Legal Document Segmentation

Project Proposal

Motivation and Problem Statement:
The surge in legal cases poses challenges for a timely judicial process. While complete automation may be impractical, automating intermediate tasks can significantly assist legal practitioners. Processing legal papers is hindered by their unstructured nature, legal jargon, and length. This project aims to create a Rhetorical Roles (RR) system for segmenting legal texts into coherent sections such as facts, arguments, legislation, issues, precedent, ruling, and ratio.

Relevance of Rhetorical Roles Prediction:
Creating a rhetorical role corpus plays a pivotal role in automating the understanding of legal documents by breaking them down into coherent units. This segmentation is foundational for various legal AI applications, including but not limited to judgment summarization, prediction of judgment outcomes, and precedent search. Predicting rhetorical roles enhances the comprehension of the structure and content of legal texts, contributing to the efficiency of automated systems within the legal domain.

Task Overview:
Given the inherent challenges posed by lengthy and unstructured legal documents, our proposed task aims to automatically segment legal judgment documents into semantically coherent text segments. Each segment is assigned a specific label, such as preamble, fact, ratio, arguments, etc., collectively referred to as Rhetorical Roles (RR). The task at hand is Rhetorical Role Prediction, involving the segmentation of a given legal document by predicting the rhetorical role label for each sentence. This task is structured as a sequential sentence classification with a single label and multiple classes.

Dataset:
The dataset comprises Indian court judgment documents, particularly from the SemEval 2023 Task - LegalEval. The task involves segmenting the document into logical sections, or Rhetorical Roles, including Preamble, Facts, Ruling by Lower Court, Issue, Argument by Petitioner, Argument by Respondent, Analysis, Statute, Precedent Relied, Precedent Not Relied, Ratio of the Decision, Ruling of Present Court, and None.

Tentative Evaluation Metric: Micro F1 Score

Tentative Approach:

- Assess the current baseline in SemEval 2023 (Task 6).
- Emphasize thorough preparation and pre-processing of the dataset for effective understanding.
- Explore diverse word embedding techniques for pre-processing, fine-tune baseline models, and attention-based text classification models. Evaluate these approaches using the provided dataset.
- Experiment with various combinations of pre-processing methods, embedding techniques, and sequence-to-sequence models, with a focus on attention-based models. Present the results obtained from these experiments.