NITK_LEGAL at SemEval-2023 Task 6: A Hierarchical based system for identification of Rhetorical Roles in legal judgements

Patchipulusu Sindhu and **Diya Gupta** and **Sanjeevi Meghana** and **Anand Kumar M**
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
{sindhu.191it137, diya.191it215, sanjeevimeghana.191it144,
m_anandkumar}@nitk.edu.in

Abstract

The ability to automatically recognise the rhetorical roles of sentences in a legal case judgement is a crucial challenge to tackle since it can be useful for a number of activities that come later, such as summarising legal judgements and doing legal searches. The task is exigent since legal case documents typically lack structure, and their rhetorical roles could be subjective. This paper describes SemEval-2023 Task 6: LegalEval: Understanding Legal Texts, Sub-task A: Rhetorical Roles Prediction (RR). We propose a system to automatically generate rhetorical roles of all the sentences in a legal case document using Hierarchical Bi-LSTM CRF model and RoBERTa transformer. We also showcase different techniques used to manipulate dataset to generate a set of varying embeddings and train the Hierarchical Bi-LSTM CRF model to achieve better performance. Among all, model trained with the sent2vec embeddings concatenated with the handcrafted features perform better with the micro f1-score of 0.74 on test data. The dataset utilised in our task is available at [1].

1 Introduction

In a country with a large population like India, the number of legal cases has been increasing rapidly. People are overloaded by the vast amount of information and documents available online, which is a result of the expansion of the internet and big data (Kemp, 2014). Legal documents are one example of this type of online information growth in the subject of law. Constitutions, contracts, deeds, orders/judgments/decrees, pleadings, statutes, and wills are some examples of these documents. These documents are very lengthy to read and comprehend and have a structure that is quite elaborate compared to a regular text.

Legal case documents are difficult to process automatically due to their length and unstructured nature If the documents could be divided up into cohesive information pieces, a legal document processing system would greatly benefit. In a legal case document, Rhetorical role labelling of a sentence means identifying the semantic function a sentence of a legal document serves, such as the case's facts, the parties' arguments, ruling from the current court, etc. Semantic search, summarization, case law analysis, and other downstream tasks can all benefit from knowing the rhetorical roles that sentences play in a legal case document. However, legal documents lack a clear organizational structure, making it challenging to comprehend the nuanced differences between rhetorical roles. Rhetorical roles(RRs) aid in the extraction of legal information, such as cases with similar facts. Comparing various rhetorical role units can also be used to find previous cases that are comparable to the one in concern. Using RRs, one may, for instance, extract the key information from the case that affects the judgement.

The goal of this paper is to showcase the proposed methodology as a part of SemEval-2023 Task 6: LegalEval: Understanding Legal Texts Sub-task A: Rhetorical Roles Prediction (RR) competition and also discuss different methods implemented to achieve this task. We secured 16th position out of 27 teams that participated in this task. The task is to segment a given legal document by accurately identifying the rhetorical role of each sentence. This task involves classifying consecutive sentences into many classes using a single label. Evaluation metric being micro F1 score is used for evaluation based on the hidden test results.

In this paper, we use Indian court judgements dataset from SemEval-2023 Task 6: Sub-task A: Rhetorical Roles Prediction (RR). We describe different methods for segmenting a legal document into coherent information units i.e, Rhetorical

[1] https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1154–1160

Roles to help with the processing of lengthy legal documents (Misra et al., 2019). Our corpus has Rhetorical Roles annotated (RRs). Rest of the report organisation is as follows: the following section 2 provides a related work on analysis of the related papers and glimpses about dataset used and kind of input and output to the system. Next section 3 shows the proposed work of the project. Followed by section 4 Analysis and application results are highlighted and in section 5 concludes the work that is discussed at the end of the paper.

2 Related works

A legal document processing system would benefit immensely from being able to segment the documents into coherent informational chunks. Authors in (Malik et al., 2021) proposed a new corpus of legal documents that have been annotated with a set of 13 labels for semantically coherent units (referred to as rhetorical roles), such as facts, arguments, statutes, issues, precedents, rulings, and ratios.

The ability to automatically recognise the rhetorical roles of words in a legal case judgement is an essential topic to work on because it's potentially useful for a broad range of activities that occur afterwards, such as summarising legal judgements and doing legal searches. Automatic Text Summarization helps people save a significant amount of time at work (Kumar et al., 2021) and stay updated on world events by condensing news stories, technical documentation, books, essays, conferences, and meetings to a much more digestible format with little data loss (Tas and Kiyani, 2007; Mehta, 2016). The study in (Moens, 2007) emphasises the significance of summary in increasing the accessibility of court decisions and presents a helpful comparison of several summarization techniques for this job. Authors in (Saravanan and Ravindran, 2010) claim that the final presentation of the summary of a legal document is proven to be improved by understanding of fundamental structures and distinct segments of the document. Authors in (Sun et al., 2019) describes the process of extending BERT, a language model that has already been trained, for text classification tasks. Assigning a label reflecting the rhetorical state of each sentence in a particular segment of a document, this paper offered a new annotation method for the rhetorical structure of legal decisions. Authors in (Xu and Hu, 2022) present a deep learning model for legal text recognition based on the combination of Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) models which could potentially be used in areas like information retrieval and processing legal documents. The creation of appropriate feature sets for the effective usage of CRFs in the process of segmenting (Vlachostergiou et al., 2017), a legal text along different rhetorical functions is also highlighted by these authors. Using a Conditional Random Field (CRF) layer with (Bi-LSTM) network, the authors of (Kim et al., 2020) proposed a model that captures contextual information and dependencies between the named entities. In the latter phases of document summarization, the segmentation of a legal text into certain rhetorical categories may well be employed to improve the outcomes. When compared to straightforward extraction-based summarization, the updated final summary improves the readability and efficiency.

2.1 Dataset

Indian court judgements dataset from SemEval-2023 Task 6: Sub-task A: Rhetorical Roles Prediction (RR) competition (Modi et al., 2023) is used in this paper. The dataset consists of 246 legal documents in json format that were judged by an Indian court as training data and 50 legal documents as test data. Each sentence of every document was classified into one of the 13 pre-defined rhetorical roles by law students from various Indian law institutes who voluntarily annotated the data. The 13 labels are: Preamble, Facts, Ruling from lower court, Argument by Petitioner, Argument by Respondant, Issues, Analysis, Statute, Precedent Relied, Precedent Not Relied, None, Ratio of decision and Ruling by present court.

2.2 Data Formatting

Data is given in a single json file, where for a given set of case documents, all sentences included in a document, the corresponding labels(basically, RR's) assigned and few other details are mentioned. The train json file contains 246 legal documents in a single json file. We converted this json file into 246 separate text files corresponding to 246 legal documents, where the i'th file contains all the sentences with their respective labels included in the i'th document. The test json file contains 50 legal documents in a single json file. Similar to train data, test data is also converted into 50 text documents. While training, for each of these 246 files, an embedding file is created that contains sentence

embeddings of all the sentences along with their assigned labels. These files are fed to the system for training. While testing, embeddings files without labels assigned for all 50 documents are sent one-by-one. The output for every embedding file is again a text file containing all the sentences with their respective predicted labels. The predicted output files from the test data is reformulated into json format, to a single json file and is submitted in the competition submissions page.

3 Methodology

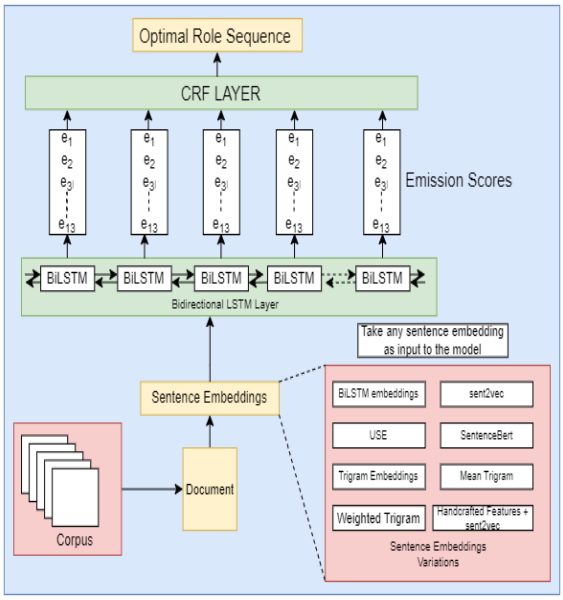

Figure 1: Hierarchical BiLSTM CRF Architecture

Systems proposed for sub - Task A of LegalEval: Understanding Legal Texts were based on Hierarchical BiLSTM CRF architecture and Roberta based transformer model. The problem of identification of rhetorical roles for the provided legal judgements was treated as a 13 - class sequence labelling problem where supervised Machine Learning models are used to predict one label (rhetorical role) for every sentence in a document and 13 classes specifically refer to the 13 rhetorical labels provided in the dataset. Input provided to the model are text files created separately for each judgement case from the provided training dataset in json format.

3.1 Hierarchical BiLSTM CRF

Conditional Random Fields (CRF): Each document i.e. legal judgement case is treated as a sequence of sentences and in a legal scenario there are some dependencies in the order in which the labels are placed; for instance, RLC often comes after FAC and RPC is always the last label. Since Conditional Random Fields (CRFs) (Lafferty et al., 2001) take into account both emission scores (probability of a label given the sentence) and transition scores (probability of a label given the previous label) when generating the label sequence, they can be utilised to describe such sequences.

We made use of Hierarchical BiLSTM (Graves et al., 2005) architecture with the CRF (Lafferty et al., 2001) layer deployed on top of it to automatically extract the sequence of rhetorical roles for the provided judgement case as shown in Figure 1. Hierarchical BiLSTM classifier takes into account the full section at once. In order to achieve this, we must feed the BiLSTM with the sequence of sentence embeddings, which generates a sequence of feature vectors whose hidden states are treated as context-aware sentence embeddings. The sentence embeddings in the BiLSTM model need to be initialised before learning can begin. Here we have tried different variations of sentence embeddings(All the sentence embeddings mentioned in Figure 1 are used in this paper):

- **BiLSTM Embeddings:** For the now dataset in form of text files, sentence embeddings are created from randomly initialised word embeddingsfor every judgement using another BiLSTM.
- **sent2vec Embeddings:** Generated sentence embeddings of 512 length using sent2vec (Naser Moghadasi and Zhuang, 2020) where pretrained weights of 'distilbert-base-uncased' were used as an underlying mechanism to encode the text data.
- **Universal Sentence Encoder (USE):** Generating sentence embeddings for the provided document using Universal Sentence Encoder of 512 length that transforms text into high-dimensional vectors that can be applied to tasks involving natural language, such as text categorization, semantic similarity,clustering, and others.
- **SentenceBERT:** Made use of SentenceBERT to generate sentence embeddings of 512 length which has an underlying Siamese architecture, which processes two sentences in

pairs during training and has two BERT architectures that are virtually identical and share the same weights.

- **Trigram Embeddings ($s_{n-1}s_ns_{n+1}$):** To capture the hierarchical information through embeddings, we merged current sentence embedding with preceding(s_{n-1}) and succeeding sentence(s_{n+1}) embeddings, in whole generating sentence embedding of 2304 length ($s_{n-1}s_ns_{n+1}$).

- **Mean Trigram** (mean_$s_{n-1}s_ns_{n+1}$)**:** Generated sentence embedding for the current sentence by taking the mean of already generated sentence embeddings of current (s_n), previous (s_{n-1}) and next (s_{n+1}) sentence through sent2vec mechanism.

- **Weighted Trigram**(weighted_$s_{n-1}s_ns_{n+1}$)**:** Generated new embedding for the current sentence (s_n) by capturing information from its preceding (s_{n-1}) and succeeding (s_{n+1}) sentence embedding and giving importance by assigning weights as $0.25s_{n-1} + 0.5s_n + 0.25s_{n+1}$.

- **Handcrafted Features + sent2vec:** Made use of Bag of Words model to capture specific entities (considering these as attributes) related to the particular label along with other set of features such as all capital letters, entity recognition. The features are created by assigning 1 if the specified entities are present in a sentence else 0 is assigned to generate a sentence embedding and concatenating the same with the already generated pretrained sent2vec embeddings. For the provided labels, entities accounting in total to 70 are considered as attributes depending on the frequency are shown in Table 1.

The latter are then forwarded to a multinomial linear layer, which generates a probability for each class for each sentence in the section i.e. emission scores represented as e_1,e_2, e_3 till e_{13} in Figure 1. Hierarchical BILTM functions better since it takes into account the full section at once. The probability scores produced by the model can be viewed as simple emission scores because they do not take label dependencies into consideration. We deploy a CRF on top of the Hierarchical BiLSTM architecture to further enliven the model. The feature vectors produced by the top-level BiLSTM are input into this CRF which finally provides the sequence of rhetorical labels as output for the given document.

In the submitted run, the model is trained over 300 epochs with training data consisting of 246 judgements. To find the optimal parameters, the model is fine tuned and Adam Optimizer with a learning rate of 0.01 is used while training. Further, to prevent overfitting, cross validation with variations in the number of folds to be 3 and 5 is used. In order to observe how well the model performs, the provided dataset is split in the ratio of 80:20 as training data and validation data for training the model and to assess its performance.

3.2 Transformers

The other proposed technique makes use of a modified pretrained RoBERTa (Majumder and Das, 2020) (Robust and Optimised BERT Pretraining Approach) encoder with an additional linear layer added to the pretrained RoBERTa base model. RoBERT was created as an improvement over BERT (Devlin et al., 2018) by offering sophisticated masked language modelling and greatly expanding the amount of training data. The resulting transformer model (a pre-trained base model from the RobertaForSequenceClassification class) is fine tuned over the training dataset to arrive at the optimum parameters.

In the submitted run, a single linear layer for classification that served as a sentence classifier was added to the original pretrained base RoBERTa model in (Liu et al., 2019). The training data were sent in batches of 16, and the test data were fed into the model in proportions of 80 and 20 , respectively. The entire pretrained RoBERTa model and the additional untrained classification layer were then fine-tuned for the specific downstream task of categorising the legal documents into one of the thirteen Rhetorical Roles. With an epsilon value set to 1e-8 and a learning rate of 2e-5, Adam Optimizer (Kingma and Ba, 2014) was used while training. The model was fine-tuned over 4 epochs using a seed value of 42.

4 Results and Analysis

During the 21 days of evaluation period (January 10 through 31, 2023), 27 CodaLab users submitted around 174 submissions for RR shared tasks. Each team has the eligibility of a maximum 20 submissions with maximum submissions of 2 per

Labels	Entities
PREAMBLE	State, appellant, Commissioner, appeal, Incometax, Criminal
FAC	accused, appellant, dated, assessee, deceased, appeal, petitioner, stated, complainant
RLC	imprisonment, offence, learned, held, sentenced, appellant, IPC, Tribunal, convicted
ISSUE	accused, circumstances, question, assessee, proves, committed, entitled
ARG_PETITIONER	learned, submitted, counsel, petitioner, accused, evidence, State, appellant, contended
ARG_RESPONDENT	accused, counsel, submitted, evidence, would, State, respondent, bail, deceased
ANALYSIS	accused, evidence, would, stated, prosecution, assessee, appellant, question
STA	person, offence, subsection, clause, State
PRE_RELIED	State, held, accused, would, evidence, question, decision, assessee, Commissioner
PRE_NOT_RELIED	decision, State, Commissioner, judgment, declaration, statement, question, Constitution, cheque
RATIO	evidence, prosecution, appellant, offence, circumstances, question, present
RPC	appeal, shall, dismissed, allowed, bail, accused, petitioner, appellant, sentence, offence, judgment, accordingly, following, result
NONE	learned, dated, judgment, JURISDICTION, APPELLATE, Criminal, counsel

Table 1: Handcrafted Features for different rhetorical roles

Method	micro f1-score
sent2vec+ Handcrafted features	**0.74**
weighted Trigram	0.731
mean Trigram	0.729
sent2vec	0.721
BiLSTM embeddings	0.696
RoBERTa Transformer	0.125

Table 2: Test scores

day. Given the intricacy of the task, we have made 10 submissions during the evaluation period. The output scores of each submission is shown in Table 2 along with the baseline model (SciBERT-HSLN). Here, our system got the highest score of 0.74, which performs competitively with the score of the baseline model.

Table 3 shows the precision, recall and micro f1-scores of different embeddings as input as mentioned in section 3.1, to the Hierarchical Bi-LSTM CRF model with cross validation of folds = 3 and 5, respectively and for random testing of the system considered 20% of the provided dataset as test data and remaining 80% as the training data. As per the experiment, sentence embeddings which consist of handcrafted features concatenated with sent2vec embeddings capturing the contextual inference outstands the other sentence embeddings in all the variations (with cross validation and Random testing) with micro f1-scores of 0.770, 0.767 and 0.774 respectively, while sentence embeddings of length 2304 capturing the information of prior and next in line sentence didn't perform well. 20% of the dataset is considered as test data in case of randome testing and the remaining 80% is used for training the model.

5 Conclusion and Future Work

In this paper, we describe our knowledge based system for the SemEval-2023 Task 6: LegalEval: Understanding Legal Texts Sub-task A: Rhetorical Roles Prediction (RR), that ranked 16 out of 27 teams participated in the task. To predict the rhetorical role of every sentence of legal documents of the supreme court of india, we have used the Hierarchical Bi-LSTM CRF model. Trained the model over different input embeddings like random weights, sent2vec, SBERT, USE, Trigram, mean Trigram, weighted Trigram and sent2vec along with handcrafted features for the better performance and ob-

Embeddings	CV (folds=3)			CV (folds=5)			Random Testing		
	P	R	micro F1	P	R	micro F1	P	R	micro F1
BiLSTM embeddings	0.730	0.743	0.734	0.744	0.753	0.743	0.743	0.751	0.742
sent2vec	0.772	0.773	0.765	0.776	0.759	0.762	0.763	0.758	0.758
SentenceBERT	0.722	0.737	0.724	0.739	0.751	0.742	0.740	0.756	0.742
USE	0.708	0.722	0.712	0.743	0.749	0.744	0.732	0.739	0.732
Trigram	0.657	0.668	0.646	0.686	0.684	0.663	0.586	0.552	0.557
mean Trigram	0.750	0.756	0.750	0.759	0.763	0.752	0.750	0.739	0.739
weighted Trigram	0.754	0.759	0.753	0.749	0.738	0.741	0.768	0.765	0.761
sent2vec+Handcrafted features	0.771	0.769	**0.770**	0.770	0.775	**0.767**	0.773	0.779	**0.774**

Table 3: P (Precision), R (Recall) and micro F1-scores of different embeddings as input to the Hierarchical Bi-LSTM CRF model for training with CV (Cross Validation) of folds = 3 and 5 and Random testing.

served that 'sent2vec with handcrafted features' performs better as compared to other embeddings. In addition to the layers in the Hierarchical Bi-LSTM CRF, we have added an attention layer to the model to give prior importance for the few parts of the data. The precision, recall and F1 scores for all these embeddings are calculated with the base model (Hierarchical Bi-LSTM CRF model) and RoBERTa transformer method is also implemented and are compared to find the best suitable model and the best embedding that gives the best labelling of the rhetorical roles of the sentences in the data. Here we have observed that out of all, the sent2vec embeddings with the Hand crafted features gave the best micro f1-score compared with the others. Furthermore, we can improve the model by adding different layers to the current model with the different embeddings for better performance and result.

References

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*.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II 15*, pages 799–804. Springer.

Richard Kemp. 2014. Legal aspects of managing big data. *Computer Law & Security Review*, 30(5):482–491.

Gyeongmin Kim, Chanhee Lee, Jaechoon Jo, and Heuiseok Lim. 2020. Automatic extraction of named entities of cyber threats using a deep bi-lstm-crf network. *International journal of machine learning and cybernetics*, 11:2341–2355.

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

Yogesh Kumar, Komalpreet Kaur, and Sukhpreet Kaur. 2021. Study of automatic text summarization approaches in different languages. *Artificial Intelligence Review*, 54(8):5897–5929.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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*.

Soumayan Bandhu Majumder and Dipankar Das. 2020. Rhetorical role labelling for legal judgements using roberta. In *FIRE (Working Notes)*, pages 22–25.

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

Parth Mehta. 2016. From extractive to abstractive summarization: A journey. In *ACL (Student Research Workshop)*, pages 100–106. Springer.

Shivanshu Misra, Siddhartha Bhattacharya, S Saravana Kumar, B Deepa Nandhini, S Christinajoice Saminathan, and P Praveen Raj. 2019. Long-term outcomes of laparoscopic sleeve gastrectomy from

the indian subcontinent. *Obesity Surgery*, 29:4043–4055.

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).

Marie-Francine Moens. 2007. Summarizing court decisions. *Information processing & management*, 43(6):1748–1764.

Mahdi Naser Moghadasi and Yu Zhuang. 2020. Sent2vec: A new sentence embedding representation with sentimental semantic. pages 4672–4680.

M Saravanan and Balaraman Ravindran. 2010. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18:45–76.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

Oguzhan Tas and Farzad Kiyani. 2007. A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213.

Aggeliki Vlachostergiou, George Marandianos, and Stefanos Kollias. 2017. From conditional random field (crf) to rhetorical structure theory (rst): Incorporating context information in sentiment analysis. In *The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28–June 1, 2017, Revised Selected Papers 14*, pages 283–295. Springer.

Hesheng Xu and Bin Hu. 2022. Legal text recognition using lstm-crf deep learning model. *Computational Intelligence and Neuroscience*, 2022.