8.83**	0.17	0.01	0	0	0
PRE	0	0.12	**15.31**	0	**0.50**	0	0
STA	0	0	0	**6.82**	0.13	0.02	0
RATIO	0.04	0.14	0.04	0.05	**37.59**	0.01	0
RLC	<u>0.50</u>	0.01	0	0	0.27	**3.16**	0
RPC	0.06	0	0	0	0.23	0	**2.81**

$$\text{Recall}_L = \frac{|\{\textit{sentences labeled L by } A_i\} \cap \{\textit{sentences labeled L by } A_j\}|}{|\{\textit{sentences labeled L by } A_i\}|} \tag{6}$$

$$\text{F-Score}_L = \frac{2 \times \text{Precision}_L \times \text{Recall}_L}{\text{Precision}_L + \text{Recall}_L} \tag{7}$$

The label-wise Average IAA F-score values among the three annotators over the 50 documents are shown in Table 3 (for both the India and U.K. datasets). Overall, the label-wise agreement values are also quite high among all the annotators. We find that for the India dataset, IAA is lowest for the label RLC (Ruling by Lower Court), which is also quite high (0.809). For the U.K. dataset, IAA is lower for the labels ARG (Argument) and RLC (Ruling by Lower Court). Agreements are high for RATIO (Ratio of the decision) and RPC (Ruling by Present Court) for both the datasets.

4.3.2 Sentence-level agreement

We have observed that IAA is lower for some specific labels such as RLC (Ruling by Lower Court). Now we delve deeper to understand why the agreements for some labels are low. To this end, we investigate the following type of questions—given that there is a low IAA for a label L (e.g., RLC), if annotator A_i has labelled a sentence as L, what label has another annotator A_j given the same sentence (owing to which there is a lower agreement for the label L)? To do such analyses, we perform a *sentence-level agreement study*.

We construct a *sentence-level agreement matrix* C (whose rows and columns are the labels) for two annotators A_i and A_j. An entry $C[x][y]$ of this matrix denotes the number of sentences which Annotator A_i labeled as L_x, but Annotator A_j labeled the *same* sentences as label L_y. Hence, the diagonal elements $C[x][y], x = y$ denote the number of sentences for which annotators A_i and A_j agree (give the same label). The

Table 5 Table showing the the sentence level agreement (%) in the *U.K. dataset* between the two annotators (A_1, A_2) who have the lowest IAA (0.886 and 0.909 as measured by the document-wise average F-score and Cohen's Kappa respectively). Each values shows the percentage of sentences, out of all sentences in the dataset. High values in the diagonals are in bold. The relatively high non-diagonal values are underlined

	FAC	ARG	PRE	STA	RATIO	RLC	RPC
FAC	**13.86**	0.03	0	0.02	0.23	0.03	0
ARG	0.11	**3.18**	0.05	0	<u>0.90</u>	0	0
PRE	0	0.01	**7.68**	0.03	<u>1.06</u>	0	0
STA	0.19	0	0.01	**6.24**	<u>0.65</u>	0	0
RATIO	0.14	0.07	<u>0.80</u>	0.23	**59.65**	0.15	0.04
RLC	0.1	0	0	0	0.08	**2.68**	0
RPC	0	0	0	0.01	0.29	0	**1.47**

Table 6 Distribution of sentences having various rhetorical labels, in the India and U.K. dataset

Labels	FAC (%)	ARG (%)	RATIO (%)	STA (%)	PRE (%)	RPC (%)	RLC (%)
India Dataset	23.13	9.00	38.63	6.88	15.65	2.79	3.63
U.K. Dataset	14.38	3.81	61.86	6.96	8.41	1.64	2.95

non-diagonal elements $C[x][y]$ where $x \neq y$ denote the number of sentences where the two annotators disagree (i.e., give different labels to the same sentence). We normalize the number of sentences in each cell of C by the total number of sentences in the corresponding dataset, and report the values in percentages. Table 4 shows this agreement matrix for the India dataset for the annotator pair (A_2, A_3) who have the *lowest IAA* as measured by the F-Scores (see Table 2). Similarly, Table 5 shows the sentence-level agreement matrix for the UK dataset for the annotator pair (A_1, A_2) who have the *lowest IAA* as measured by the F-Scores (see Table 2).

In both Table 4 and Table 5, the high values in the diagonal elements indicate that the annotators have a high overall agreement in general.

Among the non-diagonal elements, we underline the relatively high values indicating some disagreement between the annotators. We find that the label RATIO (the Court's reasoning for the final judgement) has some disagreement with other labels for both India and U.K. datasets, suggesting that the label is inherently subjective across jurisdictions. For instance, for the UK dataset, we observe in Table 5 that there is subjectivity among the label-pairs (PRE, RATIO). Since the Court's reasoning for the final judgement (RATIO) often depends on relevant prior cases/precedents (PRE), some legal expert may consider a sentence to be RATIO while some other expert may label the same sentence as a precedent (PRE).

4.4 Curation of gold standard for machine learning models

In the rest of this paper, we will apply Machine Learning models for automatically identifying rhetorical roles of sentences in case documents. For training and evaluation of such models, we need a gold standard dataset of sentences and their labels. To this end, for both India and U.K. datasets, we consider the *majority opinion* of the 3 annotators for a particular sentence as the gold standard label of that sentence. There was a clear majority verdict regarding the label (rhetorical role) of each sentence.

There are 9, 308 sentences in the India dataset and 18, 155 sentences in the U.K. dataset. Table 6 shows the distribution of sentences having various rhetorical roles in the two gold standard datasets. It is seen that the distribution of sentences across labels is similar for both datasets. For instance, the label RATIO has the maximum fraction of sentences and RPC (Ruling by Present Court) has the least fraction for both datasets. There are some differences as well, e.g., the UK documents have more discussion on the reasoning for final judgement (RATIO) while the Indian documents have more description of the facts (FAC) and precedents (PRE).

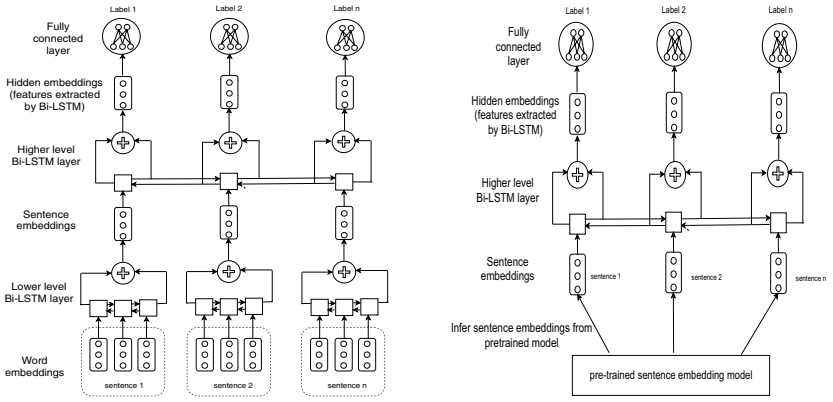
5 Methods for automatically identifying rhetorical roles

Now we describe our efforts towards automating the task of identifying rhetorical roles of sentences in a legal document. We treat this problem as a 7-class sequence labeling problem, where supervised Machine Learning models are used to predict one label (rhetorical role) for every sentence in a document. We have used statistical models (CRFs), deep neural models (BiLSTMs) and their combination for this task.

5.1 Baseline: CRF with handcrafted features

As stated in Sect. 2, this is the approach adopted in most prior works for identification of rhetorical roles of sentences in legal documents. Each document is treated as a sequence of sentences. Some dependencies exist in the corresponding sequence of labels; e.g., RLC usually follow FAC, RPC is always the end label, etc. Conditional Random Fields (CRFs) Lafferty et al. (2001) can be used to model such sequences, since they consider both *emission scores* (probability of a label given the sentence) and *transition scores* (probability of a label given the previous label) while generating the label sequence.

To implement the baseline approaches (Saravanan et al. 2008; Savelka and Ashley 2018), we represent each sentence as a vector of all features stated in these works—parts-of-speech tags (used in Savelka and Ashley (2018)), layout features (used in both Saravanan et al. (2008) and Savelka and Ashley (2018)), presence of cue phrases (used in Saravanan et al. (2008)), and occurrence of named entities like *"Supreme Court"*, *"High Court"* in the sentence (used in Saravanan et al. (2008)). The CRF works on these vectors to predict the labels (rhetorical roles).



(a) Randomly initialized word embeddings **(b)** Pretrained sentence embeddings

Fig. 1 Neural Model Hier-BiLSTM with randomly initialized word embeddings and pretrained sentence embeddings

Thus, we consider three baseline approaches: (1) CRF using the features of Saravanan et al. (2008; 2) CRF using the features of Savelka and Ashley (2018); and (3) CRF using a combination of features from both Saravanan et al. (2008) and Savelka and Ashley (2018).

5.2 Neural model 1: Hierarchical BiLSTM classifier

Handcrafting features is a tedious task and requires domain expertise. Also, it may not be possible to capture all features necessary for the task. Some features may be latent and it is difficult to perfectly encode them through manual expertise. Deep Learning models, like Bi-LSTMs, have been shown to be useful in such scenarios which require features to be automatically extracted.

We use a hierarchical BiLSTM (Bi-directional Long Short Term Memory) architecture (Graves et al. 2005) to automatically extract features for identifying the rhetorical roles. Figure 1 shows schematic diagrams of our models (two variations, as described below). Most variations of these models use two BiLSTM layers, a lower-level layer and a higher-level layer.

These models require us to feed the sequence of *sentence embeddings* to the BiLSTM, which returns a sequence of feature vectors. The BiLSTM model needs some initialization of the sentence embeddings, with which learning can start. We try several methods for obtaining sentence embeddings, as follows:

(1) *Using randomly initialized word embeddings* (Figure 1a): In this architecture, each word (represented as a vector/embedding) of a sentence is randomly initialized. It is passed through a lower-level BiLSTM layer, which combines the word representations to get the sentence representation/embedding. A higher-level

BiLSTM layer refines these sentence embeddings that capture the contextual informations. These embeddings are expected to capture all the hidden features that are otherwise difficult to encode through handcrafted features.

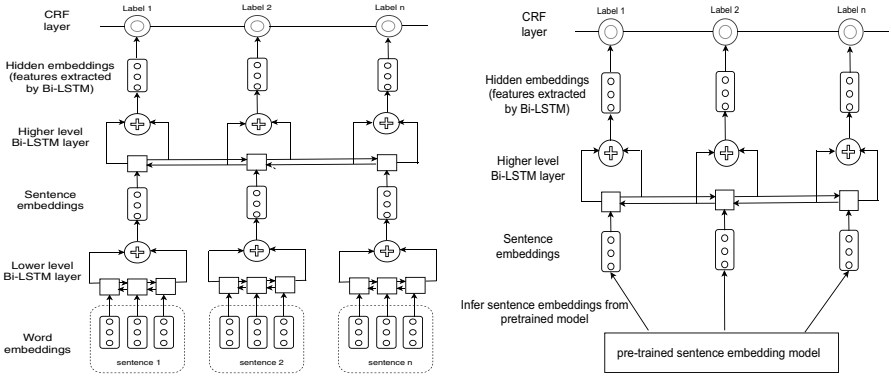
(2) *Using pretrained word embeddings* Instead of randomly initializing the word embeddings (as stated above), it is possible to initialize them using informative embeddings that have been pretrained on some text. We use two such pretrained word embeddings—(i) Google News embeddings[4] which contains 300-dimensional word vectors for 3 million words and phrases, learned on the Google News corpora, and (ii) Law2Vec embeddings[5] which contains 100-dimensional and 200-dimensional word embeddings for 492 million words, trained on legislative documents from the U.K., Europe, Canada, and many other jurisdictions. While Law2Vec embeddings are legal-context aware, Google News embeddings represent general concepts of the world.

(3) *Pretrained sentence embeddings* (Figure 1b) The two versions described earlier take as input word embeddings which are converted to sentence embeddings by the lower-level BiLSTM layer. In this variant, we directly input *pretrained sentence embeddings* to the model. Thus we no longer need the lower-level BiLSTM layer. Rather, the pretrained sentence embeddings are directly passed through the higher-level BiLSTM layer (actually the only BiLSTM layer in this model variant) to enrich the embeddings by incorporating contextual and other hidden features. We consider three kinds of pretrained sentence embeddings:

(3a) *Pretrained embeddings from a Sent2Vec model* Sent2Vec (Pagliardini et al. 2018) is an unsupervised model for learning sentence embeddings. It is as an extension of the Continuous Bag-of-Words model (C-BOW) of word-2vec. The sentence embeddings are formed by averaging both unigram embeddings and n-grams embeddings that constitute the sentence. For learning these embeddings the entire sentence is considered as the context window. The training is done to predict the words in the sentence, where possible class labels are all vocabulary words and the entire sentence is the context.

We train a *sent2vec* model over a large set of legal court case documents, to construct a sentence embedding model. To this end, we use the full set of case documents collected from the Indian and UK Supreme Courts (as described in Sect. 3, we collected more than 53K documents from the Indian Supreme Court and 700 documents from the UK Supreme Court). We ensure that the sent2vec model is *not* trained on the set of 50 test documents over which rhetorical role labeling experiments are being performed. The sent2vec model is trained on the rest of the documents. Note that we train two separate sent2vec models, one over Indian case documents, and the other over UK case documents. Once the sentence

[4] Available from https://code.google.com/archive/p/word2vec/.
[5] Avialable from https://archive.org/details/Law2Vec.



(a) Randomly initialized word embeddings **(b)** Pretrained sentence embeddings

Fig. 2 Neural Model Hier-BiLSTM-CRF with randomly initialized word embeddings and pretrained sentence embeddings

embedding model is trained, we use this model to directly infer the embeddings of the input sentences (of the 50 test documents).

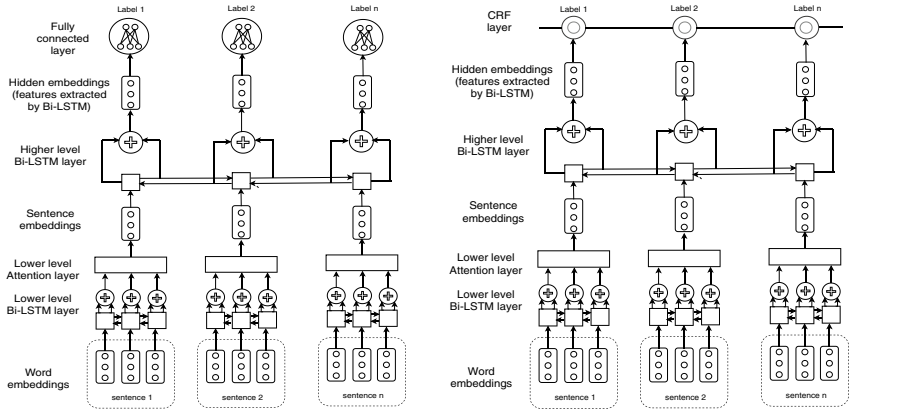
(3b) *Pretrained embeddings from Bert* Bert (Devlin et al. 2019) is a transformer model that has been trained on the BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers) for the Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks. Using this trained BERT model we can infer embeddings of sentences and words. In this work, we use the publicly available bert-base-uncased model[6] to infer the embeddings of the input sentences (of the 50 test documents) for both the India and UK datasets.

(3c) *Pretrained embeddings from LegalBert* LegalBert (Chalkidis et al. 2020) is a BERT model trained on legal documents – legislation from EU and UK, cases from the European Court of Justice (ECJ), European Court of Human Rights (ECHR), and various courts across the USA as well as US contracts, on the same tasks (MLM and NSP) as Bert. LegalBert induces legal domain-specificity to the original Bert model that was otherwise trained on general documents. We use the publicly available legal-bert-base-uncased model[7] to infer the embeddings of the input sentences (of the 50 test documents) for both the India and UK datasets.

In all the three variations described above, the sentence vectors coming out of the higher-level BiLSTM layer are then passed to a *feed-forward network* that generates probability scores for each label for each sentence. We consider a sentence to have that label for which the predicted probability score is the highest.

[6] Available at https://huggingface.co/bert-base-uncased.

[7] https://huggingface.co/nlpaueb/legal-bert-base-uncased.



(a) Hier-BiLSTM with Attention **(b)** Hier-BiLSTM-CRF with Attention

Fig. 3 Neural models Hier-BiLSTM and Hier-BiLSTM-CRF with attention

5.3 Neural model 2: Hierarchical BiLSTM CRF classifier

The probability scores generated by the BiLSTM models described above do *not* take into account label dependencies, and thus can be regarded as simple *emission scores*. To enrich the model further, we deploy a Conditional Random Field (CRF) on top of the Hierarchical BiLSTM architecture, as shown in Fig. 2. This CRF is fed with the feature vectors generated by the higher-level BiLSTM.

For this Hier-BiLSTM-CRF model as well, we experiment with all the three variations of sentence embeddings as described above—(i) randomly initialized embeddings (shown in Fig. 2a), (ii) pre-trained word embeddings, and (iii) pre-trained sentence embeddings trained over a large set of documents, i.e., Sent2Vec, Bert and LegalBert (shown in Fig. 2b).

5.4 Neural model 3: Hierarchical BiLSTM and Hierarchical-BiLSTM-CRF classifier with Attention

The *attention mechanism*, proposed by Bahdanau et al. (2014), has been found to be very effective in computing a good sequence representation of a given piece of text (e.g., a sentence) by paying attention to important words in the text (Galassi et al. 2020).

In the attention mechanism, a context vector c for understanding the importance of words in the given text (e.g., a sentence that has to be classified), is learned. The input X (a sentence in this task) consisting of a sequence of word embeddings $X = [x_1, \cdots, x_n]$ (where x_i is the embedding of the word w_i in the sentence X)[8] is

[8] The word embeddings x_i can be obtained using random initialization or Law2Vec or Google News embeddings, as discussed earlier in Sect. 5.2.

passed through a feed-forward network with *tanh* activation to get an activation sequence $[h_1, \cdots, h_n]$. We perform dot product of the learned context vector c with each h_t, to get a scalar score s_t. The set of scores $[s_1, \cdots, s_n]$ thus obtained is converted to a set of probabilities $[p_1, \cdots, p_n]$ using the standard *softmax* operation The final sentence representation of X is obtained as $X_s = p_1 x_1 + \cdots + p_n x_n$. A detailed explanation of the attention mechanism and its utility can be found in Galassi et al. (2020).

We incorporate this attention mechanism in the neural models Hier-BiLSTM and Hier-BiLSTM-CRF for rhetorical role labeling. Figure 3 shows the two model architectures after incorporating the attention mechanism. In both the models, the attention mechanism is coupled with the lower-level BiLSTM layer, so that the sentence representations learned in this layer have already attended to important words.

We experiment with the two variants of initializing word embeddings—random initializing and using pretrained word embeddings (as described earlier). The higher-level BiLSTM layer remains unaltered and learns the contextual information as before. The final classification is either by a fully connected layer (Hier-BiLSTM with Attention, as shown in Fig. 3a) or a CRF layer (Hier-BiLSTM-CRF with Attention, as shown in Figure 3b).

5.5 Neural model 4: Hier-BiGRU-CRF with pretrained sentence embeddings

This model is very similar to the Hier-BiLSTM-CRF model described earlier, with the only difference being that we use Bidirectional Gated Recurrent Units (Bi-GRUs) in place of Bi-LSTMs. The difference between GRU and LSTM is that the GRU lacks the *forget gate* and therefore has fewer parameters than LSTMs. GRUs have been shown to exhibit similar or better performance in certain tasks (Chung et al. 2014).

In this model, we replace the BiLSTM layer of Hier-BiLSTM-CRF with a BiGRU layer, to form the context-aware feature vectors/embeddings. These are then fed to the CRF layer for the final classfication.

Note that LSTMs can be replaced with GRUs also in the other models discussed above. As we will see later in Sect. 7, the Hier-BiLSTM-CRF architecture utilising pretrained sentence embeddings performs the best on the India dataset. Hence we specifically apply GRUs to create a variation (Hier-BiGRU-CRF) of this model.

5.6 Neural model 5: Tf-BiLSTM-CRF

With the advent of transformer models such as BERT, the transfer learning paradigm has gained popularity in NLP. These deep neural architectures are huge, having millions of parameters. Training these models from scratch on a small dataset like ours would result in overfitting. Therefore, we use the pre-trained models that were trained on huge datasets as a starting point. We then further train the model on our relatively smaller dataset, which is called *model fine-tuning*.

We fine-tune the transformer (Tf) models, Bert and LegalBert, for this task as follows – the bottom layer of the Tf-BiLSTM-CRF contains the pretrained transformer

models (Bert or LegalBert). Sentences are fed into this layer to get the Tf-embeddings. Similar to Hier-BiLSTM-CRF, the Tf-embeddings are passed on to the higher-level BiLSTM layer to obtain context-aware embeddings. Finally, the label classification is done using a CRF layer.

Note that, in these models, we use the pre-trained Bert and LegalBert models and fine-tune them, i.e., train them partially on the rhetorical labeling task. For this fine-tuning, we keep the weights of initial layers of the model frozen and *re-train the two higher layers.*

In this section, we have discussed several neural models for automatically identifying rhetorical roles of sentences, along with some baseline methods that relied on handcrafted features. In the subsequent sections, we will compare the performance of all these methods on the two datasets.

6 Experimental details

This section states some details about the experimental setup used by us.

6.1 Pretrained sentence embeddings using Sent2vec

As stated earlier in Sect. 5.2, we learn pretrained sentence embeddings of the two datasets using the algorithm *sent2vec* (Pagliardini et al. 2018). To this end, for the India dataset, we use 53K Indian Supreme Court case documents. Similarly, we learn pretrained sentence embeddings of the U.K. dataset using 742 documents from the UK Supreme Court. The sets of documents used to learn pretrained sentence embeddings are *disjoint* from the two test-sets of 50 documents each. We use the same preprocessing steps (using Spacy) before applying *sent2vec* (Pagliardini et al. 2018) to learn the sentence embedding model for both the datasets. The embedding dimension was 200 and other hyperparameters set to default.

6.2 Hyperparameters of the neural models

We use 100-dimensional word embeddings for random initialization of the models (wherever applicable). Also we use 100-dimensional and 200-dimensional Law2Vec word embeddings (results reported separately) and 300-dimensional Google News word embeddings. The sentence embeddings have a dimension of 200 across all experiments, except the ones involving the Bert models, where the embedding dimension is 768. Every neural model is trained for 200 epochs, using a learning rate of 0.001 for the India dataset and 0.01 for the U.K. dataset. These hyper-parameters were decided based on the model performance on the validation set. A possible reason for the different learning rates is because of the wide variation in the size of the two datasets. The dropout is 0.5 and regularization is 0.0005 for all models. The batch size is 32 for all experiments except in Tf-BiLSTM-CRF, where the batch size is 1. We use HuggingFace Pytorch-transformers for Bert and Legal-Bert. AdamW (Loshchilov and Hutter 2017) is used for optimizing the parameters of the model.

Table 7 Results of the baseline methods and all neural models for the task of rhetorical role labelling in India and U.K. datasets. The best values for each dataset are in bold

Model	Variant	India dataset			UK dataset		
		P	R	F	P	R	F
Baselines	Features in Saravanan et al. (2008)	0.414	0.331	0.405	0.455	0.388	0.319
	Features in Savelka and Ashley (2018)	0.458	0.42	0.325	0.414	0.345	0.347
	Features in Saravanan et al. (2008) +Savelka and Ashley (2018)	0.507	0.436	0.435	0.452	0.374	0.383
Hier-BiLSTM	Random Init Word Emb [we = 100, se = 200]	0.536	0.525	0.523	0.468	0.395	0.392
	Law2Vec Init Word Emb [we = 100, se = 200]	0.558	0.541	0.536	0.442	0.421	0.414
	Law2Vec Init Word Emb [we = 200, se = 200]	0.572	0.548	0.542	0.457	0.414	0.419
	Google Init Word Emb [we = 300, se = 200]	0.551	0.494	0.508	0.447	0.399	0.404
	Random Init Word Emb+Attn [we = 100, se = 200]	0.541	0.518	0.531	0.472	0.401	0.408
	Law2Vec Init Word Emb+Attn [we = 100, se = 200]	0.572	0.525	0.541	0.439	0.418	0.407
	Law2Vec Init Word Emb+Attn [we = 200, se = 200]	0.59	0.538	0.548	0.462	0.43	0.428
	Google Init Word Emb+Attn [we = 300, se = 200]	0.568	0.512	0.519	0.441	0.408	0.411
	Pretrained sent emb (Sent2Vec) [se = 200]	0.817	0.785	0.797	0.528	0.442	0.468
	Pretrained sent emb (bert) [se = 768]	0.590	0.508	0.525	0.535	0.512	0.510
	Pretrained sent emb (legal-bert) [se = 768]	0.568	0.513	0.518	0.502	0.415	0.431

Table 7 (continued)

Model	Variant	India dataset			UK dataset		
		P	R	F	P	R	F
Hier-BiLSTM-CRF	Random Init Word Emb [we = 100, se = 200]	0.652	0.552	0.578	0.492	0.388	0.416
	Law2Vec Init Word Emb [we = 100, se = 200]	0.618	0.581	0.588	0.466	0.428	0.43
	Law2Vec Init Word Emb [we = 200, se = 200]	0.620	0.59	0.591	0.468	0.417	0.431
	Google Init Word Emb [we = 300, se = 200]	0.592	0.541	0.558	0.463	0.400	0.415
	Random Init Word Emb+Attn [we = 100, se = 200]	0.630	0.574	0.581	0.516	0.404	0.421
	Law2Vec Init Word Emb+Attn [we = 100, se = 200]	0.622	0.593	0.594	0.502	0.401	0.431
	Law2Vec Init Word Emb+Attn [we = 200, se = 200]	0.615	0.582	0.587	0.498	0.421	0.438
	Google Init Word Emb+Attn [we = 300, se = 200]	0.612	0.502	0.571	0.494	0.407	0.433
	Pretrained sent emb (Sent2Vec) [se = 200]	**0.839**	**0.81**	**0.821**	0.533	0.476	0.491
	Pretrained sent emb (bert-base) [se = 768]	0.641	0.541	0.507	0.552	0.499	0.508
	Pretrained sent emb (legalbert-base) [se = 768]	0.601	0.502	0.521	0.513	0.425	0.448
Hier-BiGRU-CRF	Pretrained sent emb (Sent2Vec) [se = 200]	0.779	0.757	0.762	0.572	0.502	0.514
	Pretrained sent emb (bert-base) [se = 768]	0.687	0.558	0.596	0.520	0.514	0.501
	Pretrained sent emb (legalbert-base) [se = 768]	0.606	0.514	0.531	0.456	0.360	0.381
Tf-BiLSTM-CRF (2 layers of Bert/LegalBert Finetuned)	Bert-BiLSTM-CRF	0.688	0.615	0.635	0.653	0.567	0.589
	LegalBert-BiLSTM-CRF	0.710	0.635	0.655	**0.653**	0.587	**0.600**

We report macro-averaged Precision (P), Recall (R) and F1 (F) values. Here '*we*' implies word embedding dimension, '*se*' implies sentence embedding dimension

6.3 Evaluation setup and metrics

For both the India and U.K. datasets, we perform *5-fold cross-validation* to evaluate the models. Since we have documents from 5 legal domains in the India dataset (as was stated in Sect. 3), we ensure that in each fold there is at least one document from each domain in the train and test sets.

For a particular sentence, the label (rhetorical role) predicted by a model is considered to be correct, if it matches with the label assigned by the majority opinion of the human annotators (as described in Sect. 4.4).

Since the task of rhetorical role labelling is a multi-class classification problem, we calculate the macro-average Precision, Recall and F1 scores. These scores are calculated as the arithmetic means of individual classes' Precision, Recall and F1 scores.

7 Results and analysis

The performances of all the models for rhetorical role labeling of sentences are stated in Table 7, for both the India and U.K. case documents. The table is divided into five parts – the first part stating the performances of the baseline methods, the second part stating the performances of different variants of the Hier-BiLSTM model, the third part stating the performances of different variants of the Hier-BiLSTM-CRF model, the fourth part containing the performance of the Hier-BiGRU-CRF model and finally the performance of the transformer models (Tf-BiLSTM-CRF) Bert and Legal-Bert.

We first note that the neural models Hier-BiLSTM,Hier-BiLSTM-CRF,Hier-BiGRU-CRF and Tf-BiLSTM-CRF generally perform much better than the baseline methods employing CRF over handcrafted features. While the best F-score achieved by the baseline methods is 0.435 for the India dataset and 0.383 for the UK dataset, the best F-score achieved by the neural models is substantially higher (0.821 for the India dataset and 0.600 for the UK dataset). These results show that the latent features learnt by the neural models are much better at capturing the rhetorical labels than the handcrafted features used in the prior works of Saravanan et al. (2008) and Savelka and Ashley (2018). We now analyse in details the performance of the different neural models and their variations.

7.1 Analysis of the different neural models and their variations

In this section, we will discuss and compare the different neural models we have explored for the task of rhetorical role labelling of sentences in the India and U.K. documents.

7.1.1 Effect of using different word embeddings for initialization

The first four rows under each of the Hier-BiLSTM and Hier-BiLSTM-CRF parts of Table 7 show the performances obtained using different initializations (that were described in Sect. 5.2). We find that it is always more effective to use Law2Vec pretrained word embeddings than randomly initialized word embeddings or Google News pretrained word embeddings. Since the task is specific to legal documents, Law2Vec-based initialization make the model aware of the legal domain through the word embeddings. A notion of legal knowledge is inherent in the Law2vec embeddings since it has been trained specifically on legal documents. On the other hand, Google News embeddings have been trained on a general domain of news documents. Hence, the model is not legal-domain aware. Infusing domain-specific knowledge during pretraining is beneficial for neural models meant to address domain-specific tasks.

Also, we find that the word embedding dimension affects the performances. As we can see from Table 7, Law2Vec with 200-dimensional word embeddings perform better than 100-dimensional word embeddings.

7.1.2 Effect of using attention over different word embedding initializations

The next four rows under each of the Hier-BiLSTM and Hier-BiLSTM-CRF parts of Table 7, show the results of using the attention mechanism over different word embeddings. Attention has been shown to be useful in a wide variety of applications (Galassi et al. 2020). In the task of rhetorical role labeling as well, we find that using an attention mechanism improves performance in most cases. For instance, when using 200-dimensional Law2Vec embeddings for initialization (Law2Vec Init Word Emb [we = 200, se = 200], where '*we*' indicates the word embedding dimension, and '*se*' indicates the sentence embedding dimension), the F-Score for India and U.K. documents are 0.542 and 0.419 respectively. Whereas, when we use attention with the same initialization (Law2Vec Init Word Emb + Attn [we = 200, se = 200]), the F-Scores increase to 0.548 and 0.428 for India and U.K. documents respectively. A similar trend can be noticed for most of the other word embedding initialization settings as well.

7.1.3 Effect of using CRF with Hier-BiLSTM

Comparing the Hier-BiLSTM part and the Hier-BiLSTM-CRF part of Table 7, we can observe that using a Conditional Random Field (CRF) on top of the hierarchical BiLSTM model improves the performance slightly for most of the initialization settings. The improvement being small can be attributed to the fact that our documents consist of large sequences (average of 200 sentences per document), and we have relatively few documents; thus the CRF is unable to learn the transition scores properly. Training the models over more documents is likely to increase the improvement due to use of CRF.

7.1.4 Effect of using pretrained sentence embeddings

The last rows of the Hier-BiLSTM and Hier-BiLSTM-CRF parts of Table 7 show the performance of using pretrained sentence embeddings (Sent2Vec, Bert, Legal-Bert). For the Hier-BiLSTM model, the best result obtained using pretrained *word embeddings* (F-score **0.548** on the India dataset using Law2Vec Init word emb + Attn [we = 200, se = 200]) significantly improves to an F-score of **0.797** using pretrained sent2vec-based *sentence embeddings*. Similarly, Hier-BilSTM-CRF sees an improvement from F-score of **0.594** on the India dataset while using word embeddings (Law2Vec Init word emb + Attn [we = 100, se = 200]) to a significantly higher F-score of **0.821** using pretrained sent2vec-based sentence embeddings. This increase in performance is possibly due to the fact that the sentence embedding model has been pretrained on a huge collection of 53K Indian Supreme Court case documents. As a result of this pretraining, the neural models were aware of the language structure and dynamics of the Indian legal documents.

For U.K. documents also there is improvement in performance of both Hier-BiLSTM and Hier-BiLSTM-CRF by using pre-trained sentence embeddings, as compared to word embeddings. In Hier-BiLSTM, we find that pretrained sentence embeddings from the Bert model gives better results (F-Score of 0.510) when compared to the best word embedding intialization model (F-score of 0.428 using Law2Vec Init Word Emb+Attn [we = 200, se = 200]). In Hier-BiLSTM-CRF, when pretrained sentence embeddings are used instead of pretrained word embeddings, the performance improves from a F-score of **0.438** (using Law2Vec Init word emb + Attn [we = 200, se = 200]) to a F-score of **0.491** (using pretrained sent2vec sentence embeddings) and **0.508** (using pretrained bert embeddings) over the UK documents.

Note that the improvement when using sent2vec based sentence emebddings is much smaller for UK documents than what was noted for Indian documents. The reason behind this smaller improvement while using pretrained sentence embeddings for the UK dataset is as follows. For the U.K. dataset, we have only 742 U.K. Supreme Court case documents (in contrast to 53K Indian case documents), which is probably not sufficient for the unsupervised sent2vec algorithm to understand the language model of U.K. legal documents.

Using Bert and Legal-Bert sentence embeddings with Hier-BiLSTM, Hier-BiLSTM-CRF and Hier-BiGRU-CRF models also shows substantial improvement over that using word embeddings. Interestingly, bert-base embeddings perform better than legal-bert embeddings in most of the cases.

For the India dataset, we find that the pretrained bert-base and legal-bert embeddings could *not* perform as well as the sent2vec pretrained embeddings (which achieves the best performance of F-score 0.821). For the UK dataset, we observe that bert-base embeddings perform better than (or comparable to) the sent2vec-based sentence embeddings, while legal-bert embeddings perform worse than the sent2vec embeddings.

Table 8 Performance of the best models (for the India dataset: Hier-BiLSTM-CRF with pretrained Sent-2Vec embeddings; for the UK dataset: LegalBert-BiLSTM-CRF) on specific labels, in terms of F-Score

Dataset ↓ Labels →	FAC	ARG	RATIO	STA	PRE	RPC	RLC
India	0.839	0.592	**0.924**	0.722	0.854	0.900	0.812
UK	0.778	0.393	**0.854**	0.617	0.479	0.603	0.412

Values in bold and underline respectively indicate the label which is the most accurately and inaccurately predicted by the best models

7.1.5 Effect of using BiGRU and BiLSTM

From the performances reported in Table 7, we find that for the India dataset, using sent2vec-based pretrained sentence embeddings with Hier-BiLSTM-CRF performs the best (F-score 0.821), while using the same embeddings with Hier-BiGRU-CRF gives the second-best performance (F-score 0.762). However, for the other bert-based sentence embeddings, Hier-BiGRU-CRF performs better compared to Hier-BiLSTM-CRF.

For the UK dataset, Hier-BiGRU-CRF with sent2vec pretrained embeddings (F-score 0.514) performs better than Hier-BiLSTM-CRF with the same embeddings (F-score 0.491). But when bert and legal-bert sentence embeddings are used, the performance of Hier-BiLSTM-CRF is much better than that of Hier-BiGRU-CRF.

A probable reason for an inferior performance of BiGRU models on the UK dataset (except while using Sent2vec pretrained sentence embeddings) is that GRUs are more suitable for shorter sequences. From Table 1 we find that the average length of a document (in terms of number of sentences) is 188 for the India dataset and 363 for the UK dataset. UK documents being longer, GRUs fail to perform as well as LSTMs. Additionally, LSTMs have more parameters compared to GRUs which may contribute to its better performance for lengthier documents.

7.1.6 Performance of fine-tuned transformer models

The last part of Table 7 shows the performance of fine-tuning transformer models, Bert and Legal-Bert, for the task. In both cases, we fine-tune the last 2 layers of the transformer model. We observe that it is always beneficial to fine-tune the Bert and LegalBert models for the task (F-Score of Bert = 0.635 and LegalBert = 0.655 for India dataset; F-Score of Bert = 0.589 and LegalBert = 0.600 for UK dataset) than using the frozen / pretrained embeddings in Hier-BiLSTM-CRF (F-Score of Bert = 0.507 and LegalBert = 0.521 for India dataset; FScore of Bert = 0.508 and Legal-Bert = 0.448 for UK dataset).

We also observe that the transformer models could *not* outperform Hier-BiLSTM-CRF using pretrained sent2vec embeddings for the India dataset. However, for the UK dataset, LegalBert-BiLSTM-CRF shows the best performance, and is the overall best model for the task of rhetorical role labelling of sentences on the UK dataset, achieving an F-Score of 0.6. This difference is possibly because for the UK

Table 9 F-score of the Hier-BiLSTM-CRF model, for the different labels, and for each domain of law, for the India dataset

	FAC	ARG	Ratio	STA	PRE	RPC	RLC	*Macro average (for each domain)*
Constitutional	0.903	0.659	0.909	0.832	0.904	0.857	0.85	0.845
Labour & Industrial Law	0.776	0.505	0.929	0.423	0.728	0.783	0.681	0.689
Criminal	0.836	0.567	0.945	0.689	0.891	0.917	0.865	0.816
Land & Property	0.847	0.624	0.908	0.841	0.845	0.98	0.778	0.832
Intellectual property	0.832	0.607	0.927	0.824	0.901	0.964	0.886	0.849
Macro average (for each rhetorical role)	0.839	0.592	0.924	0.722	0.854	0.900	0.812	–

The last row indicates the macro-average F-score for each of the seven labels (rhetorical roles). The last column indicates the macro-average F-score for each of the five domains of law

dataset, we do not have a large corpus like the India dataset to train the sent2vec model. In such scenarios, use of pretrained transformer models (with suitable fine-tuning) actually proves beneficial.

7.2 Analysis of the best performing models

We now analyse in detail the performance of the best performing models, which is Hier-BiLSTM-CRF with pretrained sentence embedding for the India dataset, and LegalBert-BiLSTM-CRF for the U.K. dataset. We check how the models perform across various rhetorical roles and across various domains of law. We also check how well the rhetorical labels predicted by the models agree with expert-assigned rhetorical labels.

7.2.1 Performance across rhetorical labels

Table 8 shows the performance (in terms of F-score) of the best models for India and U.K. datasets. Some similar trends are observed for both datasets. The models are able to predict the label RATIO with very high confidence for both datasets. For the ARG (Argument) label, the models performs poorly for both the datasets, the probable reason being as follows. There are only 9% and 3.81% of sentences for the ARG label in the India and U.K. dataset respectively (refer to Table 6), which makes it difficult for the models to learn the characteristics of this label from a small amount of data. Additionally, the ARG label interleaves frequently with other labels like PRE, STA and RATIO. Hence, the model does not perform well for this label.

On the other hand, there are some differences observed between the two datasets. For the labels RLC (Ruling by Lower Court) and RPC (Ruling by Present Court) the performance of LegalBert-BiLSTM-CRF is substantially lower for the U.K. dataset (F-score 0.412 for RLC and 0.603 for RPC), but the performance of Hier-BiLSTM-CRF using pretrained sent2vec embeddings on the India dataset for these labels is quite high in comparison (F-score of 0.812 for RLC and 0.9 for RPC). This

Table 10 Label agreement matrix for labels assigned by (i) the best performing Hier-BiLSTM-CRF model with pretrained sent2vec embeddings, and (ii) majority opinion of the human annotators, for the India dataset

Human ↓ Model →	FAC	ARG	PRE	STA	RATIO	RLC	RPC
FAC	**21.34**	1.17	0.38	0.3	0.46	0.19	0
ARG	**2.85**	**4.89**	0.56	0.24	0.53	0.02	0
PRE	0.17	0.49	**14.29**	0.12	0.69	0	0.01
STA	0.61	0.25	0.59	**4.95**	0.50	0.03	0
RATIO	1.39	0.55	0.77	0.35	**35.82**	0.02	0.03
RLC	0.35	0.05	0.08	0	0.08	**2.75**	0.09
RPC	0	0	0.03	0.01	0.1	0.19	**2.48**

Each value shows the percentage of sentences, out of all sentences in the dataset. Non-diagonal values ≥ 2% are highlighted in bold, underline, indicating that the model sometimes confuses these label-pairs

Table 11 Label agreement matrix for labels assigned by (i) the best performing model, i.e., LegalBert-BiLSTM-CRF, and (ii) majority opinion of the human annotators, for U.K. dataset

Human ↓ Model →	FAC	ARG	PRE	STA	RATIO	RLC	RPC
FAC	**10.537**	0.083	0.006	0.589	**2.357**	0.804	0
ARG	0.094	**1.427**	0.215	0.017	**2.016**	0.039	0
PRE	0.066	0.149	**3.338**	0.138	**4.555**	0.16	0
STA	0.077	0.022	0.017	**4.577**	**2.269**	0	0
RATIO	1.603	1.091	**2.192**	1.697	**54.608**	0.33	0.342
RLC	0.826	0.066	0.022	0.028	0.755	**1.256**	0
RPC	0	0	0	0	0.738	0.006	**0.892**

Each value shows the percentage of sentences, out of all sentences in the dataset. Non-diagonal values ≥ 2% are highlighted in bold, underline, indicating that the model sometimes confuses these label-pairs

difference in performance of the model is due to an important difference between how the Indian and U.K. Supreme Court case documents are written. For Indian documents, even if there is more than one judge on the bench, a single paragraph is written (usually at the end of the document) which combines the reasonings of all the judges. In contrast, for U.K. documents, there are *separate paragraphs* for each "Lord" or judge where his/her judgement about the case is written. Thus, for the Indian documents, the final judgement is usually mentioned at the end of the document (which can be captured by CRF efficiently). Whereas, for the U.K. document, this may not always be the case because the judgement given by different judges is written in different paragraphs pertaining to the individual "Lord". In the absence of a particular position, it becomes difficult for CRF to capture the label efficiently for U.K. documents.

7.2.2 Performance across domains of Law

As stated earlier in Sect. 3, the India dataset has documents from five different domains of law. We now check how well the model performs on the different domains. Note that this analysis is specific to the India dataset only, as we did not have domain information for cases in the U.K. dataset.

The last column of Table 9 shows how accurately the best method is able to perform across the 5 different domains – in other words, how generalizable the model is across the five domains.[9] The model gives consistent performance (F-scores in the range [0.81, 0.85]) across all the domains, except for the domain 'Labour & Industrial law'. This performance is consistent with the human annotation, where we observed during the annotation process that the IAA between human annotators is also relatively low for the domain 'Labour & Industrial law'.

7.2.3 Comparing human-human and human-model agreement

Finally, we compare the agreement among the human annotators (the inter-annotator agreement studied in Sect. 4), and the agreement between the model and the majority opinion of the annotators. Specifically, we want to check whether the model also frequently confuses between those pairs of rhetorical labels that were found to be subjective during the human annotation process (as described in Sect. 4). To the best of our knowledge, this is the first attempt to perform such a comparative analysis, to gain more qualitative insights about the agreement of Law-AI models with human experts on this important task of rhetorical labeling.

To this end, we compute *label agreement matrices* shown in Tables 10 and 11 for the India and U.K. datasets respectively. In these tables, the rows represent the human-assigned labels (majority opinion of the annotators), and the columns represent the labels assigned by the model. Each *diagonal element* shows the percentage of sentences among all the sentences in the dataset for which the model-assigned label matches with the human-assigned label. Each *non-diagonal* element $C[i][j]$ shows the percentage of sentences where the human-assigned label i does not match the model-assigned label j. For both India and U.K. datasets, we find high values along the diagonal elements of the label agreement matrices, suggesting that there is a high human-model agreement in general.

Next, we focus on the non-diagonal elements that have relatively high values. The relatively high non-diagonal elements (that are higher than 2%) are highlighted in red underlined font. For the UK dataset (Table 11), we find that the model tends to sometimes confuse the RATIO label (reasons behind the Court's judgement) with several other labels (several values in the row and column named RATIO are highlighted). Though the performance of the model is quite high for these labels (as shown the by the analysis of F-score values in Table 8), this tendency of the

[9] Note that, during the 5-fold cross-validation, we ensured that at least one document from each domain is present in the training set (40 documents) as well as the test set (10 documents) in each fold.

Table 12 Performance of the best models (Hier-BiLSTM-CRF with pretrained sentence embedding for the India dataset, and LegalBert-BiLSTM-CRF for the U.K. dataset) on a completely unseen test set of documents (which were not part of the set on which cross validation was done)

Dataset	Statistics		P	R	F
	# docs	# sents			
India	20	2,602	0.839	0.786	0.806
UK	10	2,511	0.508	0.553	0.506

model to sometimes confuse RATIO with other labels is probably because the RATIO is often combined with facts of the case (FAC) or precedents (PRE) or arguments (ARG), thus making it difficult for the model to distinguish between these label-pairs.

Comparing the highlighted values in Table 11 (expert-model agreement) and those in Table 5 (expert-expert agreement) for the UK dataset, we find that the expert annotators also had some disagreement about sentences of the label RATIO. In fact, in both Tables 11 and 5, the highest disagreement is between the same pair of labels—RATIO and PRE. This observation suggests that the RATIO rhetorical role is somewhat subjective, and hence it is natural that the model will also face some difficulty in correctly identify RATIO sentences.

8 Analyzing the generalizability of the best performing models

In this section, we study the generalizability of the best performing models along two aspects – (i) how they perform on *unseen* documents from the same jurisdiction, and (ii) how they perform on documents from a different jurisdiction.

8.1 Performance on unseen documents from the same jurisdiction

The evaluation of the models described in Sect. 7 was done through cross-validation. We now apply the best performing models on a completely *unseen* test set of documents.

8.1.1 Curation of the unseen datasets

We use a separate set of 20 documents (having 2, 602 sentences) as the unseen India test set. The documents were randomly sampled from among the 53, 210 Indian Supreme Court case documents detailed in Sect. 3, leaving out the 50 documents that were used for training and cross-validating the models. Similarly, use 10 documents (containing 2, 511 sentences) as the unseen UK test-set. These 10 documents were randomly selected for the set of 742 UK Supreme Court documents described in Sect. 3, leaving out the 50 documents that were used for training and cross-validating the models. The rhetorical role of each sentence in these documents was labeled by law experts in the same process as described earlier. Note that none of

Table 13 Performance of the best model for rhetorical role labelling of sentences of one jurisdiction (India/UK), when applied to label sentences in documents from the other jurisdiction (UK/India)

Experiment	P	R	F
Applying the best model on UK dataset (LegalBert-BiLSTM-CRF) over India dataset	0.623	0.423	0.431
Applying the best model on India dataset (Hier-BiLSTM-CRF with Sent2Vec pretrained on India dataset) over UK dataset	0.363	0.285	0.207

the documents in the unseen test sets was part of the documents on which cross-validation was carried out.

8.1.2 Performance on unseen India test set

We employ the Hier-BiLSTM-CRF model with pretrained sent2vec sentence emdebbings model over the unseen India test set. Specifically, we use that model (out of the 5 trained models during five-fold cross validation) which had the best validation F-Score. We then evaluate the predictions given by the model and the human-assigned labels. The results are in Table 12. We find that the performance (F-Score = 0.806) of the model on the *unseen* test set is at par with the averaged cross-validation results reported in Table 7 (F-Score = 0.821). This result shows that our model can perform well in labelling sentences from unseen documents as well.

8.1.3 Performance on unseen UK test set

We label the UK test documents using LegalBert-BiLSTM-CRF model, which had the best validation FScore in the 5-fold cross-validation setup. The results are in the second row of Table 12. We observe that the performance on the unseen documents (F-Score = 0.506) is a bit lesser than the cross-validation results reported in Table 7 (F-Score = 0.600).

8.2 Cross-jurisdictional study of the best performing models

Now we study the cross-domain generalizability of the trained models. To this end, we take the best performing model for the India dataset (as measured by the validation F-Score), i.e., Hier-BiLSTM-CRF with Sent2Vec embeddings, and apply it to label sentences from the UK dataset. Conversely, we apply the best performing model for the UK dataset, i.e., LegalBert-BiLSTM-CRF, and apply it to label sentences from the India dataset. The performance of the models are stated in Table 13.

8.2.1 Applying the best model on UK dataset over India dataset

The first row of Table 13 shows the results when we apply the best performing model on the UK dataset (LegalBert-BiLSTM-CRF which had the best cross-validation performance on the UK dataset, as was stated in Table 7) to rhetorically label the sentences in the 50 documents from the India dataset which we have been using to train and cross-validate the various models (the dataset on which results were reported in Table 7). The weights of the model were learned via training over the UK dataset, and the model was then applied to infer the labels of the sentences in the India dataset. Note that, inference of the sentence embeddings and the final classification are done jointly inside the LegalBert-BiLSTM-CRF model, unlike the pretrained sentence emebedding variations of the Hier-BiLSTM/BiGRU-CRF model, where the embeddings have to be first inferred using the Sent2Vec/Bert/LegalBert models and then fed to the classification model (Hier-BiLSTM/BiGRU-CRF).

We find that the performance of the best UK model in rhetorically labeling documents in the India dataset is mediocre (F-Score = 0.431), when compared to the best performing model in the cross-validation setup (F-Score = 0.821, as reported in Table 7).

8.2.2 Applying the best model on India dataset over UK dataset

The second row of Table 13 states the results when we apply the best performing model on the India dataset (Hier-BiLSTM-CRF with pretrained Sent2Vec embeddings, which had the best F-score on the India dataset in Table 7) to rhetorically label the sentences in the 50 documents from the UK dataset. Note that, this architecture requires sentence embeddings from a Sent2vec model as input. We use the pretrained Sent2vec model of the India dataset and use it to infer the embeddings of the sentences in the UK dataset. Next, we use the best trained Hier-BiLSTM-CRF model on the India dataset, to infer the labels of the documents in the UK dataset. However, we observe that this model yields rather poor results over the UK dataset (F-Score = 0.207).

8.2.3 Role of the Sent2vec model

We also attempt to understand the role of the Sent2Vec model in the cross-jurisdictional study. Specifically, we investigate whether it is advisable to use a sent2vec model trained on text of the particular target jurisdiction on which the end-task is to be performed, or can a Sent2Vec model trained on text from a different jurisdiction achieve comparable performance. Specifically in this work, we have a much larger corpus of Indian case documents (53, 210 documents) compared to the UK corpus (742 documents). Hence, it is natural to ask – can the sentence embedding model learnt over the large Indian corpus improve the performance of rhetorical labeling over the U.K. documents?

To answer these questions, we consider the *Sent2vec model pretrained over the India dataset* and use it to infer the embeddings of the sentences of documents in the

Table 14 Average cross-validation performance on UK dataset using Hier-BiLSTM-CRF with pretrained Sent2Vec embeddings, where the embeddings of sentences in the UK dataset were inferred from the Sent2Vec model *pretrained on the India dataset*

Experiment	P	R	F
Best model on India dataset (Hier-BiLSTM-CRF with Sent2Vec pretrained on India dataset) cross-validated over UK dataset	0.544	0.483	0.493

UK dataset. Specifically, using these sentence embeddings, we train the classification model *Hier-BiLSTM-CRF with Sent2vec embeddings* over the UK dataset in a 5-fold cross-validation setup. Note that the Hier-BiLSTM-CRF model with Sent2vec embeddings is applied over the UK datasets exactly in the same way as reported in Sect. 7, with the only difference being that the Sent2vec model was pretrained over the same UK dataset in Sect. 7, while now the Sent2vec model is pretrained over the India dataset.

The performance on the UK dataset is reported in Table 14 where the validation scores are averaged across all the folds. We find that when the sent2vec model pretrained on the much larger India dataset (53K documents) is used, the performance is comparable (F-Score = 0.493) to that when the sent2vec model was trained on the much smaller UK dataset having only 742 documents (F-score = 0.491, as was reported in Table 7). This result shows that in-domain/in-jurisdiction pretraining of the sent2vec model is preferable even if there are much fewer documents available in the target domain; however, comparable results can be achieved with pretraining over large amounts of data from some different domain or jurisdiction.

8.2.4 Insights

The results in this section give important insights into the generalizability of neural models across different legal domains/jurisdictions, which is an important topic of research in recent times (Savelka et al. 2021). Two jurisdictions (such as India and UK, as in this work) may have significant differences in terms of the writing style and legal language, which leads to models trained over one jurisdiction perform relatively poorly over another jurisdiction. Since the jurisdictions are different, the prior-cases (precedents) and statute/law mentions will be different. Also the kind of names used in the jurisdictions will probably be different. Since we are relying on the *text* for the task, these differences are likely to affect the performances. Thus, it is usually beneficial to train a model for the particular 'target jurisdiction' on which the end-task is based.

However, it is possible that the target jurisdiction has a very small amount of training data available. In such a scenario, models pretrained over large amounts of data from some other jurisdiction can be applied. If such a model (pretrained over data from some other jurisdiction) is applied directly to a different target jurisdiction, then results are not likely to be good; e.g., Hier-BiLSTM-CRF with Sent2vec embeddings trained over the India dataset achieves F-score of only 0.207 when applied directly over the UK dataset (see Table 13). However, the same model achieves a much higher F-score of 0.493 over the same UK dataset, when it is *cross-validated* over the UK dataset (see Table 14). Thus, it is possible to get significant improvements in performance of models pretrained over data from some other jurisdiction, through some further training / finetuning over the target jurisdiction. As such, this is a practically useful approach for target jurisdictions that have only small amounts of training data.

9 Conclusion and future work

The objective of this work is to automate the task of rhetorical role labeling of legal case documents. We work on Indian Supreme Court and U.K. Supreme Court case documents – we perform an extensive inter-annotator study, experiment with different neural models for the task and analyse the prediction results in depth. We arrive at the following high-level take-aways:

(i) Neural models using Hierarchical BiLSTM architectures are much better in rhetorical role labeling, as compared to prior methods that use CRF over handcrafted features. Also the neural models generalize well across jurisdictions and legal domains.
(ii) Some specific pairs of rhetorical roles – such as (Ratio of the decision, Precedents) – are inherently subjective, and lead to disagreement even among law experts. The neural models too sometimes fail to distinguish between these labels.
(iii) When applying neural models to legal data mining tasks, it is better to use pretrained models for initializations. Additionally, domain-specific pretraining is more beneficial. From the experiments in this work, we noted that pretraining using Law2Vec word embeddings performs better than randomly initializing word embeddings or initialization using Google News vectors.
(iv) We observe that transformer models—Bert and LegalBert—are also well suited for the task, especially when such models are finetuned for the specific task.
(v) From the cross-jurisdictional study, we find that it is beneficial to use models pretrained on data from the same target jurisdiction on which the end-task is based. However, if a large amount of training data is not available from the target jurisdiction, then it is possible to get good results by taking a model pretrained on data from some other jurisdiction and further training the model even on small amounts of data from the target jurisdiction.

Finally, we note that neural models often lack transparency or explainablility, which is very much desired in the legal domain. Using handcrafted features aids in explanations, however we see that the performance of such handcrafted features is not as good as that of the neural models. Ultimately, this tradeoff between performance and explainability has to be decided keeping in mind the target task. For tasks such as identifying rhetorical roles of sentences, if it can be assumed that achieving good performance is more important than explainability, then it is beneficial to use the neural models proposed in this work.

The present work can be extended in future along several interesting directions. Since some label pairs are subjective even to law experts, an immediate future work would be to model the rhetorical role labelling task as a multi-label classification problem, wherein if a model's predicted label matches any one of the annotator's label, it would be considered a correct prediction. Also we would like to apply the rhetorical role labeled documents to downstream tasks such as summarization and computing similarity between two documents. Finally, while we have performed

cross-jurisdictional experiments where both jurisdictions follow the Common Law system, we would also like to explore if the models can be applied in non-Common Law settings.

Acknowledgements The authors acknowledge the anonymous reviewers whose comments helped to improve the paper. The authors also thank the Law domain experts from the Rajiv Gandhi School of Intellectual Property Law, India who helped in developing the gold standard data. The research is partially supported by SERB, Government of India, through a project titled "NYAYA: A Legal Assistance System for Legal Experts and the Common Man in India" and the TCG Centres for Research and Education in Science and Technology (CREST) through a project titled "Smart Legal Consultant: AI-based Legal Analytics". P. Bhattacharya is supported by a Fellowship from Tata Consultancy Services.

References

Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. Comput Linguist 34(4):555–596

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473

Bhattacharya P, Hiware K, Rajgaria S, Pochhi N, Ghosh K, Ghosh S (2019a) A comparative study of summarization algorithms applied to legal case judgments. In: European conference on information retrieval, Springer, pp 413–428

Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A (2019b) Identification of rhetorical roles of sentences in Indian legal judgments. In: legal knowledge and information systems–JURIX, pp 3–12

Bhattacharya P, Ghosh K, Pal A, Ghosh S (2020) Hier-spcnet: a legal statute hierarchy-based heterogeneous network for computing legal case document similarity. In: proceedings of the ACM SIGIR conference on research and development in information retrieval, pp. 1657–1660

Chalkidis I, Androutsopoulos I (2017) A deep learning approach to contract element extraction. In: legal knowledge and information systems–JURIX, pp. 155–164

Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in English. In: proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, pp 4317–4323, https://doi.org/10.18653/v1/P19-1424, https://www.aclweb.org/anthology/P19-1424

Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: the muppets straight out of law school. In: findings of the association for computational Linguistics: EMNLP 2020, pp 2898–2904, https://huggingface.co/nlpaueb/legal-bert-base-uncased

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555

Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: proceedings of NAACL-HLT 2019 pp. 4171–4186, https://huggingface.co/bert-base-uncased

Farzindar A, Lapalme G (2004) Letsum, an automatic legal text summarizing system. In: legal knowledge and information systems–JURIX, pp. 11–18

Galassi A, Lippi M, Torroni P (2020) Attention in natural language processing. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2020.3019893

Graves A, Fernández S, Schmidhuber J (2005) Bidirectional LSTM networks for improved phoneme classification and recognition. In: proceedings of the international conference on artificial neural networks (ICANN), pp. 799–804

Hachey B, Grover C (2006) Extractive summarisation of legal texts. Artif Intell Law 14(4):305–345

Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: proceedings of the eighteenth international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML 01, pp. 282–289

Liu CL, Chen KC (2019) Extracting the gist of chinese judgments of the supreme court. In: proceedings of the seventeenth international conference on artificial intelligence and law, pp. 73–82

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:171105101

Nejadgholi I, Bougueng R, Witherspoon S (2017) A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In: legal knowledge and information systems–JURIX, pp. 125–134

Pagliardini M, Gupta P, Jaggi M (2018) Unsupervised learning of sentence embeddings using compositional n-gram features. In: proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, Vol. 1, pp 528–540

Sanchez G (2019) Sentence boundary detection in legal text. In: proceedings of the natural legal language processing workshop 2019:31–38

Saravanan M, Ravindran B, Raman S (2008) Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In: proceedings of the international joint conference on natural language processing: Vol. 1

Savelka J, Ashley KD (2018) Segmenting us court decisions into functional and issue specific parts. In: legal knowledge and information systems–JURIX, pp. 111–120

Savelka J, Westermann H, Benyekhlef K, Alexander CS, Grant JC, Amariles DR, Hamdani RE, Meeùs S, Troussel A, Araszkiewicz M, Ashley KD, Ashley A, Branting K, Falduti M, Grabmair M, Harašta J, Novotná T, Tippett E, Johnson S (2021) Lex Rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In: proceedings of the international conference on artificial intelligence and law (ICAIL), pp. 129–138

Shao Y, Mao J, Liu Y, Ma W, Satoh K, Zhang M, Ma S (2020) Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In: proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20, pp. 3501–3507

Shulayeva O, Siddharthan A, Wyner AZ (2017) Recognizing cited facts and principles in legal judgements. Artif Intell Law 25(1):107–126

Venturi G (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), pp. 1–12

Walker VR, Pillaipakkamnatt K, Davidson AM, Linares M, Pesce DJ (2019) Automatic classification of rhetorical roles for sentences: comparing rule-based scripts with machine learning. In: proceedings of the workshop on automated semantic analysis of information in legal texts (with ICAIL)

Wang P, Yang Z, Niu S, Zhang Y, Zhang L, Niu S (2018) Modeling dynamic pairwise attention for crime classification over legal articles. In: the 41st international ACM SIGIR conference on research & development in information retrieval, pp. 485–494

Wang P, Fan Y, Niu S, Yang Z, Zhang Y, Guo J (2019a) Hierarchical matching network for crime classification. In: proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp. 325–334

Wang P, Fan Y, Niu S, Yang Z, Zhang Y, Guo J (2019b) Hierarchical matching network for crime classification. In: proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp. 325–334

Wyner A (2010) Towards annotating and extracting textual legal case elements. In: CEUR workshop proceedings vol. 605, pp. 9–18

Wyner AZ, Peters W, Katz D (2013) A case study on legal case annotation. In: legal knowledge and information systems–JURIX, pp. 165–174

Wyner AZ, Gough F, Lévy F, Lynch M, Nazarenko A (2017) On annotation of the textual contents of scottish legal instruments. In: legal knowledge and information systems–JURIX, pp. 101–106

Yamada H, Teufel S, Tokunaga T (2019) Neural network based rhetorical status classification for Japanese judgment documents. In: legal knowledge and information systems–JURIX, pp. 133–142

Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How does nlp benefit legal system: a summary of legal artificial intelligence. In: proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5218–5230

Authors and Affiliations

Paheli Bhattacharya[1] · Shounak Paul[1] · Kripabandhu Ghosh[2] · Saptarshi Ghosh[1] · Adam Wyner[3]

✉ Paheli Bhattacharya
paheli.cse.iitkgp@gmail.com

Shounak Paul
shounakpaul95@gmail.com

Kripabandhu Ghosh
kripaghosh@iiserkol.ac.in

Saptarshi Ghosh
saptarshi@cse.iitkgp.ac.in

Adam Wyner
a.z.wyner@swansea.ac.uk

[1] Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

[2] Department of Computational and Data Sciences (CDS), Indian Institute of Science Education and Research (IISER) Kolkata, Kolkata, West Bengal, India

[3] Law and Computer Science, Swansea University, Swansea, UK