NLP-Titan at SemEval-2023 Task 6: Identification of Rhetorical Roles Using Sequential Sentence Classification

Harsh Kataria
Cisco Systems
hkataria99@gmail.com

Ambuje Gupta
Uniphore Systems
ambujegupta99@gmail.com

Abstract

The analysis of legal cases poses a considerable challenge for researchers, practitioners, and academicians due to the lengthy and intricate nature of these documents. Developing countries such as India are experiencing a significant increase in the number of pending legal cases, which are often unstructured and difficult to process using conventional methods. To address this issue, the authors have implemented a sequential sentence classification process, which categorizes legal documents into 13 segments, known as Rhetorical Roles. This approach enables the extraction of valuable insights from the various classes of the structured document. The performance of this approach was evaluated using the F1 score, which measures the model's precision and recall. The authors' approach achieved an F1 score of 0.83, which surpasses the baseline score of 0.79 established by the task organizers. The authors have combined sequential sentence classification and the SetFit method in a hierarchical manner by combining similar classes to achieve this score.

1 Introduction

Legal cases are often complex, lengthy and intricate, posing significant challenges for researchers, practitioners, and academicians alike. These challenges become even more pronounced in developing countries like India, where the number of legal cases is growing exponentially. The National Judicial Data Grid of India (NJDG, 2020) shows more than 43 million pending legal cases. These result in legal documents that are unstructured and highly complex, making them difficult to process efficiently. This requires modern solutions like new-age technology and AI advancements to help process this vast backlog of pending legal cases efficiently.

Legal documents differ considerably from the text on which pre-trained Natural Language Processing (NLP) models are trained. Legal documents are typically large, unstructured, and include legal jargon, as well as numerous mistakes. Additionally, different categories of legal cases are fundamentally different from one another. To overcome this challenge, NLP models trained on accurate well-annotated corpus are necessary. However, creating such a corpus itself is challenging, resulting in slow growth in the Legal NLP domain.

To efficiently process long legal documents, this paper proposes a method for rhetorical role classification using NLP techniques to segment legal texts from legal documents into semantically coherent classes. The proposed method is a text segmenting task which involves classifying parts or sentences of a document. The task uses a corpus of annotated legal documents, which have the spans of the document annotated with the appropriate class or rhetorical role. Sample rhetorical roles have been depicted in figure 1. A method to merge similar classes at the base level was implemented to aide the actual task. The sequence sentence classification was carried out using sequential sentence classifier and SetFit method is used to classify the merged classes.

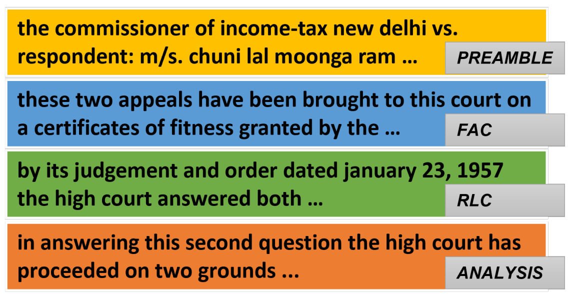

Figure 1: Sample Rhetorical Roles

The 13 rhetorical roles in the corpus include None, ANALYSIS, ARG PETITIONER, ARG RESPONDENT, FAC, ISSUE, PREAMBLE, PRE NOT RELIED, PRE RELIED, RATIO, RLC, RPC, and STA. The task aims to achieve a high F1 score

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1365–1370
July 13-14, 2023 ©2023 Association for Computational Linguistics

in the segmentation of these rhetorical roles from legal documents.

This method for rhetorical role classification using NLP techniques provides a solution to process long legal documents efficiently. However, creating an accurate and well-annotated corpus of data is a significant challenge in the Legal NLP domain. The proposed method can help to overcome this challenge by segmenting legal texts from legal documents into semantically coherent classes, enabling the efficient processing of legal documents.

2 Background Research

With the development in the field of Deep Learning, legal text processing has been an active topic of research. Some of the examples of work in legal domain are summarization of legal documents (Jain et al., 2021) summarization for Legal Judgment and segmenting Rhetorical roles (Saravanan and Ravindran, 2010), headnote generation (Mahar et al., 2021)

The classification of scientific articles based on their rhetorical role is an active research area in natural language processing. This approach involves identifying the function or purpose of individual sentences or paragraphs within a scientific article and then using this information to categorize the article into a specific rhetorical role. Teufel and Moens (Teufel and Moens, 2002) made a significant contribution to scientific discourse by introducing the concept of rhetorical role for scientific articles. In their research paper, they presented a hierarchical model that identified different levels of granularity, such as sections, paragraphs, and sentences within scientific articles. This model was used to classify scientific articles into one of six rhetorical zones representing distinct communicative functions. In the field of scientific discourse analysis, another method for classifying scientific articles based on rhetorical roles involves the use of argumentation schemes. Palau and Moens (Palau and Moens, 2009) proposed an approach that employed argumentation schemes to determine the function of individual sentences within scientific articles. Their research paper demonstrated the effectiveness of this approach in identifying the rhetorical roles of scientific articles. By utilizing argumentation schemes, Palau and Moens' method provides a structured and systematic way of analyzing the persuasive elements of scientific discourse. In (Ghosh and Wyner, 2019), authors categorized rhetorical roles into seven distinct categories and compared the effectiveness of neural network-based approaches with traditional methods. These results revealed that the neural network-based approach outperformed the traditional methods in identifying the rhetorical roles. The rhetorical role task can also be seen as sequence sentence classification, i.e. classifying a document with various classes. Previously, hierarchical sequence encoders such as LSTMs (Hochreiter and Schmidhuber, 1997) were utilized for encoding individual sentences and contextualizing these encodings. Afterwards, a CRF was applied on top of the contextualized encodings as seen in (Jin and Szolovits, 2018) which used medical data from (Dernoncourt and Lee, 2017).

3 Experimental Setup

3.1 Data

The task organizers for this task provided the data and a baseline result on it. The dataset was meticulously annotated, consisting of two JSON files, train and dev. Additional details regarding this are available in the task description paper (Modi et al., 2023) by the task organizers. The training data comprised 247 cases, while the dev data included 30 cases. The cases belonged to either the 'Tax' or 'Criminal' category. Every case includes the full case text, which was the complete case document. It also had spans of categorized rhetorical roles at a character level, with the role labels and the sentence itself.

The authors implemented a text cleaning pipeline, which included the removal of HTML tags, punctuations, accent characters, URLs, mentions, tags, special characters, and stopwords, along with the expansion of contractions and conversion of all text to lowercase, followed by lemmatization. Upon following this process, no significant improvement of the F1 score was observed, due to which the authors decided to move on with discarding this pipeline.

The legal judgment documents comprise the dataset JSON files used in this study. These were extracted from the archives of Indian legal cases and originated from various states and courts, including the Supreme Court of India, high courts, and district-level courts. The composition of the dataset in terms of the number of spans identified as a particular rhetorical role is presented in Table 1. It is observed here that the PREAMBLE role

Rhetorical roles	Count of Text
['ANALYSIS']	401
['ARG_PETITIONER']	214
['ARG_RESPONDENT']	60
['FAC']	2234
['ISSUE']	155
['NONE']	437
['PRE_NOT_RELIED']	4
['PRE_RELIED']	38
['PREAMBLE']	3665
['RATIO']	12
['RLC']	262
['RPC']	49
['STA']	53
Grand Total	7584

Table 1: Count of rhetorical roles in corpus

and FACT role account for 48.3 percent and 29.4 percent respectively of the total spans. More information about the definition of rhetorical roles can be found on Legal-NLP-EkStep's rhetorical role baseline github repository (Legal-NLP-EkStep, 2021).

3.2 Experiment Environment

The authors conducted their experiments in a Python environment using a combination of scripting and various Python libraries such as NumPy, pandas, torch, and transformers. They utilized Jupyter Notebook as their Python IDE and version 3.8 of Python. To train their models, they also employed Google Colab, including both GPU and TPU. They also utilized a Tesla V100 hosted on an AWS server.

3.3 Evaluation metrics

The authors evaluated the model's performance using F1 score, precision, and recall (Bernier-Colborne and Goutte, 2020). The task organizers (Modi et al., 2023) also applied F1 score to rank the challenge responses.

4 System Overview

The proposed method in this paper explores the usage of modern Natural language processing techniques and models, namely, Hierarchical Sequential Labeling Network (HSLN) (Jin and Szolovits, 2018) and BERT (Devlin et al., 2019) base embeddings to accomplish the task of sequential sentence classification into rhetorical roles from a legal document. The following text describes the above-mentioned techniques.

4.1 Bert and SetFit

BERT (Devlin et al., 2019) is a deep learning model that is pre-trained and bidirectional, has achieved remarkable performance on numerous NLP tasks. BERT-Base-Uncased is a variation of the BERT model that has been trained on an extensive collection of uncased text.

SetFit (Tunstall et al., 2022) is a classification model that is capable of few-shot learning. It is composed of a sentence transformer model followed by a classification head. Using the sentence transformer model, SetFit generates dense embeddings on paired sentences, enabling effective classification. During the initial fine-tuning phase, the Sentence Transformer model utilizes contrastive training to train on a limited set of labeled input data. This involves generating positive and negative pairs through in-class and out-class selection and subsequently training on these pairs (or triplets) to generate dense vectors per example. In the subsequent stage, a classification head is trained on the encoded embeddings along with their corresponding class labels. During inference, an unseen example undergoes fine-tuning through the Sentence Transformer model, generating an embedding that is then input to the classification head, resulting in a predicted class label.

4.2 Methodology

To start with, the authors combined 13 categories into 10 by combining those that exhibited similarities. The classes that were merged are shown in table 2.

Combined Class	Original Classes
AST	STA
	Analysis
REL	Pre-Relied
	Pre Not Relied
RXC	RLC
	RPC

Table 2: Combined Classes from Original Classes

The authors have utilized a **two-stage approach** to achieve their results and the approach is explained below:-

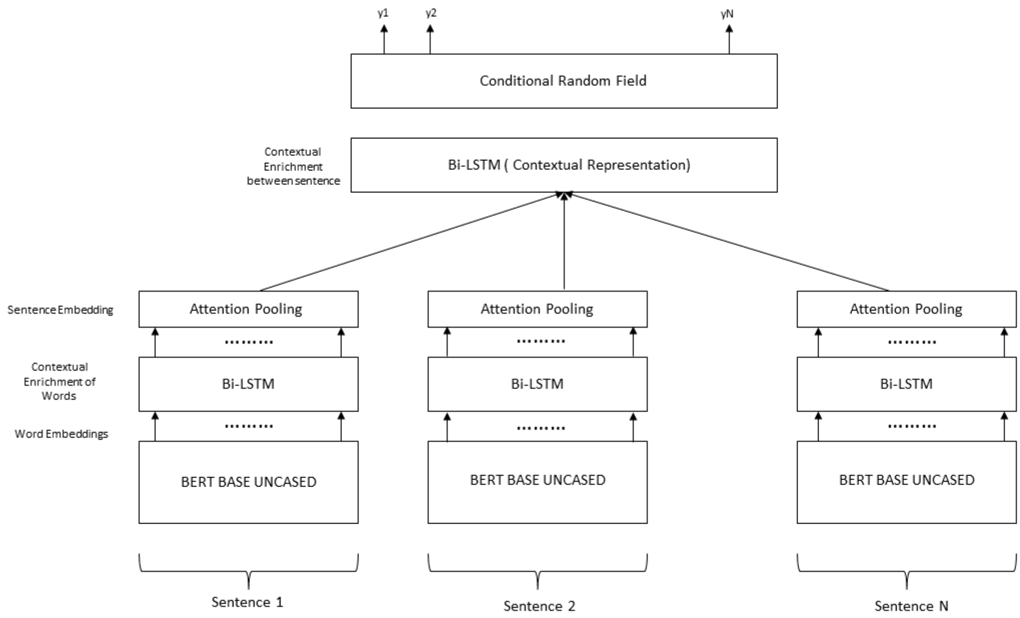

Figure 2: HSLN Architecture

(i) **Sequence Sentence Classifier**: The objective of the sequence sentence classification algorithm is to classify sentences in a legal document (sent1, sent2, ..., sentn) into corresponding labels (e.g., Preamble) by utilizing the contextual information of surrounding sentences. In this work, after merging, there were a total of 10 classes. The sequence sentence classification algorithm is based on a Hierarchical Sequential Labeling Network (HSLN) (Jin and Szolovits, 2018). The architecture of the HSLN can been seen in the figure 2. The authors selected HSLN as the base approach due to its capability to handle documents of any length, which is a crucial consideration in the domain of legal documents that tend to be extensive. The proposed HSLN algorithm employs a multi-step approach. The proposed HSLN model's architecture is explained as follows:-

   (a) WordEmbedding: - First, words (t1i, t2i, ..., tni) within a sentence (Si) are transformed into word embeddings to encode sentences into a numerical representation. Bert Base Uncased model has been used to obtain these word embeddings.

   (b) Sentence Representation: The output from the BERT model is fed into a Bi-LSTM network to capture the contextual meaning within the sentence. An attention pooling layer is applied to derive sentence embeddings

   (c) Contextual Representation: Finally, these sentence embeddings are input into another Bi-LSTM network to extract contextual information from surrounding sentences and obtain a contextual representation of the input sentence.

   (d) CRF: These are then fed to a CRF layer which helps in prediction using the contextual data.

   To mitigate overfitting, the authors have implemented Dropout after each layer in the model. To minimize computational complexity (GPU), the parameters of the Bert-base-uncased model were frozen.

(ii) **SetFit** : The SetFit methodology was used to classify the merged classes back to their original categories. The three merged classes namely AST, REL, and RXC were trained using SetFit. These are binary classes for STA-Analysis, Pre-relied-Pre-not-relied and RLC-RPC, respectively. The authors have developed distinct SetFit models for each merged class. SetFit enables us to perform efficient classification on these binary classes and helps achieve higher accuracy. The result from these three SetFit models combined with the results from the first step provides us with the final predicted Rhetorical roles for the document.

In conclusion, as per figure 3, the complete process entails submitting the legal document to the sequential sentences classifier, which has been trained with merged classes. Following prediction, the merged classes are conveyed to their corresponding SetFit classifiers to generate the results. Finally, both results are integrated to obtain an accurate classification of legal documents based on their specific Rhetorical Roles.

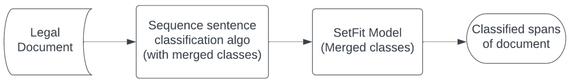

Figure 3: Process Flow

5 Results and Conclusion

Prior to attaining optimal performance, diverse transformer-based models were employed. The model underwent training over a span of 70 epochs, utilizing an Adam optimizer with a learning rate of 3e-4. To preclude overfitting and excessive GPU usage, the authors implemented a scheduler mechanism whereby the learning rate automatically diminishes if the F1 score does not exhibit improvement for five consecutive epochs.

Model	Transformer Name	Merged Classes	F1 score (unseen data)
Baseline	LegalBert	No	0.79
Model 1	LegalBert	No	0.795
Model 2	LegalBert	Yes	0.81
Model 3	Bert-base-uncased	Yes	0.8301
Model 4	XLM-Roberta	Yes	0.82

Table 3: Experimental results

Table 3 presents a compilation of distinct models and their corresponding F1 scores. The results indicate that all the models exceeded the baseline performance, with the optimal approach being the third model in which class merging was employed, followed by SetFit, utilizing the Bert-base uncased model, resulting in a 5% increase in the F1 score relative to the provided baseline. This approach yielded an F1 score of 0.8301.

The classification report for the training data is presented in Table 4 (best model). The overall F1 score for the training data was 0.8322,

Classes	Precision	Recall	F1
ARG_PETITIONER	0.5075	0.4857	0.4964
ARG_RESPONDENT	0.75	0.3158	0.4444
AST	0.7971	0.8814	0.8372
FAC	0.805	0.8828	0.8421
ISSUE	0.8333	0.8	0.8163
NONE	0.9508	0.9158	0.933
PREAMBLE	0.996	0.9902	0.9931
RATIO	0.5968	0.5286	0.5606
REL	0.8289	0.4091	0.5478
RXC	0.7725	0.6232	0.6898
accuracy			0.8322
macro avg	0.7838	0.6832	0.7161
weighted avg	0.832	0.8322	0.8254

Table 4: Classification report for train data (Best model)

with the Preamble and None classes exhibiting the highest F1 scores of 0.99 and 0.93, respectively. The FAC, AST, and Issue classes followed with F1 scores of 0.84, 0.83, and 0.81, respectively. However, the model encountered challenges differentiating between the ARG_PETITIONER and ARG_RESPONDENT classes. The authors identified two potential reasons for this difficulty: insufficient training data and the high similarity between these classes, which could result in interchangeable data depending on the specific case.

The tasks carried out in this paper were part of Task 6: LegalEval in Semeval 2023. The aim was to classify rhetorical roles in a legal document. The authors were able to score the 6th rank in the competition by applying the above mentioned approach. It was identified that merging of the classes at a base level helps in classification. Also note that text cleaning on this data had minimum improvements.

With further research and development, the proposed method can potentially improve the efficiency of legal systems in developing countries like India, where the backlog of pending legal cases is ever-increasing. The method for segmentation of rhetorical roles can also further help in data extraction tasks, summarization tasks, judgment prediction and other such tasks as legal entity recognition.

References

Gabriel Bernier-Colborne and Cyril Goutte. 2020. Challenges in neural language identification: NRC at VarDial 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 273–282, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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

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.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. *arXiv preprint arXiv:1808.06161*.

Legal-NLP-EkStep. 2021. Rhetorical role baseline appendix.

Sarmad Mahar, Sahar Zafar, and Kamran Nishat. 2021. Headnote prediction using machine learning. *Int. Arab J. Inf. Technol.*, 18(5):678–685.

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).

NJDG. 2020. National judicial data grid.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

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.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.