HiCuLR: Hierarchical Curriculum Learning for Rhetorical Role Labeling of Legal Documents

Santosh T.Y.S.S[1], **Apolline Isaia**[1,2], **Shiyu Hong**[1,3], **Matthias Grabmair**[1]

[1] Technical University of Munich, Germany

[2]Télécom Paris

[3]Future Technology School; South China University of Technology

Abstract

Rhetorical Role Labeling (RRL) of legal documents is pivotal for various downstream tasks such as summarization, semantic case search and argument mining. Existing approaches often overlook the varying difficulty levels inherent in legal document discourse styles and rhetorical roles. In this work, we propose HiCuLR, a hierarchical curriculum learning framework for RRL. It nests two curricula: Rhetorical Role-level Curriculum (RC) on the outer layer and Document-level Curriculum (DC) on the inner layer. DC categorizes documents based on their difficulty, utilizing metrics like deviation from a standard discourse structure and exposes the model to them in an easy-to-difficult fashion. RC progressively strengthens the model to discern coarse-to-fine-grained distinctions between rhetorical roles. Our experiments on four RRL datasets demonstrate the efficacy of HiCuLR, highlighting the complementary nature of DC and RC.

1 Introduction

Rhetorical Role Labeling (RRL) of legal documents involves assigning the functional role played by each sentence of the document in the legal discourse (e.g., preamble, fact, evidence, reasoning). RRL for long legal case documents is a precursor task for several downstream tasks, such as case summarization (Hachey and Grover, 2006; Saravanan et al., 2008; Kalamkar et al., 2022; Farzindar and Lapalme, 2004), semantic case search (Nejadgholi et al., 2017; Ali et al., 2021), case law analysis (Savelka and Ashley, 2018) and argument mining (Walker et al., 2019; Ali et al., 2023).

Initial works on RRL of legal judgements have regarded the task as straightforward classification of sentences without modeling any contextual dependency between them (Walker et al., 2019) which later got to be viewed as sequence labeling (Bhattacharya et al., 2021). Initial works (Saravanan et al., 2008; Farzindar and Lapalme, 2004;

Hachey and Grover, 2006) performed RRL using hand-crafted features as part of a summarization pipeline. Further works (Walker et al., 2019; Savelka and Ashley, 2018) used Conditional Random Fields on these hand-crafted features. Recently, deep learning-based methods have been applied to this task on Japanese documents (Yamada et al., 2019), Indian documents (Bhattacharya et al., 2021; Ghosh and Wyner, 2019; Malik et al., 2022; Kalamkar et al., 2022). These methods adopt a hierarchical approach to account for the sequential sentence classification nature of the task, drawing context from surrounding sentences. This architecture, initially based on word embeddings (Bhattacharya et al., 2021; Ghosh and Wyner, 2019), has transitioned to BERT-based contextualized embeddings (Malik et al., 2022; Kalamkar et al., 2022), becoming the standard approach for RRL tasks. Recently, (Santosh et al., 2023) reformulated the task as span-level sequential classification that segment the document into sets of contiguous sequence of sentences (spans) and assign them labels. Further (Tyss et al., 2024) proposed contrastive and prototypical learning to effectively leverage knowledge from semantically similar instances (neighbours).

All these current approaches typically perform fine-tuning by presenting all training examples in a completely random order during training. However, the difficulty levels of discourse structures in legal documents can vary significantly, with some following a standard format identifiable by simple lexical clues, while others require deeper analysis due to unconventional writing styles. Similarly, identifying rhetorical roles such as preamble, which encompass broad definitions, is easier compared to finer-grained roles like the ratio of the decision, which is often confused with analysis or ruling by the present court. In this work, we leverage these observations to employ curriculum learning (CL) (Bengio et al., 2009), where the model's training process mimics a meaningful learning order in-

spired by human cognition. By excluding the negative impact of difficult examples in the early training stages, CL can guide learning towards a better local minima in the parameter space (Hacohen and Weinshall, 2019). CL has demonstrated success in various NLP tasks such as machine translation (Kocmi and Bojar, 2017; Zhang et al., 2018; Zhou et al., 2020; Platanios et al., 2019), NLU (Xu et al., 2020; Maharana and Bansal, 2022), AMR Parsing (Wang et al., 2022), summarization (Kano et al., 2021; Sotudeh et al., 2022; Sun et al., 2023), data-to-text generation (Chang et al., 2021), information retrieval (Su et al., 2021; Zhu et al., 2022; Zeng et al., 2022), pre-training (Ranaldi et al., 2023; Nagatsuka et al., 2021) and even in legal tasks such as statutory retrieval for mining negatives (Santosh et al., 2024).

In this work, we propose a hierarchical curriculum learning approach that nests two complementary curricula: (i) Document-level curriculum, which orders input documents based on their difficulty. We explore various difficulty metrics, such as deviations from standard discourse structure, to order these documents. We employ a baby-step scheduler (Spitkovsky et al., 2010), gradually exposing the model to more difficult samples over time. (ii) Rhetorical role-level curriculum, which utilizes similarity-based scheduling (Dogan et al., 2020). Ordering rhetorical roles in a sequential easy-to-hard sequence, as in the document-level curriculum, is infeasible since we cannot disentangle sentences from a document, which is our input. Hence we use a similarity-based curriculum where the model is initially allowed to belong to similar rhetorical roles to a lesser extent, instead of solely belonging to its ground-truth role and it gets refined over training. That is, in the beginning small mistakes of similar classes are less corrected than bigger mistakes, resembling a teaching process in which broad concepts are explained before subtle differences are addressed later. Our experimental results on four RRL datasets demonstrate the effectiveness of our various difficulty metrics for both DC and RC, as well as our model-agnostic HiCuLR framework, which incorporates both curricula.

2 Preliminaries

Task : Given a judgment document $x = \{x_1, x_2, \ldots, x_m\}$ with m sentences as the input, where $x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ represents the ith sentence containing n tokens, the task of RRL is to predict sequence of $l = \{l_1, l_2, \ldots, l_m\}$ where l_i is the rhetorical role corresponding to sentence x_i and $l_i \in$ L, set of predefined rhetorical roles.

Baseline : We use Hierarchical Sequential Labeling Network from prior works (Kalamkar et al., 2022; Malik et al., 2022) and demonstrate effectiveness of our model-agnostic HiCuLR framework in conjunction with it. Initially, each sentence x_i is encoded independently using a BERT model (Devlin et al., 2018) to derive token-level representations. These token-representations are passed through a Bi-LSTM layer (Hochreiter and Schmidhuber, 1997), followed by an attention pooling layer (Yang et al., 2016), to yield sentence representations. These are passed through Bi-LSTM layer to obtain contextualized sentence representations to encode information from surrounding sentences which are passed through a CRF layer (Lafferty et al., 2001) that predicts the best sequence of labels. The model is trained end-to-end with standard cross entropy loss.

3 HiCuLR framework

A curriculum learning framework consists of two main components - Difficulty scoring function which quantifies the data based on relative easiness and a pacing function which arranges the transitioning of data from easy to difficult examples for training (Bengio et al., 2009). In this section, we describe our document-level difficulty estimators along with the baby-step scheduler (3.1) and rhetorical role-level curriculum with similarity based scheduling (3.2). Further, we describe combining both of these complementary curricula in a nested hierarchical fashion (3.3).

3.1 Document-level Curriculum (DC)

We investigate four different strategies to obtain the relative difficulty across documents.

(a) Rhetorical Shifts : We hypothesize that documents with more consecutive shifts in rhetorical roles are harder to comprehend. We calculate the difficulty score as the number of consecutive shifts of rhetorical roles in a document normalized by the number of sentences in that document.

(b) Deviation from Expert Discourse: While there is no universally agreed-upon guideline for the discourse structure (sequence of rhetorical roles) in legal judgments, experts suggest an inherent structural pattern that is typically followed. Utilizing an

expert-provided discourse structure[1], we posit that documents deviating more from this structure are more challenging. We quantify the deviation based on the number of inversions required in the merge sort algorithm to align the document's rhetorical role structure with the expert-provided one.

(c) Deviation from Data-based Best Discourse: Instead of relying on expert-provided discourse structure, we derive the best structure by computing transition matrix for every pair of rhetorical roles using the training data for each dataset. We calculate the deviation score based on this sequence.

(d) Data-based Probabilistic Discourse: We utilize the transition matrix probabilities to compute the difficulty score as the log-likelihood of the sequence of labels for each document, which is then normalized by length. Lower log-likelihood scores indicate greater difficulty.

Based on the difficulty score, we partition the dataset into buckets and adopt baby-step scheduling for training (Spitkovsky et al., 2010). Initially, data from the easiest bucket is used and subsequent buckets are merged after a fixed number of steps, until the entire dataset is utilized.

3.2 Rhetorical Role-level Curriculum (RC)

Given that the RRL task operates on entire documents as input, sequentially exposing the model only to sentences with easy rhetorical roles followed by difficult ones in a sequential fashion becomes impractical. To address this, we employ a similarity-based scheduler (Dogan et al., 2020) for RC. This approach relies on a pairwise similarity matrix between rhetorical labels, which we obtain using two different strategies.

(a) Confusion Matrix: We hypothesize that more mistakes a model makes between pairs of rhetorical roles, the more similar and confusing they are to the model. Therefore, we obtain the confusion matrix from the validation dataset by using the trained model in a non-curriculum (random) order as our pairwise similarity matrix.

(b) Embedding Similarity: We pass the rhetorical role and their descriptions[2] to obtain the embeddings from the LegalBERT model (Chalkidis et al., 2020) to compute the semantic similarity between the rhetorical roles, resulting in a pairwise matrix.

---

[1] Expert-given discourse structure is obtained from Build respository at https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline

[2] Descriptionsare obtained from https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline

The similarity-based curriculum (Dogan et al., 2020) employs a probability distribution over rhetorical roles as the target label, unlike traditional one-hot encoding. This allows each sentence to be associated with similar rhetorical roles to a lesser extent, rather than being solely assigned to its ground-truth role. The class probabilities are initialized using a normalized similarity matrix to reflect the closeness among classes, and each row corresponds to a rhetorical label. This matrix is used in place of target probabilities when computing the cross-entropy loss. As the training progresses, this similarity matrix is gradually shifted towards the standard one-hot-encoding (diagonal matrix). Each element in similarity matrix is updated as:

$$v_{ij}^{t+1} = \begin{cases} \frac{1}{1+\varepsilon \sum_{k \neq j} v_{ik}^t} & i = j \\ \frac{\varepsilon v_{ij}^t}{1+\varepsilon \sum_{k \neq j} v_{ik}^t} & i \neq j \end{cases} \quad (1)$$

$\varepsilon \in (0,1)$ controls the convergence rate of labels to one hot vectors. v_{ij}^t represents probability of j-th element of i-th row at training step t. This approach aims to penalize major mistakes more heavily in the beginning, such as when the predicted class and target class are dissimilar, compared to minor mistakes when both the predicted and target class are similar. This resembles a teaching process where broader, easier rhetorical roles are learned before finer, closely associated rhetorical roles.

3.3 Combining both Complementaries

Our HiCuLR framework hierarchically nests both of these curricula, with the rhetorical-level on the outside and the document-level curriculum on the inside. Within each step of the rhetorical role-level curriculum, the document-level curriculum is operated, starting from easy buckets and progressively expanding to the entire dataset, until the next step of the rhetorical curriculum takes place and the entire procedure is repeated iteratively.

4 Experiments & Results

We experiment on four RRL datasets: (i) Build (Kalamkar et al., 2022), (ii) Paheli (Bhattacharya et al., 2021) (iii) M-CL and (iv) M-IT (Malik et al., 2022), derived from Indian legal judgments. Build has 13 rhetorical roles annotated, while the others have 7 each. Detailed dataset descriptions and implementation details are in App. A and B. Table 1 reports macro-F1 and micro-F1 scores.

		Build		Paheli		M-CL		M-IT	
		Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1
0	Baseline	60.02	78.43	61.53	66.48	58.42	66.21	63.54	68.15
1	Rhetorical Shifts	61.06	79.94	63.14	68.15	60.24	67.73	65.20	69.76
2	Dev. from Expert Disc.	60.78	79.77	62.77	67.86	59.06	66.87	63.48	68.54
3	Dev. from Data-based Best Disc.	60.64	79.64	63.10	68.35	61.15	68.88	64.22	69.47
4	Data-based Prob. Disc.	*61.14*	*80.17*	*63.52*	*69.13*	*61.85*	*69.46*	*65.29*	*70.20*
5	Confusion Matrix	60.82	80.66	*64.12*	69.27	61.74	70.13	*66.29*	71.24
6	Embedding Similarity	*61.44*	*81.28*	63.26	68.24	*62.27*	*71.42*	65.98	*71.49*
7	Hierarchical: RC (5), DC(4)	**63.61**	**82.11**	**65.28**	70.12	**63.15**	**72.08**	**69.42**	71.98
8	Hierarchical: RC (6), DC(4)	62.77	81.82	64.96	**70.48**	62.86	71.46	68.75	**72.27**
9	Sequential: DC(4), RC (5)	61.39	81.47	64.02	69.35	62.16	*71.65*	*68.12*	*72.06*
10	Sequential: RC(5), DC (4)	61.26	81.52	63.56	69.10	61.41	70.22	66.59	71.92
11	Reverse hier.: DC(4), RC (5)	*62.44*	*81.84*	*64.42*	*69.72*	*62.88*	71.47	67.21	71.48

Table 1: Performance comparison of various curriculum-based methods. Dev., Disc., hier. indicate deviation, disourse and hierarchical respectively. Numbers in bracket for entries (7-11) denote the index number in first column, indicating the specific variant. Overall best and best in each sub-group are bolded and italicized respectively.

4.1 Results

Document-level Curriculum (DC): Our analysis reveals that all four variants of difficulty scoring in DC led to improvements compared to the baseline across all datasets, highlighting the efficacy of using DC for RRL. Particularly, the data-based probability method (4) consistently outperformed other DC methods. While the deviation from expert discourse (2) performed slightly better or comparably to the data-based approach (3) on the Build dataset, its underperformance on other datasets suggests that the discourse structure derived from the Build annotation experts may not generalize well across different datasets. This suggests that a data-driven structure is a better proxy for expert-given discourse, facilitating easier adoption. Notably, the simple metric of rhetorical shifts (1), which does not incorporate specific label information, yielded better results than deviation-based methods (2, 3) in three out of four datasets, indicating its effectiveness as a strong signal for capturing document difficulty. This also signals the underutilization of the label information in the best sequence methods (2, 3), which is captured effectively using the data-based probabilistic discourse method (4). It suggests that using probabilistic method (4) facilitates the variability of possible discourse sequence styles as opposed to single best sequence in (2, 3).

Rhetorical Role-level Curriculum (RC): Both the RC methods perform better than the baseline. However, there is no clear winner between them. We attribute this to the differences in the label space of the datasets - Build dataset has more labels compared to others, indicating their fine-grained nature, which is captured effectively by embedding sim-

ilarity. Overall, RC performs slightly better than DC, indicating that the ordering of output labels is more important than that of input documents.

HiCuLR: We use the data-based probabilistic method (4) from DC and vary both the RC methods to create two HiCuLR variants (7,8). Overall, HiCuLR shows improvement compared to DC and RC alone, highlighting their complementary nature. Between them, confusion matrix takes the lead on challenging macro-F1. We also observe a trend reversal in HiCuLR compared to RC; for example, in the Build dataset, embedding similarity (6) in RC performs better, but within the HiCuLR, the confusion matrix (7) takes the lead. While embedding similarity reflects the label definition without considering the input, the confusion matrix captures the interplay of inputs and the associated rhetorical roles, making it a better proxy for overall dataset.

Ablation on combining DC and RC: We experiment with other strategies to combine DC and RC. Sequential strategy executes these curricula in a pipeline manner. Variant (9) executes DC first, followed by RC. DC-RC (9) outperforms RC-DC (10), suggesting that similarities between rhetorical roles towards the end of learning improve model performance. Additionally, combining both curricula in reverse hierarchical order, with RC on the outside and DC on the inside, performs better than sequential versions (9, 10), highlighting curricula interaction. Variant (11) underperforms compared to HiCuLR (7), indicating gradual exposure to rhetorical roles is preferable.

5 Conclusion

In this work, we proposed a novel hierarchical curriculum learning framework, HiCuLR, for RRL of legal documents. HiCuLR integrates complementary curricula, with Rhetorical Role-level Curriculum (RC) on the outer layer and Document-level Curriculum (DC) on the inner layer We investigate different difficulty scoring metrics in DC and similarity based strategies in RC, observing notable improvements in performance. Particularly, the data-based probabilistic method within DC and the confusion matrix approach within RC stood out as effective strategies. Our experiments on four RRL datasets, verify the effectiveness of HiCuRL.

Limitations

One limitation of the current task setup is its constraint to assign single label per sentence, which may not fully account for the complexity of lengthy sentences that can encompass multiple rhetorical roles. To overcome this constraint, one alternative could be to rethink the task as a multi-label classification, enabling each sentence to be associated with more than one rhetorical role. Another avenue worth exploring is to move away from sentence-level segmentation towards a more detailed approach at the phrase or sub-sentence level. This would involve assigning rhetorical roles to individual phrases or sub-sentences (Tokala et al., 2023) and also specifying the dependency relationships among these segments, similar to Discourse Dependency Parsing (Carlson et al., 2003).

Additionally, our evaluation is confined to datasets containing Indian legal documents. These datasets may share common vocabulary and writing style specific to the country's legal practices, potentially limiting the generalizability of our findings to legal documents from other jurisdictions. Since legal documents from different countries and regions may exhibit significant variations in language and structure, it's essential to broaden the assessment to include diverse legal contexts across different countries and regions.

Ethics Statement

The scope of this study is to introduce technical methodologies and corresponding empirical validations aimed at advancing rhetorical role labeling, a fundamental task at the forefront of legal document processing. Our experiments have been conducted on four datasets sourced from various Indian courts, made available through earlier works. Whenever applicable, consent for data usage was obtained according to the terms and conditions provided by the dataset providers. While these datasets contain real names of involved parties and lack anonymization, we anticipate no adverse effects resulting from our experimentation. We assert that our research makes a constructive contribution to the overarching objectives of advancing legal NLP and fostering the creation of AI-driven tools to improve productivity of legal professionals.

References

Basit Ali, Ravina More, Sachin Pawar, and Girish K Palshikar. 2021. Prior case retrieval using evidence extraction from court judgements. In *ASAIL/LegalAIIA@ ICAIL*, pages 1–11.

Basit Ali, Sachin Pawar, Girish Palshikar, Anindita Sinha Banerjee, and Dhirendra Singh. 2023. Legal argument extraction from court judgements using integer linear programming. In *Proceedings of the 10th Workshop on Argument Mining*, pages 52–63.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2021. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current and new directions in discourse and dialogue*, pages 85–112.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 727–733.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ürün Dogan, Aniket Anand Deshmukh, Marcin Bronislaw Machura, and Christian Igel. 2020. Label-similarity curriculum learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 174–190. Springer.

Atefeh Farzindar and Guy Lapalme. 2004. Letsum, an automatic legal text summarizing. In *Legal knowledge and information systems: JURIX 2004, the seventeenth annual conference*, volume 120, page 11. IOS Press.

Saptarshi Ghosh and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 3. IOS Press.

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

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Ryuji Kano, Takumi Takahashi, Toru Nishino, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2021. Quantifying appropriateness of summarization data for curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1395–1405.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–992.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2021. Pre-training a bert with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996.

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

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Poczós, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172.

Leonardo Ranaldi, Giulia Pucci, and Fabio Massimo Zanzotto. 2023. Modeling easiness for training transformers with curriculum learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 937–948.

TYS Santosh, Kristina Kaiser, and Matthias Grabmair. 2024. Cusines: Curriculum-driven structure induced negative sampling for statutory article retrieval. *arXiv preprint arXiv:2404.00590*.

TYS S Santosh, Philipp Bock, and Matthias Grabmair. 2023. Joint span segmentation and rhetorical role labeling with data augmentation for legal documents. In *European Conference on Information Retrieval*, pages 627–636. Springer.

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

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

Sajad Sotudeh, Nazli Goharian, Hanieh Deilamsalehy, and Franck Dernoncourt. 2022. Curriculum-guided abstractive summarization for mental health online posts. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 148–153.

Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021. Dialogue response selection with hierarchical curriculum learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1740–1751.

Shichao Sun, Ruifeng Yuan, Jianfei He, Ziqiang Cao, Wenjie Li, and Xiaohua Jia. 2023. Data selection curriculum for abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7990–7995.

Yaswanth Sri Sai Santosh Tokala, Sai Saketh Aluru, Anoop Vallabhajosyula, Debarshi Kumar Sanyal, and Partha Pratim Das. 2023. Label informed hierarchical transformers for sequential sentence classification in scientific abstracts. *Expert Systems*, 40(6):e13238.

Santosh Tyss, Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306.

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

Peiyi Wang, Liang Chen, Tianyu Liu, Damai Dai, Yunbo Cao, Baobao Chang, and Zhifang Sui. 2022. Hierarchical curriculum learning for amr parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 333–339.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Neural network based rhetorical status classification for japanese judgment documents. In *Legal Knowledge and Information Systems*, pages 133–142. IOS Press.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. Curriculum learning for dense retrieval distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1979–1983.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.

Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the association for computational linguistics*, pages 6934–6944.

Yutao Zhu, Jian-Yun Nie, Yixuan Su, Haonan Chen, Xinyu Zhang, and Zhicheng Dou. 2022. From easy to hard: A dual curriculum learning framework for context-aware document ranking. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2784–2794.

A Dataset

We experiment on four RRL datasets:

(i) Build (Kalamkar et al., 2022) includes judgments from Indian supreme court, high court and district courts, with publicly available train and validation splits. It comprises 184 and 30 documents respectively, totaling 31865 sentences, average of 115 per document, annotated with 13 rhetorical role labels. Due to the absence of a public test dataset, we use the training dataset for both training and validation, evaluating performance on the validation partition.

(ii) Paheli (Bhattacharya et al., 2021) features 50 judgments from the Supreme Court of India across five domains, with 7 rhetorical roles annotated. It contains 9380 sentences (average of 188 per document). We split into 80% train, 10% validation, and 10% test set at the document level.

(iii) M-CL / (iv) M-IT (Malik et al., 2022) encompasses judgments from the Supreme Court of India, High Courts, and Tribunal courts, with two subsets: M-CL, comprising 50 documents related to Competition Law, and M-IT, with 50 documents related to Income Tax cases. Both subsets are annotated with

7 rhetorical role labels. M-CL has 13,328 sentences (average of 266 per document) and M-IT has 7856 sentences (average of 157 per document). We split M-CL / M-IT into 80% train, 10% validation, and 10% test set at the document level.

B   Implementation Details

We follow the hyperparameters for baseline as described in Kalamkar et al. 2022. We use the BERT base model to obtain the token encodings. We employ a dropout of 0.5, maximum sequence length of 128, LSTM dimension of 768, attention context dimension of 200. We sweep over learning rates {1e-5, 3e-5, 5e-5. 1e-4, 3e-4} for 40 epochs with Adam optimizer (Kingma and Ba, 2014). For HiCuLR, we vary interval step size for target similarity matrix updates in RC over $\{5, 10, 15, 20, 25\}$, decay factor ε in RC over $\{0.8, 0.9, 0.95, 0.99, 0.999\}$, number of difficulty buckets in DC over $\{3, 5, 7, 10, 12, 15\}$, max training epochs during each baby step in DC over $\{2, 4, 6, 8, 10\}$. We derive the best model based on validation set performance.