IndLegal - Legal Text Summarizer

Abstract

This project report describes the design, implementation, assessment, and possible uses of a BART summarizer system that has been adapted for Indian legal cases. Such a summarization model has been designed with the purpose of distilling long legal manuscripts into brief overviews while saving the key words, the important dates, and other necessary information.

The report includes a description of the architecture of the model and the training flow, explaining how specific cases and the corresponding summaries were prepared and used to adapt the BART model for the specific features of Indian legal language and content. From the design of the system, it is possible to expect that raw legal text can be easily processed, and an appropriate summary with the relevance of significant legal terms, case specific details and other information relevant to the objective of practitioners and scholars is produced.

Observations and analysis of model assessments and extensive dataset evaluations, impact of legal case types in the construction of the summaries is addressed in this report. Furthermore, scope and limitations of this system are explained, for instance, its use in enhancing legal research and case law studies as well as increasing the effectiveness of the judiciary by quickly presenting the summary of a given case.

Finally, this project describes an innovation in summarization of legal documents in the year 2022; it points out improvements that can be achieved, particularly expansion of languages supported, enhancement on the accuracy of keywords, and linking up with legal portals for easy access to data, increasing acceptability in the legal profession.

Introduction

In the introduction to this project report, the problem of the necessity for a better and more advanced operating system to process, comprehend and summarize Indian legal documents is raised. Legal cases are usually quite lengthy running into several hundreds of pages containing descriptions, a timeline of events, arguments within a case, legal provisions and the decision of a court. For attorneys, magistrates, and scholars, such long documents are uneconomical in terms of time and resources spent reviewing them. Thus, the principal aim of this project is to develop a strategy based on natural language processing (NLP) and machine learning capable of shrinking the workload of legal researchers by providing them with summaries of case files.

This project proposes a BART combiner summarizer with the emphasis on accurate extraction of legal information. BART – Bidirectional and Auto-Regressive Transformers is a particular sequence-to-sequence model that is well designed to perform summarization. BART was first trained as a denoising autoencoder; it is now widely applied for high-level summary purposes on long-form text due to its ability to keep the text organization and

factual information intact, which are very important aspects in the law industry. The BART model was initially pre-trained on a large amount of data and then the BART-base model was chosen for this project to fine-tune on Indian laws cases because it provides the most favorable tradeoff of cost and quality of index. Using BART-base makes it possible for the model to handle large volumes of legal texts while losing only a minimal amount of case-related information, an aspect that is very vital in the legal field which enterprises a lot of information precision and completeness.

During the fine-tuning stage, the process of designing and organizing a database of court case summaries from the Indian legal system was undertaken. This database aims to facilitate the translation of the BART architecture into the use of the Indian legal language, legal concepts, and practice. As necessary, each document was modified for the preservation of important case elements like key dates, names, relevant law, as well as the decision, incurring little loss of important aspects. As expected, the main reason for training BART with this data was the desire to emphasize the language style and the vocabulary linked to Indian legal provisions used in the summarizer herself.

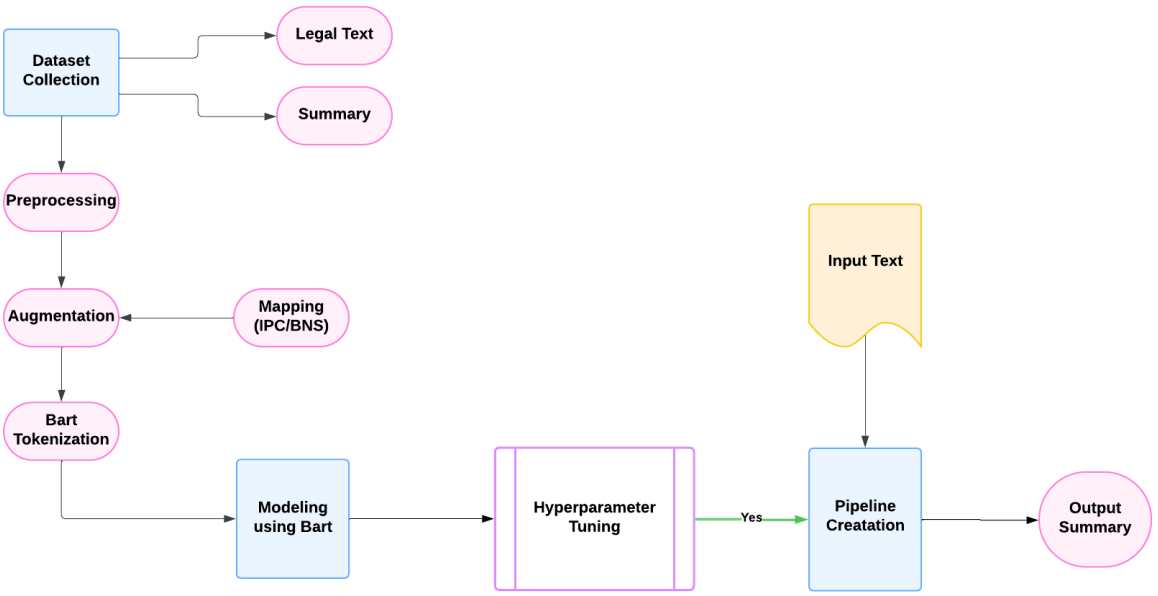
In addition to the model architecture, this project gives importance towards hyperparameter tuning and data management aspects of the problem, which are also two factors that help in making the summarizer more efficient and accurate. It was important to select hyperparameters, such as batchsize, learning rate, warmup steps, gradient accumulation steps, etc., in such a way that they allow for stable training of the model, while also being efficient in terms of memory and time. A Data Collator for Seq2Seq was also used to accommodate the other span of lengths since legal documents tend to differ in sizes whilst ensuring consistency in the input-output order during training. The training regime incorporated evaluation snapshots and gradient accumulation enabling the model to be trained on large batches of data without excessive computational resources.

The performance of the model attained after the training started, was assessed on a comprehensive test dataset of Indian legal cases, focusing on metrics measuring the ability of the model to summarize the document by retaining significant details while making it concise. The performance of the summarizer proved to be satisfactory through both the number-based surveys and the analysis of the short, related and purposeful summaries produced by the model in question. The assessment consisted of the monitoring of the summarization quality within the parameters of different cases, the extent to which the model learned to handle different types of legal narratives, and the quality of elicitation itself in terms of preservation of important legal elements such as judgment dates, statutes, and case names.

When it comes to implementation perspectives, this technological system is very promising in terms of improving the effectiveness of legal research, enabling lawyers to reach the gist of lengthy cases promptly, as well as in assisting the courts in expediting the processes of reviewing documents. The BART-based model automates and eliminates laborious and time-consuming tasks of summarizing legal documents, therefore, making it possible to harmonize information retrieval with the ongoing trends in the legal profession, which emphasizes the need for digital transformation. This solution goes beyond the challenge of providing access to legal information to invest in the future by developing things like

supportive multilingual capabilities, enhancing the domain specific keyword mining, and possibly even having an interface with legal text databases for retrieval and summarization.

This project signifies progress in the area of processing legal documents. It presents a system which advocates for the necessity of legal texts and the application of their simple versions. This summarization model illustrates the potential of NLP and machine learning in enhancing the management of legal information as it focuses on the main story and captures all the crucial details of every case.



Dataset

IL-TUR is abbreviated as Indian Legal Text Understanding and Reasoning, and is designed as a way to establish a benchmark for the performance evaluation and also for the training of the AI/NLP models in the Indian legal environment. IL-TUR encompasses eight core tasks each of which brings into play different distinct thoughts, context, and reasoning due to intricacies in legal language and legal reasoning in the jurisdiction of Indian Laws. Tasks that include the use of Indic languages alongside English, make the dataset wider and more applicable in terms of linguistic demographics in the country.

Some important features of the IL-TUR dataset include:

- Single Structure: The dataset repository presents a consistent data structure, thereby facilitating effective management of diverse tasks and different data portions. The users can either browse the dataset using a viewer or use Python codes to download and make use of the dataset, making it convenient to work with and analyze the dataset as needed.

- Data Characteristics: The data set contains a large number of legal documents containing cases enriched for performing tasks such as legal summary, legal classification, legal case prediction, etc. These documents are provided with fields such as id, document, and summary where each document corresponds to a particular real-world case with its illustrations, judgment, and other details which are crucial to that case such as the citations of the case. Annotations contain num_doc_tokens (number of tokens in the whole document) and num_summ_tokens (number of tokens in the summary), which shed light on the size of the document and its token ratio.

To conclude, IL-TUR dataset come as a response to the rising demand of AI that is trained on Indian law with a comprehensive and flexible material for going on with the construction of legal NLP models. For instance, researchers looking to use IL-TUR are expected to include citations for the dataset and appropriate task specific studies that can enhance cooperation and creativity in the area of legal oriented AI within the country's India.

id string · lengths	document list · lengths	summary list · lengths	num_doc_tokens int64	num_summ_tokens int64
1 4	4 4.04k	1 995	94 140k	0 27.8k
6712	["IVIL Appeal No.768 (NT) of 1977 etc.", "From the judgment Order dated 9.10.1975 of the Madhya Pradesh High Court in M.C.C…	["The appellant, a manufacturer of cement, entered into an agreement with the Cement Manufacturing Company of India…	1,385	509
2858	["Appeal No. 36 of 1967.", "Appeal by special leave from the judgment and order dated August 25, 1966 of the Punjab High…	["The appellant executed a usufructuary mortgage of his house and continued to reside in it as a tenant under a lease obtaine…	4,867	708
6466	["ivil Appeal No. 2072 of 1990.", "From the Judgment and Order dated 13.5.", "1986 of the Calcutta High Court in Suit No. 2479…	["The appellant had filed a suit in the High Court of Calcutta for a declaration that the properties set out in the schedule…	6,074	1,134
6818	["ivil Appeal No. 3504 of 1991.", "the Judgment and Order dated 5.10.1990 of the Bombay High Court in W.P. No. 210 of 1990.",…	["The appellant was injured in a road accident on 22.1.1989, and a claim petition was filed belatedly on 15.3.1990 with a…	2,683	850
5824	["Civil Appeal No. 3189/1989.", "From the Judgment and Order dated 11.3.1987 of the Delhi High Court in C.W.P. No. 875 of…	["Some persons were plying their business by squatting on pavement in front of a hospital in Delhi and had put up stalls…	4,182	1,173

Methodology

The methodology for developing a summarization model for Indian legal documents revolves around a systematic, layered approach, from data collection to model evaluation. This step-by-step explanation of the methodology provides context for each phase, ensuring that each process aligns with the overall goal of creating a robust, contextualized legal document summarization tool.

1. **Data Collection**

   To start off, we sourced information from the Hugging Face Hub specifically the Indian Supreme Court Judgement dataset that is part of IL-TUR dataset collection. This dataset was selected due to its applicability on the Indian judicial systems and the fact that it comes already split into training and testing sets. These were transformed into Pandas DataFrames to simplify handling and processing. The access this pre-structured dataset provides to real-world folders containing legal documents in their complete forms, proved to be an immense asset for training a model designed to work with complex legal text.

2. Data Preprocessing

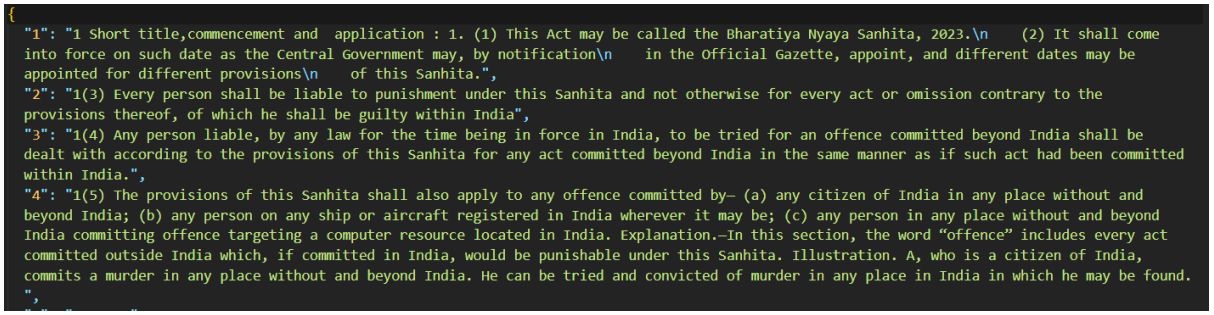
The data preprocessing stage is essential for refining and transforming raw data into structured formats suitable for the model. Each step in preprocessing plays a role in ensuring the model understands and can learn effectively from the legal documents.

- String Merging: The dataset stores both data and summaries as arrays of strings. These arrays are concatenated into single strings for simplified processing, allowing for easier operations across document structures.

- Extracting Sections and Articles: Since the Indian Penal Code (IPC) and Bhartiya Nayay Sanhita (BNS) laws involve numerous sections and articles, our preprocessing includes identifying and extracting these entities from the documents. This is vital as many case summaries refer to specific sections that require tracking within summaries and mappings.

- Mapping of IPC and BNS: At the core of this project is the mapping between IPC and BNS laws, including sections, acts, punishments, and articles. This mapping allows us to track changes across laws, capture removed or added provisions, and ensure continuity and relevance in generated summaries. For this mapping:

    - ➢ **OCR (Optical Character Recognition)** was used to capture the exact text from scanned legal documents.
    - ➢ **Nesting Object Identification** was applied to map the hierarchical relationships within legal sections, building structured data for easy retrieval.
    - ➢ **Structured JSON Storage** was implemented to store the mapping, allowing for flexibility, quick updates, and reliable integration during model training and summarization tasks.



Mapping IPC to BNS

- Text Cleaning with Stopword Removal: To remove noise from the text, common words with little informational value, known as stopwords, were filtered out. Using the Natural Language Toolkit (NLTK), this process

significantly improved the quality of the text input, enhancing the focus on meaningful terms.

- Lemmatization: Lemmatization reduced words to their root forms, ensuring consistency by standardizing variations of a word. This step allowed the model to interpret different forms of a word with the same meaning, enhancing comprehension of legal terminology.

- Tokenization: Tokenization split the text into individual words or tokens, preparing the data for analysis by breaking down sentences into manageable units. This essential step was also performed with NLTK, making future processing and structuring more efficient.

3. Model Development:

In this project, the BART (Bidirectional and Auto-Regressive Transformers) model is used, which is a powerful sequence-to-sequence model well-suited for text summarization tasks. Here's a breakdown of the model development process:

- Data Tokenization and Collation: Preprocessed texts and summaries were tokenized using the BART tokenizer. We utilized a special collator to prepare each text into the model's required format by setting up input_ids, attention_masks, and labels. Given BART's 1024-token limit, we ensured each entry complied with this constraint, helping manage large legal texts.

- Dataset Preparation: The processed data was further structured into Hugging Face Datasets for compatibility with the model's training pipeline. Each training entry was tokenized and stored in dictionaries, then transformed into list-based formats (summ_pt for training and summ_test_pt for evaluation) before converting them into Hugging Face-compatible datasets. This design supported smooth batch integration and processing.

- Data Collation for Sequence-to-Sequence Tasks: We used the DataCollatorForSeq2Seq for efficient batch processing, ensuring dynamic padding across varying text lengths. By aligning batch inputs to the same size per training step, this collator simplified the preparation for sequence-to-sequence tasks, effectively managing data padding and masking.

- Model Training with Hugging Face Trainer API: Training was conducted using the Trainer API from Hugging Face, which unifies model, datasets, training configurations, and evaluation steps into a streamlined training loop. It automatically handled logging, evaluation, and checkpointing, ensuring efficient model updates while mitigating risks of data loss or interruption. This robust pipeline allowed for continual model improvements and checkpoint-based recoverability.

4. **Fine-Tuning:**

Hyperparameter tuning is perhaps one of the most important steps involved in enhancing the performance of the model. It fine-tunes parameters that affect the training process. For this BART-based summarization model, great effort has been undertaken to fine-tune hyperparameters efficiently in such a manner that both efficiency and model performance and learning stability can be achieved during the training process.

Number of Training Epochs (num_train_epochs): it is set to a very minimal value of 1 during initializations to represent the behavior of early performance. This can then be extended in later runs even with the most extreme amount of learning. The number of one epoch pass is usually taken as a check for the model's readiness to be validated and to determine if more training is needed.

Warmup steps- warmup_steps-Setup at 500, it sets the learning rate to increase linearly from 0 on the first steps. Warmup is helpful to stabilize training as sharp weight updates are avoided in the start. This smooths out the problem of convergence that the model might face.

Batch Size: The train and evaluation batch sizes are set to 2, balancing the usage of memory with reasonable training speed. For high memory requirements, smaller batch sizes are often needed in big models like BART.

Weight Decay (weight_decay): It is 0.01. This is a form of regularization that prevents overfitting by penalizing large model weights. Thus, the model generalizes better to new data.

Logging Steps (logging_steps): It is set at 10 to ensure that the loss and gradients are logged in plenty of frequency, providing a clue about the training process and the ability to monitor at real-time for anomalies.

Evaluation Strategy: It is set to steps with eval_steps at 500, meaning that during the process of training, it will be evaluated continuously to provide an opportunity to monitor the performance and check against overfitting without having to wait for the entire epoch.

Save Steps (save_steps): Saved at 1e6 It is a frequency set for saving model checkpoints during training. For short training runs, high save intervals keep storage usage low, while regular evaluations still give some performance data.

Gradient accumulation steps (gradient_accumulation_steps): Set to 16, which simulates a larger batch size by accumulating gradients over multiple steps before updating the model weights. It's very helpful in cases where it's not possible to implement a larger batch size due to the limits of hardware, but still highly efficient in learning from smaller per-device batch sizes.

Together, these hyperparameters will govern stable, efficient, and scalable model training to achieve optimized performance for legal document summarization with BART. All the parameters used were based on memory constraints, model stability, and convergence speeds. Further fine-tuning would be possible in order to improve further performance after initial results are analyzed.

5. **Model Evaluation**

The model's effectiveness was measured through both **quantitative and qualitative evaluations.**

1. **Quantitative Evaluation with Rough Metrics**: ROUGE scores were the primary metric used to gauge the model's summarization quality. Key results were:
    - **ROUGE-1**: 0.76 (indicating strong unigram recall)
    - **ROUGE-2**: 0.69 (reflecting bigram effectiveness)
    - **ROUGE-L and ROUGE-Lsum**: 0.38 (suggesting reasonable context retention)

These metrics indicate that the model effectively captures essential information and context, although further enhancements could improve contextual retention.

6. **Qualitative Analysis**
   The qualitative analysis of the study suggests that the BART model has a great capability in effectively summarizing Indian legal documents when used. The summaries produced by the model are sensible and contextually appropriate as was anticipated at the beginning for such clarity and simplicity. As such, legal practitioners have pointed out a more contextual approach especially in complex litigations which has not been addressed by the model. This assessment puts into perspective the importance of further enhancements and iterations, which will assist the model in catering to different types of legal professionals and users who require accurate and detailed summaries.

Result and Discussion

The project was designed with the aim of creating an summarization model for Indian legal documents. The BART model was chosen for this task due to its proven effectiveness in text summarization tasks. Datasets contain Indian court cases along with their summaries. The training process involved optimizing the model parameters to minimize the loss function, which measures the difference between the model's predictions and the actual summaries.

Conclusion

This project is developed a model that summarizes Indian legal documents using Bart (bidirectional auto-regressive transformers). The model was trained on a dataset that contained various cases from the Supreme Court and High Court and there corresponding summaries.

We aim to improve all ROUGE scores, with a particular focus on enhancing the ROUGE-L score. Building on my current use of the BART model, and plan to incorporate advanced architectures like Longformer, LED, and BigBird to better handle long contexts.

But results are encouraging; there is still potential for improvement. Further work could look to generate a summary in **different languages** that make it easily accessible to a wide range of people.

We are planning to fine-tune our model using **QLora** (quantized low rank adaptation) for better legal text summarization with BART because it enables efficient fine-tuning on limited hardware, reduces resource costs, and allows for quick adaptation to specific legal contexts. This approach maintains high performance while providing relevant and accurate summaries of complex legal documents.

We are also planning use the lightRAG for the retrieval augmented generation from the legal documents, as lightRAG is one of most optimal RAG, which uses the dual-retrival methodology which eventually captures both low level and high level context.

This project establishes a strong foundation for further research in summarizing legal documents, which could greatly enhance the accessibility of legal processes.