Text Summarization in Legislation

Pratham.G.Rao, Prahladh.N.C

Dr. Sowmya Lakshmi B.S

Department of Artificial Intelligence and Machine Learning, BMS College of Engineering, Bengaluru 560050, Karnataka, India

ABSTRACT

Automated text summarization using Natural Language Processing (NLP) can significantly improve the accessibility of legislative documents. This approach includes extractive methods, which select key sentences directly from the text, and abstractive methods, which generate new sentences to convey the document's meaning. Extractive methods are straightforward and maintain the original wording, while abstractive methods use advanced models like BERT and GPT-3 for more coherent summaries. Evaluated using metrics such as ROUGE and BLEU, these techniques can enhance the efficiency of navigating complex legal texts, aiding legal professionals and the public alike. Future research aims to refine these models for better accuracy and adaptation to legal language.

1. INTRODUCTION

The burgeoning volume and intricate nature of legislative documents pose considerable challenges for legal professionals, policymakers, and the public. Traditional manual review of these extensive texts is often impractical due to the sheer time and effort required. Consequently, there is a growing need for automated solutions to distill essential information efficiently. Natural Language Processing (NLP), a branch of artificial intelligence focused on the interaction between computers and human language, offers a promising approach to this problem. By leveraging NLP techniques, automated text summarization can significantly streamline the process of understanding and navigating complex legislative documents.

Text summarization in the context of legislation can be broadly categorized into extractive and abstractive methods. Extractive summarization identifies and extracts the most important sentences or phrases from the original document, thus creating a summary that retains the exact wording of the source text. Techniques such as frequency-based methods, TextRank, and machine learning models like Support Vector Machines (SVMs) and neural networks are commonly employed in this approach. While extractive summarization is relatively straightforward and ensures high accuracy by preserving the original language, it may not always capture the deeper context or nuanced meaning of the document.

2. RELATED WORKS

Text summarization of legal documents has seen significant advancements, focusing on both extractive and abstractive techniques. Extractive summarization methods are popular for their straightforwardness and reliability. Frequency-based methods, where sentences containing the most frequent terms are selected, ensure key topics are covered. TextRank, inspired by PageRank, builds a graph of sentences and ranksSpanish was developed first when seen its development for them by importance. Machine learning models like SVMs and neural networks are also used; Hirao et al. (2013) utilized structured models to capture discourse structures in legislative texts, while Litvak and Last (2008) proposed a graph-based, centroid summarization approach.

Abstractive summarization, while more complex, offers coherent and comprehensive summaries. Transformer-based models such as BERT, GPT-3, and T5 have shown great promise. Liu and Lapata (2019) fine-tuned BERT for summarization tasks, improving summary quality significantly. Zhang et al. (2020) demonstrated that GPT-3 could generate contextually relevant summaries, and Raffel et al. (2020) used T5 for state-of-the-art summarization results, including legal texts.

Several studies have focused on domain-specific adaptations due to the unique nature of legal language. Zhong et al. (2020) introduced a hierarchical attention network tailored for legal document summarization, improving performance on legal datasets. Galgani et al. (2012) combined machine learning and heuristic methods for summarizing legal cases, integrating domain knowledge with data-driven techniques for accurate and informative summaries.

3.PROPOSED ARCHITECTURE

The proposed architecture is designed to automate the process of summarizing and simplifying legal articles and bills. It leverages web scraping, text processing, and both extractive and abstractive summarization techniques to provide concise and understandable summaries of legal texts. Additionally, it incorporates translation capabilities to support multiple languages.

Components

1. **Web Scraping Module**:
   - **Purpose**: To collect data from various sources, such as PDFs and websites, containing legal articles and bills.
   - **Technologies**: BeautifulSoup, Requests
   - **Functionality**:
     - Extracts raw text data from the specified URLs or PDF files.

        - Parses the HTML content to retrieve relevant sections of the articles or bills.

2. **Data Processing Module**:
    - **Purpose**: To clean and preprocess the extracted text data.
    - **Technologies**: NLTK, SpaCy
    - **Functionality**:
        - Tokenizes and normalizes the text.
        - Removes any unwanted characters, stop words, and performs stemming or lemmatization if necessary.
        - Segments the text into meaningful units (sentences, paragraphs).
3. **Summarization Module**:
    - **Purpose**: To generate summaries of the preprocessed text using both extractive and abstractive techniques.
    - **Technologies**: Transformers (Hugging Face), Pegasus, BERT
    - **Sub-components**:
        - **Extractive Summarizer**:
            - Identifies the most important sentences or paragraphs in the text.
            - Uses techniques like TextRank or BERT-based models.
        - **Abstractive Summarizer**:
            - Generates new sentences that capture the essence of the original text.
            - Uses Pegasus or other state-of-the-art models for generating coherent and concise summaries.
4. **Simplification Module**:
    - **Purpose**: To simplify the generated summaries for better readability.
    - **Technologies**: Simplification models (e.g., T5, mBART), Textstat
    - **Functionality**:
        - Uses sequence-to-sequence models trained on simplified text corpora.
        - Evaluates and adjusts the readability score using metrics like Flesch Reading Ease.
5. **Translation Module**:
    - **Purpose**: To translate the simplified summaries into multiple languages.
    - **Technologies**: Hugging Face transformers, MarianMT
    - **Functionality**:
        - Translates text from English to target languages like Hindi, Kannada, and Malayalam.
        - Ensures the translation maintains the meaning and readability of the original summary.
6. **User Interface**:
    - **Purpose**: To provide an easy-to-use interface for users to interact with the system.
    - **Technologies**: Streamlit
    - **Functionality**:
        - Allows users to input keywords or article numbers to search for relevant articles or bills.
        - Provides options to choose the level of summarization and target language for translation.
        - Displays the summarized and translated text along with the original text for comparison.
7. **Backend Infrastructure**:
    - **Purpose**: To handle the server-side processing and integration of various modules.
    - **Technologies**: Flask/Django, AWS/GCP
    - **Functionality**:
        - Manages API requests and responses.
        - Ensures scalable and efficient processing of summarization and translation tasks.
        - Handles data storage and retrieval.

4. IMPLEMENTATION

· Extractive Summarization:

- · Frequency-based Methods: Calculate term frequency and select sentences with the highest frequencies.
- TextRank: Construct a graph of sentences and apply the PageRank algorithm to rank and select the most important sentences.
- Machine Learning Models: Train models like SVMs or neural networks on annotated legal documents to identify key sentences for extraction.

· Abstractive Summarization:

- · Transformer Models: Fine-tune pre-trained models like BERT, GPT-3, and T5 on legal text summarization tasks to generate new, concise sentences.

· Domain-Specific Adaptations:

- · Hierarchical Attention Networks: Implement networks designed to capture the hierarchical structure of legal documents.
- Hybrid Methods: Combine machine learning and heuristic approaches to integrate domain knowledge into the summarization process.

· Evaluation:

- · Use ROUGE and BLEU metrics to compare generated summaries against reference summaries for quality assessment.

4.1 Methodolgy

Data Collection

- Scraping Sources: Gather data on constitutional articles and bills from PDFs and various internet sources. They are available at the official website of the ministry of parliamentary affairs.
- Tools Used: Use web scraping techniques with libraries like BeautifulSoup to collect and parse the required data.

Text Processing

- Normalization: Clean and normalize the scraped text to ensure consistency and accuracy.
- Preprocessing: Tokenize, remove stop words, and preprocess the text for better summarization results.

Feature Extraction and Model Training:

- Use NLP techniques to extract key features from the texts, such as term frequency, sentence importance, and named entities.
- Train machine learning and deep learning models, including transformer-based models like BERT, FalconsAI, Pegasus, for both extractive and abstractive summarization.

Summarization Models

- Extractive Summarization: Utilize an extractive summarization model to highlight key sentences directly from the text.
- Abstractive Summarization: Implement an abstractive summarization model (e.g., Pegasus) to generate a concise and coherent summary by interpreting and paraphrasing the original text.

5. RESULT

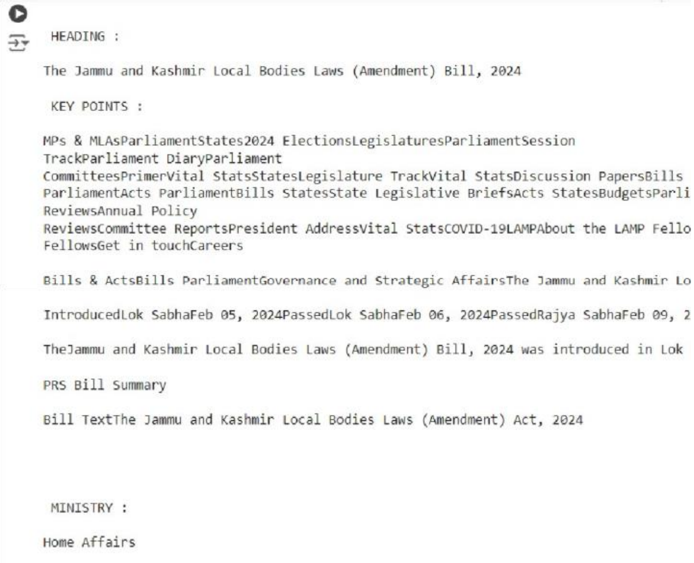
1.**Improved Access to Legal Information**: Enhanced ability for legal professionals, policymakers, and the public to access and understand legislative information through concise and accurate summaries.

2. **Increased Efficiency in Legal Research and Compliance**: Reduced time and effort required for legal research and compliance by providing quick access to essential information from extensive legislative texts.

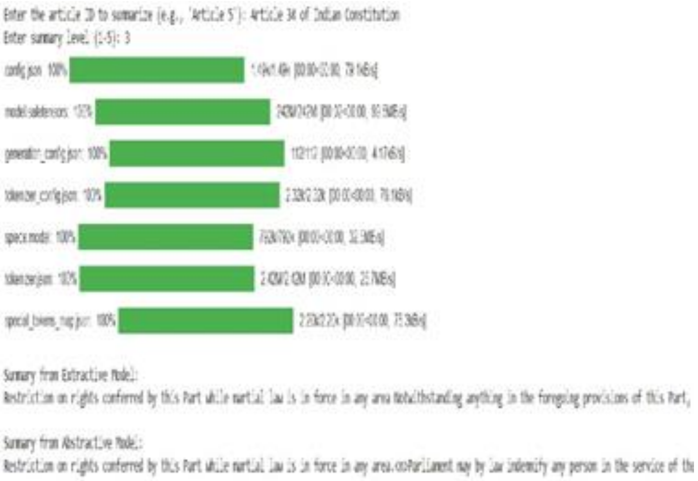
3. **Support for Decision Making and Policy Analysis**: Better-informed decision making and policy analysis by providing clear and accurate summaries of legislative texts.

4. **Enhanced Public Engagement**: Increased public engagement and understanding of legislative processes through accessible and readable summaries.

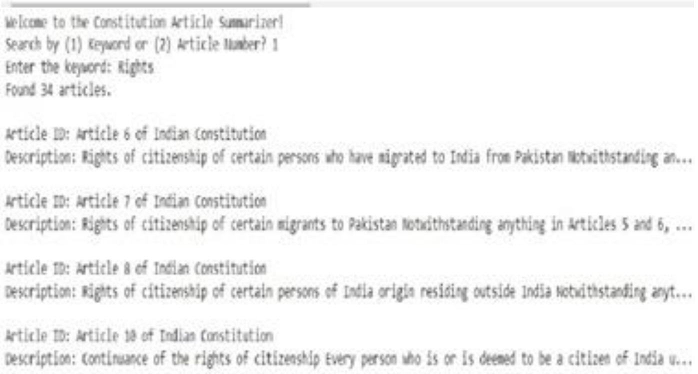
By addressing these challenges and objectives, the development of effective text summarization systems for legislation can significantly improve the accessibility and usability of legal information, benefiting a wide range of stakeholders.



```
 HEADING :

The Jammu and Kashmir Local Bodies Laws (Amendment) Bill, 2024

 KEY POINTS :

MPs & MLAsParliamentStates2024 ElectionsLegislaturesParliamentSession
TrackParliament DiaryParliament
CommitteesPrimerVital StatsStatesLegislature TrackVital StatsDiscussion PapersBills
ParliamentActs ParliamentBills StatesState Legislative BriefsActs StatesBudgetsParli
ReviewsAnnual Policy
ReviewsCommittee ReportsPresident AddressVital StatsCOVID-19LAMPAbout the LAMP Fello
FellowsGet in touchCareers

Bills & ActsBills ParliamentGovernance and Strategic AffairsThe Jammu and Kashmir Lo

IntroducedLok SabhaFeb 05, 2024PassedLok SabhaFeb 06, 2024PassedRajya SabhaFeb 09, 2

TheJammu and Kashmir Local Bodies Laws (Amendment) Bill, 2024 was introduced in Lok

PRS Bill Summary

Bill TextThe Jammu and Kashmir Local Bodies Laws (Amendment) Act, 2024



 MINISTRY :

Home Affairs
```



```
Enter the article ID to summarize (e.g., 'Article 5'): Article 34 of Indian Constitution
Enter summary level (1-5): 3
config.json: 100%                  1.49k/1.49k [00:00<00:00, 78.1kB/s]
model.safetensors: 100%            242M/242M [00:02<00:00, 93.9MB/s]
generation_config.json: 100%       112/112 [00:00<00:00, 4.17kB/s]
tokenizer_config.json: 100%        2.32k/2.32k [00:00<00:00, 76.1kB/s]
spiece.model: 100%                 792k/792k [00:00<00:00, 32.3MB/s]
tokenizer.json: 100%               2.42M/2.42M [00:00<00:00, 23.7MB/s]
special_tokens_map.json: 100%      2.20k/2.20k [00:00<00:00, 75.3kB/s]

Summary from Extractive Model:
Restriction on rights conferred by this Part while martial law is in force in any area Notwithstanding anything in the foregoing provisions of this Part,

Summary from Abstractive Model:
Restriction on rights conferred by this Part while martial law is in force in any area.<n>Parliament may by law indemnify any person in the service of the
```



```
Welcome to the Constitution Article Summarizer!
Search by (1) Keyword or (2) Article Number? 1
Enter the keyword: Rights
Found 34 articles.

Article ID: Article 6 of Indian Constitution
Description: Rights of citizenship of certain persons who have migrated to India from Pakistan Notwithstanding an...

Article ID: Article 7 of Indian Constitution
Description: Rights of citizenship of certain migrants to Pakistan Notwithstanding anything in Articles 5 and 6, ...

Article ID: Article 8 of Indian Constitution
Description: Rights of citizenship of certain persons of India origin residing outside India Notwithstanding anyt...

Article ID: Article 10 of Indian Constitution
Description: Continuance of the rights of citizenship Every person who is or is deemed to be a citizen of India u...
```

6. Conclusion

Text summarization in legislation represents a vital advancement in managing and understanding complex legal documents. As the volume and complexity of legislative texts continue to grow, effective summarization tools become indispensable for legal professionals, policymakers, researchers, and the general public.

Key Insights and Achievements

1**. Addressing Complexity**: By leveraging advanced natural language processing (NLP) and machine learning techniques, summarization models can effectively handle the complex and dense language typical of legal texts. These tools help distill essential information while preserving the context and nuances crucial for legal accuracy.

2. **Enhancing Accessibility:** Summarization tools make legislative information more accessible to a wider audience. Non-experts, including citizens and stakeholders in various sectors, can better understand the implications of laws and regulations through clear, concise summaries.

3. **Improving Efficiency**: Legal professionals and researchers benefit from reduced time and effort in navigating extensive legislative documents. Summarization tools streamline legal research and compliance processes, enhancing productivity and decision-making.

4. **Customization and User-Centric Design:** By offering customizable summaries, these tools cater to diverse user needs. Whether a detailed analysis for legal experts or a high level overview for the general public, the ability to adjust summaries based on user preferences significantly increases their utility.

5.**Scalability and Real-Time Updates**: Modern summarization systems are designed to handle large volumes of data and provide timely updates, ensuring that users always have access to the most current legislative information. This capability is critical in fast paced legal and regulatory environments.

The development and implementation of effective text summarization in legislation are transformative, offering significant benefits across various domains. By making legislative texts more accessible, understandable, and manageable, these tools not only enhance legal practice and policymaking but also foster greater public engagement and transparency in the legislative process. As technology continues to evolve, the potential for even more sophisticated and user-centric summarization tools will expand, further revolutionizing how we interact with and understand legal information.

7. Future Work

The field of text summarization in legislation is rapidly evolving, and several promising areas for future research and development can further enhance the effectiveness, accuracy, and usability of these tools. Here are key directions for future work:

1. **Advanced Natural Language Processing (NLP) Techniques**:

   ○Contextual Understanding: Improving models' ability to understand the broader context of legislative texts. This includes better handling of long-range dependencies and the integration of external knowledge bases to provide more informed summaries.

   ○ Semantics and Pragmatics: Enhancing the models to better grasp the semantics and pragmatics of legal language, which often involves complex and nuanced meanings that are critical for accurate summarization.

2. **Interdisciplinary Approaches:**

○ Legal Expertise Integration: Incorporating insights from legal experts to refine summarization algorithms. This can help in developing models that are more attuned to legal reasoning and the specific needs of different legal domains.

○Cognitive Science and Linguistics: Leveraging findings from cognitive science and linguistics to better understand how humans summarize complex texts, which can inform the development of more natural and effective summarization methods.

3.**Ethical and Responsible AI:**

○Bias Mitigation: Addressing biases in summarization models to ensure fairness and impartiality. This involves developing techniques to detect and mitigate biases that may arise from training data or model architecture.

○ Transparency and Explainability: Enhancing the transparency and explainability of summarization models so that users can understand how summaries are generated and trust their outputs. This includes developing methods to provide interpretable explanations for the decisions made by the models.

4.**Personalization and Customization:**

○User-Eccentric Summaries: Creating systems that can tailor summaries based on user preferences, such as the level of detail, specific topics of interest, and preferred format. This personalization can make summaries more relevant and useful for different user groups.

○Interactive Interfaces: Developing interactive interfaces that allow users to adjust and refine summaries in real-time, providing greater control over the summarization process.

5.**Scalability and Real-Time Processing:**

○Efficient Algorithms: Designing more efficient algorithms that can handle large-scale legislative databases and provide real-time summarization capabilities. This includes optimizing computational resources and improving the speed of model training and inference.

○Continuous Learning: Implementing continuous learning frameworks that enable summarization models to adapt to new legislative texts and evolving legal language dynamically.

7.REFERENCES

[1] Otter, D.W., Medina, J.R., Kalita, J.K. (2020). A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 32(2): 604-624. https://doi.org/10.1109/TNNLS.2020.2979670

[2] Costa-Jussà, M.R. (2018). From feature to paradigm: Deep learning in machine translation. Journal of Artificial Intelligence Research, 61: 947-974. https://doi.org/10.1613/jair.1.11198

[3] Nair, L.R., Peter, S.D. (2012). Machine translation systems for Indian languages. International Journal of Computer Applications, 39(1): 24-31. https://doi.org/10.5120/4785-7014

[4] Chaudhary, J.R., Patel, A.C. (2018). Machine translation using deep learning: A survey. International Journal of Scientific Research in Science, Engineering and Technology, 4(2): 145-150.

[5] Sutskever, I., Vinyals, O., Le, Q.V. (2014). Sequence to sequence learning with neural networks. NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2: 3104-3112.

[6] Basmatkar, P., Holani, H., Kaushal, S. (2019). Survey on Neural Machine Translation for the multilingual translation system. 3rd International Conference on Computing Methodologies and Communication (Erode, India), pp. 443-448. https://doi.org/10.1109/ICCMC.2019.8819788

[7] Unnikrishnan, P., Antony, P.J., Dr Soman, K.P. (2010). A novel approach for English to south Dravidian language statistical machine translation system. International Journal on Computer Science and Engineering, 2(8): 2749-2759.

[8] Khan, N.J., Anwar, W., Durrani, N. (2017). Machine Translation Approaches and Survey for Indian Languages. ArXiv abs/1701.04290.

[9] Revanuru, K., Turlapaty, K., Rao, S. (2017). Neural machine translation of Indian languages. Compute 2017: 10th Annual ACM India Conference (Bhopal, India), pp. 11-20. https://doi.org/10.1145/3140107.3140111

[10] Verma, C., Singh, A., Seal, S., Singh, V., Mathur, I. (2019). Hindi-English neural machine translation using attention model. International Journal of Scientific and Technology Research, 8(11): 2710-2714.

[11] Das, A., Yerra, P., Kumar, K., Sarkar, S. (2016). A study of attention-based neural machine translation model on Indian languages. Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (Osaka, Japan), pp. 163-172.

[12] Shah, P., Bakrola, V. (2019). Neural machine translation system of Indic languages - An attentionbased approach. 2nd International Conference on Advanced Computational and Communication Paradigms (Gangtok, India), pp. 1-5. https://doi.org/10.1109/ICACCP.2019.8882969

[13] Shivakumar, K.M., Nayana, S., Supriya, T. (2015). A study of Kannada to English baseline statistical machine translation system. International Journal of Applied Engineering Research, 10(55): 4161-4166.

[14] Reddy, M.V., Hanumanthappa, M. (2013). Indic language machine translation tool: English to Kannada/Telugu. Proceeding of 100th Science Congress (Kolkata, India), 213: 35-49. https://doi.org/10.1007/978-81-322-1143-3_4

[15] Kodabagi, M.M., Angadi, S.A. (2016). A methodology for machine translation of simple sentences from Kannada to the English language. 2nd International Conference on Contemporary Computing and informatics (Noida, India), pp. 237-241. https://doi.org/10.1109/IC3I.2016.7917967

[16] Saini, S., Sahula, V. (2018). Neural machine translation for English to Hindi. Fourth International Conference on Information Retrieval and Knowledge Management (Malaysia), pp. 1-6. https://doi.org/10.1109/INFRKM.2018.8464781

[17] Verma, A.A., Bhattacharyya, P. (2017). Literature survey: Neural machine translation. CFILT, Indian Institute of Technology Bombay, India

a