SCaLAR_NITK @ JUSTNLP Legal Summarization (L-SUMM) Shared Task - Rhetorical Role based Abstractive Hierarchical Summarization of Indian Legal Documents

Arjun T D and Anand Kumar Madasamy
Department of Information Technology
National Institute of Technology Karnataka (NITK), Surathkal, India
{arjuntd.243it001, m_anandkumar}@nitk.edu.in

Abstract

This paper presents the systems we submitted to the JUST-NLP 2025 Shared Task on Legal Summarization (L-SUMM). Creating abstractive summaries of lengthy Indian court rulings is challenging due to transformer token limits. To address this problem, we compare three systems built on a fine-tuned Legal Pegasus model. System 1 (Baseline) applies a standard hierarchical framework that chunks long documents using naive token-based segmentation. System 2 (RR-Chunk) improves this approach by using a BERT-BiLSTM model to tag sentences with rhetorical roles (RR) and incorporating these tags (e.g., [Facts]...) to enable structurally informed chunking for hierarchical summarization. System 3 (WRR-Tune) tests whether explicit importance cues help the model by assigning importance scores to each RR using the geometric mean of their distributional presence in judgments and human summaries, and finetuning a separate model on text augmented with these tags (e.g., [Facts, importance score 13.58]). A comparison of the three systems demonstrates the value of progressively adding structural and quantitative importance signals to the model's input.

1 Introduction

Automatic text summarization of legal documents is a critical, high-impact challenge in applied NLP. It offers the potential to help legal professionals quickly distill lengthy and complex case judgments, thereby improving judicial efficiency (Shukla et al., 2022). In a multilingual nation like India, this task is further complicated by the need to ensure access to justice across different languages (Datta et al., 2023). As the volume of legal text continues to grow, the development of robust benchmarks and models for the Indian legal domain has become an active area of research (Joshi et al., 2024).

The JUST-NLP 2025 Legal Summarization (L-SUMM) shared task provides a key benchmark for this problem, focusing on the abstractive summarization of Indian court judgments. A primary difficulty in this task, as noted by Sharma et al. (2023), is the extreme length of legal documents, which often exceeds the input-token limitations of modern transformer models like PEGASUS. This necessitates intelligent strategies beyond naive truncation.

A promising avenue for handling long documents is to leverage their inherent logical structure. Prior work has shown the value of semantic segmentation of legal texts (Kalamkar et al., 2022). A powerful way to represent this structure is through the identification of rhetorical roles (e.g., *Facts*, *Reasoning*, *Decision*), a technique that has been successfully applied to legal texts for summarization and analysis (Saravanan et al., 2008; Bhattacharya et al., 2019; Malik et al., 2022).

In this paper, we present our three systems submitted to the L-SUMM task, which explore a progressive integration of this structural information. Our baseline system uses a standard hierarchical, token-based chunking method. Our second system introduces a more context-aware hierarchical approach, using "rhetorical chunks" based on semantic roles. Our final, most advanced system fine-tunes a model on text embedded with data-driven importance scores for each rhetorical role, explicitly teaching the model to weigh information based on our analysis of the summary-generation process.

2 Methodology

2.1 Base Model and Fine-Tuning

All three systems leverage the **Legal Pegasus** (`nsi319/legal-pegasus`) model, chosen for its fine-tuning on legal text. We fine-tuned this model on the full 1200-document **InLSum** training set. Due to GPU memory constraints with long sequences, we employed memory-saving techniques:

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 121–126
December 24, 2025 ©2025 Association for Computational Linguistics

the Adafactor optimizer, a per-device batch size of 1, and gradient accumulation steps of 8 (effective batch size of 8). The learning rate was set to 2×10^{-5}.

2.2 Handling Document Length

A primary challenge in legal summarization is the extreme length of judgments, often exceeding the 1024-token limit of models like PEGASUS. We implemented a two-pronged strategy:

- **Short Documents:** Judgments approximated as shorter than 1024 tokens were summarized directly by feeding the entire text to the model.
- **Long Documents:** Judgments exceeding the threshold were processed using a hierarchical summarization approach, detailed differently for each system below.

2.3 System 1: Baseline Hierarchical Summarization

Our baseline system addresses the challenge of long documents using a standard hierarchical summarization technique, illustrated in Figure 1. For documents exceeding the model's input limit (approximated by a 4096-character threshold), the text is first divided into overlapping chunks of approximately 900-1000 tokens. This naive token-based splitting often disrupts the natural semantic flow of the legal text. Each chunk is then summarized independently by our fine-tuned Legal Pegasus model, capturing primarily local context. These initial summaries are recursively combined in pairs (or small groups) and re-summarized, building a tree structure where each ascending level incorporates context from a wider portion of the original document. This step is repeated recursively until the final summary is generated, which will have context of all the initial chunks.

2.4 System 2: Hierarchical Summarization with Rhetorical Chunking (RR-Chunk)

System 2 enhances the hierarchical approach by incorporating semantic structure. We first preprocess the entire dataset using a BERT-BiLSTM model, which uses BERT model which is finetuned on Indian legal corpus for word embeddings and two Bi-LSTMs for sentence and document level context followed by a classification head to tag each sentence with one of seven rhetorical roles (e.g., *Facts*, *Decision*). A Legal Pegasus model is then fine-tuned on this enriched data, learning to recognize text prepended with role tags (e.g., `[Facts] The petitioner...`). During inference for long documents, instead of splitting by token count, we employ **rhetorical chunking**: consecutive sentences sharing the same role are grouped into a single chunk. This preserves the logical units of the judgment (like keeping all facts together) and provides more coherent segments to the summarizer. But if the chunk itself is longer than 1024 tokens then the chunk is again split using a 900-1000 token threshold. The same multi-level hierarchical summarization process depicted in Figure 1 is then applied, using these semantically meaningful rhetorical chunks as the base units and the RR-tuned model for summarization at each level. This approach aims to guide the language model more effectively by leveraging the inherent structure of the legal document, hypothesizing that summaries generated from complete rhetorical units will be superior. Short documents are summarized directly using the RR-tuned model.

2.5 System 3: Fine-Tuning with Weighted Rhetorical Roles (WRR- Tune)

Our third system investigates whether explicitly signaling the data-driven importance of each rhetorical role during fine-tuning can further enhance summarization.

Algorithm 1 RR Importance Scoring

```
Require: Tagged Judgments $J$, Tagged Summaries $S$
Initialize $C_j[r] \leftarrow 0$, $C_s[r] \leftarrow 0$ for all roles $r$
Initialize $T_j \leftarrow 0$, $T_s \leftarrow 0$
for all document $d$ in $J \cap S$ do
    for all sentence $sent$ in $d_{judg}$ do
        $r \leftarrow$ get_role($sent$)
        $C_j[r] \leftarrow C_j[r] + 1$; $T_j \leftarrow T_j + 1$
    end for
    for all sentence $sent$ in $d_{summ}$ do
        $r \leftarrow$ get_role($sent$)
        $C_s[r] \leftarrow C_s[r] + 1$; $T_s \leftarrow T_s + 1$
    end for
end for
for all $r$ in all unique roles do
    $P_j \leftarrow (C_j[r]/T_j) \times 100$
    $P_s \leftarrow (C_s[r]/T_s) \times 100$
    $Ret \leftarrow P_s/P_j$
    $Score[r] \leftarrow \sqrt{Ret \times P_s}$
end for
return $Score$ sorted descending
```

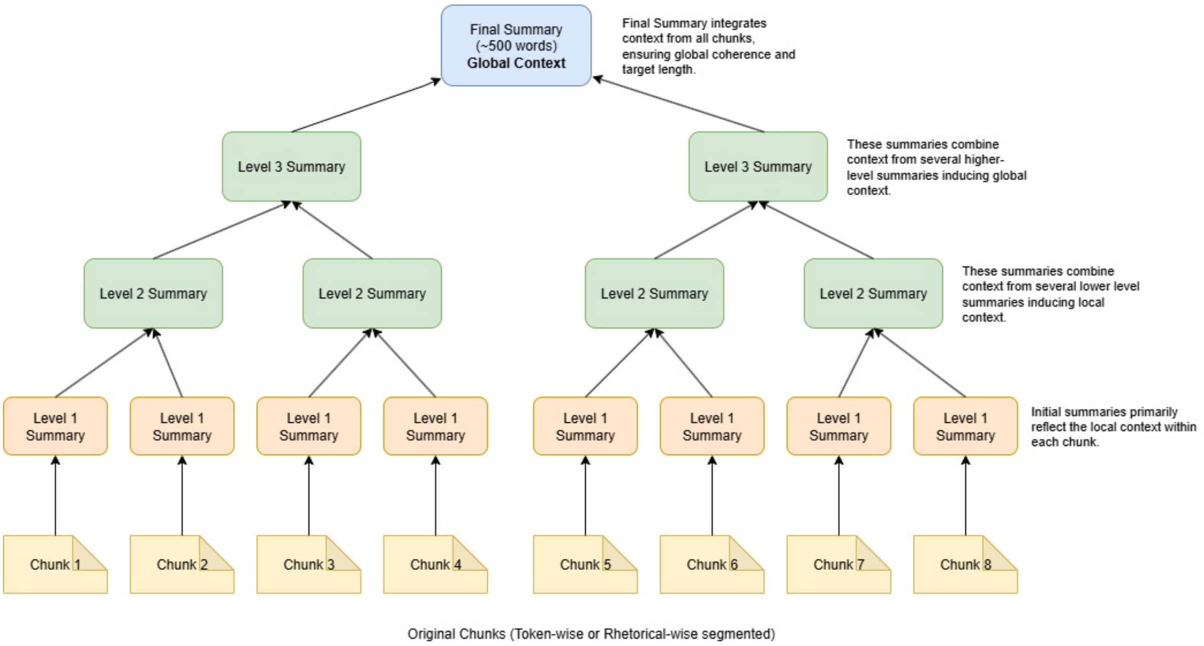
Final Summary
(~500 words)
Global Context
Final Summary integrates context from all chunks, ensuring global coherence and target length.
Level 3 Summary
Level 3 Summary
These summaries combine context from several higher-level summaries inducing global context.
Level 2 Summary
Level 2 Summary
Level 2 Summary
Level 2 Summary
These summaries combine context from several lower level summaries inducing local context.
Level 1 Summary
Level 1 Summary
Level 1 Summary
Level 1 Summary
Level 1 Summary
Level 1 Summary
Level 1 Summary
Level 1 Summary
Initial summaries primarily reflect the local context within each chunk.
Chunk 1
Chunk 2
Chunk 3
Chunk 4
Chunk 5
Chunk 6
Chunk 7
Chunk 8
Original Chunks (Token-wise or Rhetorical-wise segmented)

Figure 1: Hierarchical summarization tree. Initial summaries (Level 1) capture local context from base chunks (token-based or rhetorical). Subsequent levels combine these summaries, progressively incorporating broader context. until the final summary integrates information globally.

Algorithm 1 estimates how important each rhetorical role is to human-written summaries. Let J denote the set of tagged judgments and S the set of tagged summaries. For every document d that appears in both sets ($d \in J \cap S$), we iterate over all judgment sentences d_{judg} and summary sentences d_{summ}, and obtain their rhetorical roles. We maintain two role-count distributions: $C_j[r]$ records how many judgment sentences belong to role r, and $C_s[r]$ records the same for summaries. We also track total sentence counts T_j and T_s across judgments and summaries respectively. After accumulating these counts over all documents, we compute the percentage frequency of each role in judgments as $P_j = (C_j[r]/T_j) \times 100$ and in summaries as $P_s = (C_s[r]/T_s) \times 100$. The retention ratio $Ret = P_s/P_j$ captures how strongly role r is preserved from judgment to summary. Finally, the overall importance score for each role is given by $Score[r] = \sqrt{Ret \times P_s}$, which emphasizes roles that are both frequently included in summaries and retained at a high rate. Roles are then ranked in descending order of $Score[r]$.

We first derived an importance score for each rhetorical role based on its representation in the training data, as detailed in Algorithm 1. The core idea is that roles significantly more concentrated in human-written summaries (relative to their presence in full judgments) are more important for summarization . To balance this **retention** factor with the role's absolute **presence (volume)** in the summary, we employed the **geometric mean** (`sqrt(Retention * Summary percentage(volume)`), ensuring high scores are assigned only to roles that are both highly retained and substantially present. This analysis yielded the scores shown in Table 1, identifying roles like *Facts* and *Decision* as most important. The seven rhetorical roles are Facts, Reasoning, None, Decision, Arg Petitioner, Arg Respondent and Issue. As show in table 1, each sentence is classified into one of the seven rhetorical roles along with it's importance score. We then created a new version of the training dataset where the input text embedded these scores within the role tag using a human-readable format, for example, `[Facts, importance score 9.48] The petitioner....` This descriptive tag provides a clearer linguistic signal to the model about the score's meaning compared to just embedding rhetorical role. A separate Legal Pegasus model (**WRR-Tune**) was then fine-tuned from scratch on this new weighted-role dataset, learning to directly associate the explicit importance score with the summarization task during training. For inference, System 3 uses the same **rhetorical chunking** hierarchical method as System 2, but utilizes this specialized WRR-Tuned model to generate summaries at each level, thus leveraging the

Role	Imp Score	Retention	Summary %
Facts	9.48	2.83x	31.73%
Reasoning	5.61	2.18x	14.43%
None	4.93	0.64x	37.87%
Decision	4.77	3.60x	6.33%
Arg. Petitioner	2.09	0.63x	6.92%
Arg. Respondent	0.73	0.23x	2.28%
Issue	0.60	0.81x	0.44%

Table 1: Rhetorical Role Importance Scores.

learned importance weights throughout the process. This system tests whether the model can effectively learn and utilize explicit, data-driven importance signals provided directly within the input during fine-tuning.

3 Results and Discussions

3.1 Computational Environment

All experiments were conducted using a virtual machine equipped with an Intel(R) Xeon(R) CPU @ 2.00GHz and an NVIDIA Tesla P100 GPU with 16GB VRAM. The models were implemented in Python using the PyTorch deep learning framework, along with the Hugging Face Transformers library for BERT-based architectures. All the experiments being reported in the paper including the comparative studies were done by us, in this computational setup.

3.2 Results

The performance of our three systems on the validation set and the performance of System 3 on the test set are presented in Table 2.

As shown in Table 2, there is a clear and consistent improvement across all metrics on the validation set as we progressed from System 1 to System 3. System 2 (RR-Chunk), which utilized rhetorical roles for more coherent chunking, significantly outperformed the baseline System 1, highlighting the benefit of incorporating semantic structure into the hierarchical process. System 3 (WRR-Tune), which fine-tuned the model with explicit, data-driven importance scores embedded in the input, achieved the best performance on the validation set by a considerable margin, particularly demonstrating strong gains in ROUGE-2 and BLEU scores. This confirms our hypothesis that providing the model with quantitative importance signals during fine-tuning is a highly effective strategy for legal summarization. On the final test set, System 3 maintained strong performance, achieving an average score of 20.74 and a ROUGE-L score of 25.93.

Overall, the steady rising trend in all three systems indicates that summarization performance is greatly improved by combining both structure information (by rhetorical roles) and quantitative salience cues (via significance ratings). The advancements show that models that are informed by both explicit markers of content relevance and linguistic structure are beneficial for legal abstraction.

3.3 Discussion

The results clearly demonstrate a progressive improvement from System 1 to System 3 across all evaluation metrics. System 1, which relies on a standard hierarchical summarization pipeline with naive token-based chunking, provides a reasonable baseline but is limited by its inability to preserve the semantic structure of legal documents. As a result, the model often receives context fragments that do not align with coherent discourse units, reducing the effectiveness of the hierarchical encoder–decoder process.

System 2 (RR-Chunk) shows a noticeable increase in ROUGE-2, ROUGE-L, and BLEU, which highlights the advantage of using rhetorical roles for segmentation. Since Indian court judgments follow a well-defined argumentative structure, grouping text according to rhetorical roles leads to more meaningful chunks. This enables the model to better capture fact-heavy and decision-relevant sections, improving the overall quality of the generated summaries.

System 3 (WRR-Tune) achieves the highest performance on both the validation and test sets. By fine-tuning the model with explicit importance scores embedded directly into the input, the system gains an additional signal that helps it prioritize legally salient content during generation. These importance cues guide the model toward focusing on segments that contribute more substantially to accurate and coherent summaries. The stronger gains in ROUGE-2 and BLEU suggest that importance-weighted fine-tuning enhances the model's ability to reproduce key multiword expressions and legally significant phrasing.

Overall, the consistent upward trend across all three systems confirms that incorporating both structural information and quantitative salience cues significantly boosts summarization performance. The improvements indicate that legal abstraction benefits from models that are guided not only by linguistic structure but also by explicit indicators of content relevance.

Dataset	System	AVG	ROUGE-2	ROUGE-L	BLEU
Validation Set Results					
Validation	System 1 (Baseline Hierarchical)	18.65	18.81	24.43	12.70
Validation	System 2 (RR-Chunk)	19.93	20.37	25.16	14.26
Validation	**System 3 (WRR-Tune)**	**21.53**	**22.57**	**26.28**	**15.75**
Test Set Results					
Test	**System 3 (WRR-Tune)**	20.74	21.86	25.93	14.43

Table 2: Performance comparison of the three systems on the InLSum validation set and the final test set performance for System 3. Scores are reported as provided by the shared task organizers.

4 Conclusion and Future Work

In this paper, we presented three systems for the JUST-NLP 2025 Legal Summarization shared task, all based on a fine-tuned Legal Pegasus model. Our methods progressed from a standard hierarchical baseline (System 1) to a semantically-aware model using rhetorical-role-based chunking (System 2), and finally to a novel system fine-tuned on text embedded with data-driven importance scores (System 3). Our experiments on the validation set show a clear and consistent performance improvement at each stage, with System 3 achieving the highest scores across all metrics. This confirms our hypothesis that progressively enriching the model's input with both structural-semantic information (rhetorical roles) and quantitative, data-driven signals (importance scores) is a highly effective strategy for producing more accurate and coherent summaries of complex legal judgments. For future work, we plan to address the limitations of hierarchical chunking by experimenting with end-to-end long-context models such as LED or Long-T5, which can process entire documents at once. We also plan to work on explainablity and ensembling with other well performing LLMs like BART and LED.

Limitations

The hierarchical summarization framework, used in all three systems, is a necessary workaround but is not ideal. It still risks context loss at chunk boundaries and, more significantly risks the coherency of the final summary . Furthermore, the performance of Systems 2 and 3 is fundamentally dependent on the accuracy of the upstream BERT-BiLSTM model used for rhetorical role tagging. Any errors from this classifier are propagated and potentially amplified by the summarization model, which has been trained to trust these (sometimes incorrect) structural and weighted tags.

Ethics Statement

This research was conducted using publicly available legal dataset released for academic and research purposes. No private or personally identifiable information was involved at any stage. The primary goal of this work is to explore abstractive summarization for legal documents. While the proposed models show promising results, they reflect patterns present in the training data. Any biases, inaccuracies, or limitations in the dataset may influence model predictions. Therefore, these models should not be seen as replacements for human legal reasoning. We strongly encourage users to apply this work responsibly and ethically, keeping in mind the sensitive nature of legal decision making.

Acknowledgments

This work was supported by the ANRF-SERB-CRG project titled “A Deep Explainable Framework for Semantically Similar Document Retrieval and Summarization of Legal Text,” under the supervision of Dr. Anand Kumar M, Department of Information Technology, National Institute of Technology Karnataka, Surathkal.

References

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identifying rhetorical roles of sentences in indian legal judgments. In *Proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX)*.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.

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*.

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. Association for Computational Linguistics.

M. Saravanan, B. 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*. IJCNLP.

Saloni Sharma, Surabhi Srivastava, Pradeepika Verma, Anshul Verma, and Sachchida Nand Chaurasia. 2023. A comprehensive analysis of Indian legal documents summarization techniques. *SN Computer Science*, 4(5).

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.