LTRC at SemEval-2023 Task 6: Experiments with Ensemble Embeddings

Pavan Baswani, Hiranmai Sri Adibhatla, Manish Shrivastava
Language Technologies Research Center, KCIS
IIIT Hyderabad, India
pavan.baswani@research.iiit.ac.in, hiranmai.sri@research.iiit.ac.in, m.shrivastava@iiit.ac.in

Abstract

In this paper, we present our team's involvement in *SemEval-2023 Task 6: LegalEval: Understanding Legal Texts*. The task comprised three subtasks, and we focus on subtask A: Rhetorical Roles prediction. Our approach included experimenting with pre-trained embeddings and refining them with statistical and neural classifiers. We provide a thorough examination of our experiments, solutions, and analysis, culminating in our best-performing model and current progress. We achieved a micro F1 score of 0.6133 on the test data using fine-tuned LegalBERT embeddings.

1 Introduction

Rhetorical role labeling is a critical aspect of analyzing and interpreting legal documents, particularly in the field of judgment. The labeling of rhetorical roles involves identifying the semantic purpose of each sentence and categorizing the different roles played by each part of a judgment, including the parties involved, the arguments presented, and the decision made. Identifying the different rhetorical roles of sentences in a legal case document can enhance the document's comprehension (Malik et al., 2021b) and aid in various downstream tasks, such as classification (Chalkidis et al., 2021), text summarization (Jain et al., 2021), case law analysis, and entity extraction (Skylaki et al., 2021; Chalkidis and Androutsopoulos, 2017; Chalkidis et al., 2017).

Despite its potential benefits, the lack of structure, technical jargon, multiple themes, and high specificity makes it difficult even for legal experts to identify the rhetorical roles effectively. Therefore, the task of rhetorical labeling in legal documents is a taxing and challenging one that requires advanced NLP techniques. By understanding the various rhetorical roles at play, legal professionals can more effectively analyze and interpret the content of legal documents, and ultimately, make informed legal decisions (Chalkidis et al., 2019). This process can help lawyers to understand the implications of the judgment, and ensure that their clients' interests are protected in the legal system.

While it may be challenging to automate the entire judicial process, automating intermediate tasks can be valuable in assisting legal practitioners and accelerating the system. Nevertheless, legal documents possess unique characteristics, and the application of existing NLP models and techniques can be difficult due to the difference between legal texts and commonly occurring texts used to train such models. Consequently, the development of legal domain-specific techniques is essential to enhance the application of NLP in the legal domain. As a step in that direction, Task 6 (LegalEval: Understanding Legal Texts) of SemEval (Modi et al., 2023) proposed 3 shared subtasks which will act as building blocks in developing legal AI applications. In this paper, we describe our efforts on Sub-Task A: Rhetorical roles (RR) prediction[1].

2 Related Work

In recent years, there has been a surge in research in the field of legal text processing, leading to the creation of several datasets, tasks, and applications. Examples include but are not limited to Prior case retrieval (Al-Kofahi et al., 2001; Jackson et al., 2003), summarization (Bhattacharya et al., 2019), events and named entities extraction (Kalamkar et al., 2022a; Lagos et al., 2010), and judgment prediction (Xiao et al., 2018; Chalkidis et al., 2019; Malik et al., 2021b).

Rhetorical role (RR) labeling is a process of assigning specific labels such as Fact, Argument, Final Judgment, and more to individual sentences within a court case document. This task is a critical component of legal analytics, as it can improve the readability of lengthy case documents and aid in

[1] https://github.com/pavanbaswani/rhetorical_roles

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 841–846

various downstream tasks. Moreover, it can add a structural framework to the document, thereby enabling an organized and systematic analysis of the legal document. Data mining and machine learning approaches (Walker et al., 2019) were employed for classifying a sentence to its respective rhetorical role. However, the dataset size and the number of classes or roles are less. Several neural models (Bhattacharya et al., 2021) were also built using LSTMs (Hochreiter and Schmidhuber, 1997) and CRF (Lafferty et al., 2001). Legal judgments from the Supreme Court of India and the Supreme Court of the United Kingdom (U.K.) were considered and seven rhetorical roles were identified. Transformer-based models (Kohli et al., 2021) like RoBERTa (Liu et al., 2019) with attention mechanism were employed to capture long-range relations. It has been demonstrated that incorporating both span segmentation and sequence classification (Santosh et al., 2023) can enhance accuracy in various tasks. A new corpus of legal documents from diverse legal domains has been created (Malik et al., 2021a), which includes more detailed and fine-grained annotations of 13 rhetorical role labels. This corpus forms the baseline for the shared task.

3 System Description

3.1 Dataset

The dataset (Kalamkar et al., 2022b) contains legal documents related to criminal and tax cases, focusing on Indian legal documents in English. This corpus contains 12 fine-grained RRs and a NONE label. Figure 1 presents the category-wise distribution of criminal and tax cases. This dataset comprised legal judgments from various courts, including the Supreme Court of India, High Courts, and Tribunal courts.

The data is split into 247 training documents and 30 dev documents. However, one of the training documents is a duplicate, resulting in 150 duplicated annotation samples within the training data. Table 1 details the statistics of the dataset. All plots are computed after removing the duplicate document(s) from the training data. Figure 2 and 3 highlight the RR labels and their corresponding distribution across train and dev respectively.

3.2 Models

Massively pre-trained Transformer-based language models have led to significant progress in natural language processing (NLP), demonstrating state-of-the-art performance in various tasks. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), a pre-trained Language model, is based on the transformer architecture and is optimized for sequence-to-sequence tasks such as language modeling and machine translation. In this paper, we explore

	Train	Dev
# Documents	247	30
# Duplicate Docs	1	0
# Duplicate Samples	150	0
# Unique Samples	28836	2890
# Unique Labels	13	13
sentence length (min)	1	1
sentence length (max)	695	487
sentence length (avg)	36.98	35.48

Table 1: Train and Dev data Statistics (where Dev = development)

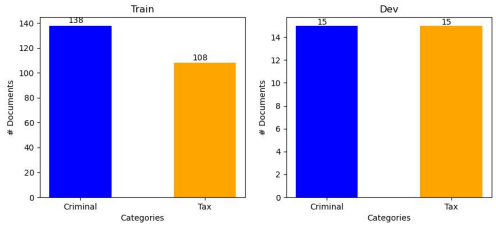

Figure 1: Category wise Documents' distribution

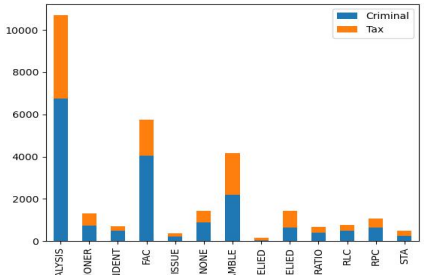

Figure 2: Train Data Label Distribution

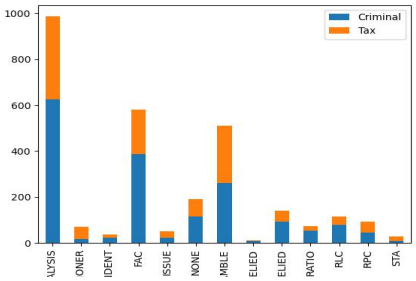

Figure 3: Dev Data Label Distribution

two variants of neural models with and without fine-tuning along with various classification heads to automatically identify the rhetorical roles of sentences in legal documents. Further, each of these neural models has variations on the choice of pre-trained embeddings.

Embeddings

In our preliminary experiments, we explored variants of pre-trained embeddings and ensemble embeddings from bert-based models with various statistical and neural classification heads:

LegalBERT: We use LEGAL-BERT-BASE (Chalkidis et al., 2020) which is fine-tuned on BERT (Devlin et al., 2019) for legal domain and has shown substantial improvement in challenging downstream tasks like multi-label-classification.

InCaseLawBERT: In contrast to LegalBERT, InCaseLawBERT (Paul et al., 2022) was trained on 5.4 million Indian legal documents using the same configuration as the bert-base-uncased model. These case documents were gathered from the Indian Supreme Court and High Courts. The model was initialized with LegalBERT (Zheng et al., 2021) model and further fine-tuned on Indian legal documents for 300K steps, focusing on Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks.

LegalBERT+LDA: Latent Dirichlet Allocation (LDA) (Jelodar et al., 2019) is a popular approach that defines each topic by a collection of words. These topic modeling approaches have an advantage as the distribution of topics over documents is obtained. By combining sentence vectors obtained by LDA and LegalBERT which have been fine-tuned on legal contracts, we achieve context-aware topic identification. The two-sentence vectors are concatenated with a weight hyper-parameter to balance the relative importance of information from each source.

LegalBERT+RoBERTa: Contextualized embeddings that are pre-trained have proven to be effective word representations for tasks involving structured prediction. It has been discovered through recent research that even better word representations can be achieved by combining various types of embeddings. Nonetheless, the process of selecting the optimal combination of embeddings to create the most effective concatenated representation can differ depending on the task and available embedding options. [CLS] is a special classification token and the last hidden state of BERT corresponding to this token (h[CLS]) is used for classification tasks. We concatenated the [CLS] embedding from the LegalBERT with the sentence embedding (mean pooled) of RoBERTa (Liu et al., 2019) and evaluated on the task dataset. Since the concatenated vector is in a high-dimensional space, where information would be sparse, autoencoder techniques are used to learn a lower-dimensional representation. By utilizing these dimensionality reduction techniques (autoencoder), we are able to reduce the embedding dimensions from 768 to 512 and 256.

3.2.1 ML based Classifiers

The reduced embeddings are fed as input to various machine learning (ML) classification models without fine-tuning them. Specifically, we evaluated the performance of Logistic Regression (LR), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and XGBoost.

3.2.2 Neural Classifiers

We conducted experiments using the three different variants of embeddings (pre trained LegalBERT, LegalBERT + LDA, InCaseLawBERT) with different neural classification heads (Linear, Dense, and Multiple Binary Classifiers (MBC)). During training, we fine-tuned the models and updated the weights using the Cross-entropy loss function. Compared to traditional machine learning models, these fine-tuned models exhibited superior performance. The fully connected linear and dense layers combine input features via flattened inputs and matrix multiplication. Activation functions such as ReLU or sigmoid introduce nonlinearity into the model. These layers serve as the classification head for many classification models.

Alternatively, the MBC approach involves training N binary classifiers, where N is the number of classes in a given classification problem. Each binary classifier predicts the probability of an input belonging to one of the N classes. This approach is also known as "one-vs-all" or "one-vs-rest" classification.

4 Evaluation

4.1 Experimental Setup

We experiment with and without fine-tuning pre-trained transformers LegalBERT, InCaseLawBERT, ensemble embeddings LegalBERT+LDA and LegalBERT+RoBERTa. The train set contains

246 documents and dev set contains 30 documents (refer Table 1). Given that the maximum sequence length in the annotated sample is less than 512 as observed from Figure 4, we have set the sequence length for our models at 256. Our experiments involved reducing the embedding dimension to both 256 and 512, and we found that the results considerably improved with the 512 embedding dimension as information loss due to dimensionality reduction is lower. The specifications used for training are GPU Name: Nvidia P100, GPU Memory: 16GB, GPU Clock: 1.32GHz, CPU Cores: 2, RAM: 12GB, PLATFORM: Kaggle.

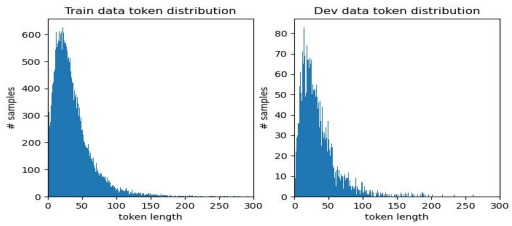

Figure 4: Train and Dev set token distribution

	LR		SVM		MLP		XGBoost		RF	
	Train	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*
LB-L	0.579	0.551	0.587	0.529	0.51	0.509	0.574	0.539	0.55	0.492
LB-AE	0.488	0.375	0.511	0.376	0.365	0.327	0.455	0.368	0.489	0.361
LB+ LDA-L	0.589	0.556	0.592	0.537	0.494	0.495	0.577	0.54	0.569	0.512
LB+ LDA-AE	0.4947	0.377	0.511	0.375	0.385	0.342	0.458	0.374	0.48	0.359
LB+ RB-L	0.61	0.564	0.546	0.546	**0.624**	**0.601**	0.565	0.525	0.526	0.463
LB+ RB-AE	0.491	0.356	0.499	0.353	0.368	0.313	0.44	0.35	0.48	0.354

Table 2: ML Model Experimental Results with embedding dimension 512;
Notation: LB=LegalBERT, **RB**=RoBERTa-base and **-L**, **-AE** denotes the Linear Layer, AutoEncoder Layer as Classifier heads respectively

	Dense		Linear		MBC	
	Train	*Dev*	*Train*	*Dev*	*Train*	*Dev*
LegalBERT (CLS)	0.5768	0.64	0.68	0.66	0.558	0.5439
LegalBERT (meanpool)	0.389	0.361	0.718	0.6584	0.3589	0.341
LegalBERT +LDA (CLS)	0.367	0.341	0.664	0.655	0.447	0.404
LegalBERT + LDA (meanpool)	0.715	0.649	0.672	0.658	0.351	0.341
InCaseLawBERT (CLS)	0.695	*0.66*	**0.717**	**0.669**	0.368	0.341
InCaseLawBERT (meanpool)	0.693	0.648	0.7136	0.6557	0.356	0.333

Table 3: Neural Model Experimental Results

4.2 Results

Table 2 are the results obtained from pre-trained embedding variants without further fine-tuning for the specific task. These models are straightforward to implement and don't require extensive resources. Among the different combinations explored, concatenating LegalBERT embeddings with RoBERTa sentence embeddings and MLP as a classifier has performed the best. However, it's worth noting that these results may be better for tasks with fewer class labels. Table 3 shows the results of the neural classifiers tested, and we achieved a **micro F1 score of 0.6133** on the test data using the LegalBERT (CLS) classifier. In our ongoing experiments, we have found that concatenating LegalBERT embeddings with LDA representation and fine-tuning InCaseLawBERT embeddings trained on Indian judicial documents has resulted in improvement on the dev data.

4.3 Observations

During our examination of the errors made by our models on the dev set, we identified certain patterns of failure. One such example was the role of "PRE_NOT_RELID", which was underrepresented in the dataset and posed difficulties for the model in terms of accurate classification due to poor generalization. Another role, "STA", was also frequently misclassified. This role is a subcategory of "ANALYSIS", and the model often erroneously predicted "ANALYSIS" instead of "STA". Additionally, statements categorized as "NONE" or other groups but containing multiple named entities were often mistakenly labeled as "PREAMBLE". The models struggled to accurately classify certain classes, such as "FAC" and "RATIO", which were often mislabeled as "ANALYSIS". This could be attributed to the model's limited understanding and lack of sufficient patterns for these classes.

5 Conclusion

Automated intermediate tasks have the potential to enhance the capabilities of legal practitioners and accelerate the legal system, particularly in populous countries like India where the number of legal cases is increasing at an almost exponential rate. Our solution for the sub-task involved experimenting with ensemble embeddings and indicated that richer embeddings help in better accuracy and overall task results. However, given the class imbalance observed for a few rhetorical labels, the models misclassify samples belonging to classes that have lower representation. Enriching the embeddings and adding a CRF layer for the classifier would improve the results.

References

Khalid Al-Kofahi, Alex Tyrrell, Arun Vachher, and Peter Jackson. 2001. A machine learning approach to prior case retrieval. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 88–93.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 413–428. Springer.

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.

Ilias Chalkidis and Ion Androutsopoulos. 2017. A deep learning approach to contract element extraction. In *JURIX*, volume 2017, pages 155–164.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 19–28.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

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

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

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

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022a. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022b. Corpus for automatic structuring of legal documents. *arXiv preprint arXiv:2201.13125*.

Guneet Singh Kohli, Prabsimran Kaur, and Jatin Bedi. 2021. Automatic detection of rhetorical role labels using ernie2. 0 and roberta.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nikolaos Lagos, Frederique Segond, Stefania Castellani, and Jacki O'Neill. 2010. Event extraction for legal case building and reasoning. In *Intelligent Information Processing V: 6th IFIP TC 12 International Conference, IIP 2010, Manchester, UK, October 13-16, 2010. Proceedings 6*, pages 92–101. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Shubham Kumar Nigam, Angshuman Hazarika, Arnab Bhattacharya, and Ashutosh Modi. 2021a. Semantic segmentation of legal documents via rhetorical roles. *arXiv preprint arXiv:2112.01836*.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021b. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Guha, Sachin Malhan, and Vivek Raghavan. 2023. SemEval-2023 Task 6: LegalEval: Understanding Legal Texts. In *Proceedings of the*

17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, Canada. Association for Computational Linguistics (ACL).

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022. Pre-training transformers on indian legal text. *arXiv preprint arXiv:2209.06049*.

TYSS Santosh, Philipp Bock, and Matthias Grabmair. 2023. Joint span segmentation and rhetorical role labeling with data augmentation for legal documents. *arXiv preprint arXiv:2302.06448*.

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

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. *ASAIL@ ICAIL*, 2385.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.