IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12721

# A Survey of Legal Document Summarization Methods

**Sheetal Ajaykumar Takale**

Professor, Information Technology Department, VPKBIET, Baramati

0000-0002-6081-2524

**Abstract**: Legal document Summarization is one of the major applications of Artificial Intelligence for Law. This paper presents a survey of various types of approaches. Legal document summarization approaches are mainly categorized as: Extractive vs. Abstractive, Supervised vs. Unsupervised. Recently, Legal domain specific vs. General Domain Large Language Models for legal document summarization are developed. This paper also presents an overview of state-of-the-art technology using LLMs for Legal document summarization. An innovative approach of Knowledge Representation using Ripple-Down-Rules for document summarization is also presented. The paper also presents evaluation of methods.

**Keywords:** Include at least 4 keywords or phrases.

## I. INTRODUCTION

If we consider the number of pending court cases in India, we find that the Artificial Intelligence for Law is the most promising application area. Complex structure of Indian Judgment document and length of the document proves the complexity of Indian Judgment summarization problem. Complexity of automatic judgment summarization is due to the NP-hard sentence selection task in summarization.

A legal practitioner must refer to all the judgments relevant to the case. Legal judgment summary which is also called as "headnote", helps to identify the important or informative portion of judgment and reduces burden of a legal practitioner. A legal judgment is a lengthy and complex document to read it. Legal editors manually prepare summaries for Lawyers and Judges. Manual summarization of the legal judgment is a tedious and backbreaking task. Hence, the need of automatic legal text summarization for the lawyers, Judges and legal experts is evident. Document Summarization approaches proposed for Summarization of legal documents can be categorized based on the summarization method or the algorithm adopted. Summarization method can be: Extractive or Abstractive. Summarization algorithms are Supervised or Unsupervised. In Extractive summarization, detailed extracts, such as the key phrases or sentences, are selected from the source documents, which are further used to build the summary. Whereas, the abstractive summarization generates new sentences that precisely convey the same meaning. Supervised algorithm for legal document summarization requires labelled training dataset for ranking of sentences. It involves human efforts for labelling of sentences in the legal judgement document. Unsupervised algorithm involves sentence clustering and sentence ranking approach. The Artificial Intelligence (AI) based approach using Ripple-Down-Rules (RDR) and Rhetorical Roles is also proposed for legal document summarization.

Abstractive summarization methods based on deep neural networks require large training datasets. The task of collecting such large datasets is prolonged and overpriced task in case of Legal judgement summarization task. Large human efforts are involved for generating Gold standard summaries. To deal with the problem of low-data or data scarcity, pre-trained Large Language Models (LLMs) are used. In addition to general domain pretrained models, Legal domain specific LLMs are also developed. Major advantage is summarization without further training. In this paper we propose a survey of various approaches proposed to summarize the legal document.

## II. SUPERVISED EXTRACTIVE SUMMARIZATION APPROACHES

Rhetorical Role labelling based legal judgement summarization Approach has been an instrumental idea for this problem. Rhetorical role of a sentence represents the semantic function of the sentence for the legal document.

First rhetorical role based classifier for legal text summarization was developed by Hachey [1] which was based on work by Teufel et al. [2]. Teufel and Moens [2] proposed a supervised learning algorithm for summarization of scientific articles. Summary was generated using the extracted sentences with the rhetorical role. They proposed a rhetorical annotation scheme having non-overlapping and non-hierarchical seven labels. Each sentence was assigned to exactly one category. Seven rhetorical roles for sentences used in this work were:

DOI: 10.17148/IJARCCE.2023.12721

*Aim, Textual, Own, Background, Contrast, Basis, Other*. Naive Bayesian supervised learning model was used for learning of rhetorical rules. Training dataset was generated by three task oriented and trained human annotators.

SUM project by Hachey et al. [3] is a system for summarizing the legal judgments of the House of Lords (HOLJ) using rhetorical status classifier. The supervised learning algorithms used in this implementation were: C4.5 decision tree, Naive Bayesian classifier, Winnow Algorithm and Support Vector Machines (SVM). The features extracted for sentences were: *Location, Thematic Words, Sentence Length, Quotation, Named entities and Cue Phrases*.

Bhattacharya et al. [4] have explored the Hierarchical Bi-directional Long Short Term Memory (BiLSTM) and CRF model for identifying rhetorical role as used by Saravanan et al. [5]. Deep Learning models are used for automatic feature extraction. The features generated by the Neural Network as used by the CRF classifier.

Kavila et al. [6] have proposed a hybrid approach for summarization of legal documents which is combination methods from AI. They have proposed thirteen different rhetorical roles for the legal document summarization.

## III. KNOWLEDGE REPRESENTATION FOR EXTRACTIVE SUMMARIZATION

Another approach is to identify the rhetorical roles of sentences in the legal documents using the incremental knowledge acquisition framework built using the ripple down rules. In this approach human intelligence and efforts are utilized to build the rules for knowledge acquisition. RDR is called as an incremental knowledge acquisition framework because; the knowledge base is built with incremental refinements. The RDRs [7] are generated by the subject experts. The refinement or the new rule is recommended by the subject expert for the case which generated an error. This newly added rule in the Knowledge Base corrects the error. RDR is a knowledge acquisition approach proposed by Compton and Jansen. RDR are generated with help of domain expert. The process of knowledge acquisition is incremental, and failure driven. Every failure or knowledge error is patched by adding a new rule by the subject expert. Two types of structures of RDR are: SCRDR [8] and MCRDR [9], [10].

SCRDR: has both true (except) branches and false (if-not) branches. MCRDR: has only true (exception) branches. If at a node, condition evaluates to true, conditions for all children node of that node are tested. The last node on the path which evaluates true provides the conclusion. Hence, for a MCRDR, conclusion is a conjunction of all conditions on the path. Knowledge acquisition using RDR has always been compared with supervised machine learning approach for document classification. The overhead of generating labelled training and testing dataset has always been major disadvantage of supervised approach. For RDR KA, major advantage is error correction ability. Every knowledge error can be patched with the newly added rule.

Legal document summarization using the human expert knowledge in the form of RDR to identify the rhetorical role of each sentence has proved to be an instrumental approach. As compared to Machine Learning models, a very little work has been done for legal intelligence using progressive or improving knowledge acquisition with RDR. Galgani et al. have proposed LEXA, an approach for automatic legal citation classification [11] using knowledge acquisition methodology using RDR. They have designed a knowledge base of 72 RDR rules to recognize distinguished citations. In their work, they concluded with the advantages of knowledge acquisition approach by comparing it with machine learning techniques.

Galgani et al. [12] have proposed an approach for legal case report categorization using RDR. The need of maintaining a large training dataset is avoided in this approach with the help of human generated RDRs to build the incremental knowledge bases.

Galgani et al. [13] have proposed a novel legal document summarization technique using RDR knowledge acquisition to combine different summarization techniques. A rule in the knowledge base is having conjunction of any number of conditions followed by conclusion. Conclusion decides the relevance of sentence for summary. Conditions in the rule are defined using 16 attributes which are defined at term level and sentence level.

For text summarization, Hoffmann and Pham [14] have proposed framework using two data structures: *"level-of-detail tree"*- using tree rule-base and *"discourse structure graph"*- using the graph rule-base. Both rule bases follow Single Classification Ripple Down Rules (SCRDR). Format for rules is, *condition → conclusion*. " *For graph rule, conclusion of the rule is relation between two sentences. For tree rule, conclusion is Boolean value. Rule base tested in this implementation is having 116 rules."*

DOI: 10.17148/IJARCCE.2023.12721

Hartadi and Budi [15] have proposed an approach using RDR for extraction of punishment provision from Indonesian Law text. Knowledge base is build using the *Ridor- the Ripple Down Rule Learner*. *Ridor* is a version of Induct RDR which is available with *WEKA* (Java Data Mining Software). In this work, *Ridor* is used as a classifier Format of rules created by Ridor is:

if condition then conclusion except if.... else if...

The work proposed by Galgani et.al [11] makes use of regular expression to define the condition part of ripple down rule. Rule is expressed as *Pattern → Conclusion*. Whereas, the work proposed by Galgani et al. [16] makes use of attributes defined at sentence level and document level. In this work, for the RDR, condition part is either the regular expression or conjunction of constraints defined using features.

**Table 1: Indian Judgement Document Structure**

| Sr. No. | Details |
|---|---|
| 1 | Beginning of the Judgment : Name of Court, Bench, Judicature, Appeal Number, Appellant, Respondent, name and designation of the Judge concerned, Date of delivery of judgment |
| 2 | Introduction : Involves Preliminary issues, Summary of the Appellant's case, Summary of the defendant's case and Issues to be determined |
| 3 | Evidence and Fact findings : Argument of the appellant, Argument of the defendant, Evidence from either side Judges evaluation of the evidence and the arguments |
| 4 | Ratio Decidendi: The principles of law on which the court reaches its decision. |
| 5 | Conclusion and Final Decision |

Considering the structure of the Indian legal judgment, the problem of legal judgment summarization is addressed in a different way with a different approach by Sheetal Takale et.al [17]. Structure of Indian legal judgment is as stated in table 1. In this work a framework for acquisition of knowledge in the form of RDR rules is developed. This framework builds a set of RDR rules for identifying the Rhetorical roles as listed in Table 2. RDR Rules specify a set of conditions using the features such as: Word Relevance: Key Frequency (KF), Keyword Inverse Sentence Frequency (KISF), Coverage and Diversity: Sentence Length, Ratio of Stop Words, Ratio of Cue-phrases, Ratio of Keywords, Ratio of Proper Noun, Ratio of Capitalized Words, Ratio of Numerical Data Sentence Relevance and Informative: Sentence Sentiment, Ratio of Quoted Text, Ratio of Date, Sentence Position, Sentence Relevance: Sentence Similarity Score.

Format for rules is, *condition → conclusion*: if the given condition is satisfied, the label or rhetorical role is assigned to the corresponding sentence. Condition of a rule is specified using three types of attributes: document level, sentence level and term level attributes.

Major variation of this approach from RDR approach is: the tree structure is not followed for maintaining a record of the rules.

**Table 1: Rhetorical Roles**

| Rhetorical Role | Description |
|---|---|
| Year | Provides year |
| Petitioner | Petitioner name |
| Court, Bench | Court, Bench Name of the court and bench |
| Writ Petition Number | Case number |
| Respondent | Respondent name |
| Advocate for Petitioner | Advocate for the appellant |
| Advocate for Respondents | Advocate for respondent name |
| Coram | Name of the judge |
| Dates | Relevant dates for the case |
| Facts | Facts of the case provided in judgement document |
| Sections and Rules | Sections and rules in judgment document |
| Names | Identify Person Names Involved |
| Judgment | Final decision Final decision given by judge |

  

DOI: 10.17148/IJARCCE.2023.12721

For the legal document summarization problem, in addition to the sentence selection and sequencing, major concern is to extract informative sections. This approach is a combination of extractive and abstractive summarization approaches. The Rhetorical Roles associated with sentences in the judgement become the tool for sentence and content selection from source texts. The knowledge base contains a set of RDRs having format of *Condition → Conclusion*. A portion of text which is satisfying the given Condition is annotated by the Conclusion of rule. Conclusion specifies the annotation with the rhetorical role.

## IV. LARGE LANGUAGE MODEL BASED LEGAL DOCUMENT SUMMARIZATION

Abstractive summarization methods based on deep neural networks require large training datasets. The task of collecting such large datasets is prolonged and overpriced task in case of Legal judgement summarization task. Large human efforts are involved for generating Gold standard summaries. To deal with the problem of low-data or data scarcity, pre-trained Large Language Models (LLMs) [18] are used. In addition to general domain pretrained models, Legal domain specific LLMs are also developed. Major advantage of these pre-trained LLMs is, they can be used for document summarization without further training.

Pretrained LLMs impose a restriction on the size of the maximum input length. Maximum allowed input length varies for the different models. LegPegasus allows length of 1024 tokens, DistilBERT allows 512 tokens, Legal LED allows 16384 tokens, Longformer allows 4096 tokens and BERT allows 512 tokens. Major challenges in case of long document summarization task are: for neural model the computational complexity is increased and large length of the document adds noise to the task of summarization. To deal with the limitations imposed by the maximum input length restriction and the challenges involved in long document summarization, we have to follow the divide-and-conquer approach [19], [20] to summarize the long legal judgements. We split the input legal judgement into chunks of allowed input size.

For abstractive summarization, the Sequence-to-Sequence model with Encoder-Decoder architectures that use RNNs is the most widely recognised framework. Haifeng Wang et.al [18] have a detailed review on Pre-trained Language Models and their application for Natural Language Processing (NLP) task.

Dimitris Mamakas et.al. [21] have proposed use of Pretrained Transformers such as LegalBERT and Longformer for processing Long Legal Documents. However, for Indian legal document abstractive summarization, length of document and availability of training dataset becomes the bottleneck.

Modern Deep Neural Network based abstractive summarization methods include, CNN / Daily Mail based PointerGenerator [22], pre-trained BERT model based abstractive summarization technique: BERTSumAbs [23], Pegasus [24], BART [25], and Longformer [26]. However, the input token limit of the majority of abstractive summarization models is typically lower than the length of legal case documents.

Bajaj et al. [27] built a two-stage extractive abstract technique for summarising lengthy documents. For lengthy document summarization, Gidiotis and Tsoumakas [19] suggested a “Divide- and-Conquer” strategy.

LegalSumm [28] is the method for legal document abstractive summarization. However, Only 200 tokens may be generated using LegalSumm, which is considerably fewer than the desired summaries.

Due to unavailability of large training dataset of Indian legal documents for summarization task, a novel methodology using two text summarization models BART and PEGASUS without the need for a large training dataset was proposed by Satyajit Ghosh et.al. [29]. The abstractive summarization approach for Indian Legal Text Summarization proposed by Satyajit Ghosh et.al. [29] includes collecting Indian legal documents, extracting texts using Optical Character Recognition (OCR), removing noise, normalizing text, constructing dictionaries, and applying models to summarize the documents.

Sequence-to-sequence model for abstractive summarization of Dutch court judgements proposed by Schraagen, Marijn et.al. [30] involves a reinforcement learning and a deep learning algorithms. Different datasets were used for different models, for BART-CNN/Daily mail, RL-Rechtspraak and BARTRechtspraak. The results of the paper showed that BART performed better than the RL model using ROUGE scores and other evaluation metrics.

The use of Document Context Vector and Recurrent Neural Networks in a Seq2Seq-based generative model for abstractive and extractive text summarization is proposed by Chandra

DOI: 10.17148/IJARCCE.2023.12721

## V. CONCLUSION

This paper presents an overview of various approaches proposed for summarization of legal document. Approaches are categorized as supervised approaches, Extractive Approaches, Abstractive Approaches, Artificial Intelligence based Knowledge Representation approaches and Large Language Model based or Pre trained models based approaches. Some of the observations are: Supervised approach is dependent on gold standard corpus with accurately labelled dataset. It is also observed that the Abstractive models always outperform the extractive models, however, suffer due to issues like inconsistencies and hallucinations in the generated summaries. About the newly emerging field of LLMs for Legal domain it can be said that pre-trained abstractive summarization models and LLMs are not yet ready for fully automatic summarization in a complex domain such as Law. However, an AI based approach involving human- intelligence-in-the-loop approach is more suitable where a legal expert can monitor the quality of the summaries generated by these methods.

## REFERENCES

[1] B. Hachey and C. Grover, "A rhetorical status classifier for legal text summarisation," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 35–42. [Online]. Available: https://www.aclweb.org/anthology/W04-1007

[2] S. Teufel and M. Moens, "Summarizing scientific articles: Experiments with relevance and rhetorical status," *Computational Linguistics*, vol. 28, p. 2002, 2002.

[3] B. Hachey and C. Grover, "Extractive summarisation of legal texts," *Artif. Intell. Law*, vol. 14, no. 4, pp. 305–345, 2006.

[4] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, and A. Wyner, "Identification of rhetorical roles of sentences in indian legal judgments," in *Legal Knowledge and Information Systems - JURIX 2019: The Thirtysecond Annual Conference, Madrid, Spain, December 11-13, 2019*, ser. Frontiers in Artificial Intelligence and Applications, M. Araszkiewicz and V. Rodr´ıguez-Doncel, Eds., vol. 322. IOS Press, 2019, pp. 3–12. [Online]. Available: https://doi.org/10.3233/FAIA190301

[5] M. Saravanan, B. Ravindran, and S. Raman, "Automatic identification of rhetorical roles using conditional random fields for legal document summarization," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008. [Online]. Available: https://www.aclweb.org/anthology/I08-1063

[6] S. D. Kavila, V. Puli, G. S. V. Prasada Raju, and R. Bandaru, "An automatic legal document summarization and search using hybrid system," in *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, S. C. Satapathy, S. K. Udgata, and B. N. Biswal, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 229–236.

[7] P. Compton and B. Jansen, "Knowledge in context: A strategy for expert system maintenance," in *Australian Joint Conference on Artificial Intelligence*, 1988.

[8] P. Compton and R. Jansen, "A philosophical basis for knowledge acquisition," *Knowledge Acquisition*, vol. 2, no. 3, p. 241–257, 1990.

[9] D. Richards, "Two decades of ripple down rules research," *The Knowledge Engineering Review*, vol. 24, pp. 159–184, Jun. 2009.

[10] B. H. Kang, W. Gambetta, and P. Compton, "Verification and validation with ripple-down rules," *Int. J. Hum. Comput. Stud.*, vol. 44, no. 2, pp. 257–269, 1996.

[11] F. Galgani, P. Compton, and A. G. Hoffmann, "LEXA: building knowledge bases for automatic legal citation classification," *Expert Syst. Appl.*, vol. 42, no. 17-18, pp. 6391–6407, 2015.

[12] F. Galgani, P. Compton, and A. Hoffmann, "Knowledge acquisition for categorization of legal case reports," in *Knowledge Management and Acquisition for Intelligent Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 118–132.

[13] F. Galgani, P. Compton, and A. G. Hoffmann, "Combining different summarization techniques for legal text," in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 115–123. [Online]. Available: https://aclanthology.org/W12-0515

[14] A. G. Hoffmann and S. B. Pham, "Towards topic-based summarization for interactive document viewing," in *K-CAP*. ACM, 2003, pp. 28–35.

[15] B. Hartadi and I. Budi, "Punishment provision extraction from indonesian law texts with knowledge acquisition rules," in *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, ser. 2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017. United States: Institute of Electrical and Electronics Engineers Inc., 2018, pp. 204–209, 9th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017 ; Conference date: 28-10-2017 Through 29-10-2017.

  

DOI: 10.17148/IJARCCE.2023.12721

[16] F. Galgani, P. Compton, and A. G. Hoffmann, "Combining different summarization techniques for legal text," in *Proceedings of HYBRID12*, 2012, p. 115–123.

[17] S. A. Takale, S. A. Thorat, and R. S. Sajjan, "Legal document summarization using ripple down rules," in *2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2022, pp. 78–83.

[18] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, 2022.

[19] A. Gidiotis and G. Tsoumakas, "A divide-and-conquer approach to the summarization of long documents," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 3029–3040, 2020.

[20] A. Deroy, K. Ghosh, and S. Ghosh, "How ready are pre-trained abstractive models and llms for legal case judgement summarization?" 2023.

[21] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis, "Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer," in *Proceedings of the Natural Legal Language Processing Workshop 2022*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 130–142. [Online]. Available:https://aclanthology.org/2022.nllp-1.11

[22] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, R. Barzilay and M. Kan, Eds., 2017, pp. 1073–1083.

[23] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., pp. 3728–3738.

[24] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," *CoRR*, vol. abs/1912.08777, 2019.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-tosequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[26] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The longdocument transformer," *CoRR*, vol. abs/2004.05150, 2020.

[27] A. Bajaj, P. Dangati, K. Krishna, R. Uppaal, B. Windsor, E. Brenner, D. Dotterrer, R. Das, and A. McCallum, "Long document summarization in a low resource setting using pretrained language models," in *Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, Online, JUli 5-10, 2021*, 2021, pp. 71–80.

[28] D. de Vargas Feijo and V. P. Moreira, "Improving abstractive summarization of legal rulings through textual entailment," *Artificial Intelligence and Law*, pp. 1–23, 2021.

[29] S. Ghosh, M. Dutta, and T. Das, "Indian legal text summarization: A text normalisation-based approach," *CoRR*, vol. abs/2206.06238, 2022.

[30] M. Schraagen, F. Bex, N. Van De Luijtgaarden, and D. Prijs, "Abstractive summarization of dutch court verdicts using sequence-to-sequence models," in *Proceedings of the Natural Legal Language Processing Workshop 2022*, 2022, pp. 76–87.

[31] C. Khatri, G. Singh, and N. Parikh, "Abstractive and extractive text summarization using document context vector and recurrent neural networks," *CoRR*, vol. abs/1807.08000, 2018.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[33] S. Takale, S. Payal, S. Jagtap, P. Jagtap, and A. Khan, "Legal data assistive tool using deep-learning," in *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2023, pp.791–796.

## BIOGRAPHY

**Dr. Sheetal A. Takale** has completed her Ph.D. in Computer Science and Engineering from Walchand College of Engineering, Sangli. She is working as Professor and Head of Information Technology Department. Her research interest areas are Information Retrieval and Artificial Intelligence for Legal Domain.