International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (online): 2347-5552, Volume-11, Issue-4 July 2023
https://doi.org/10.55524/ijircst.2023.11.4.11
Article Id IRP1399, Pages 61-66
www.ijircst.org

Knowledge Representation for Legal Document Summarization

Sheetal Ajaykumar Takale

Professor, Department of Information Technology, Vidya Pratishthan Vidya Pratishthans Kamalnayan Bajaj Institute of Engineering and Technology (VPKBIET), Baramati, Pune, Maharashtra, India

Correspondence should be addressed to Sheetal Ajaykumar Takale; sheetaltakale@gmail.com

ABSTRACT- This paper presents a novel approach for legal document summarization. Proposed approach is based on Ripple-Down Rules (RDR). It is an incremental knowledge acquisition method. RDR allows us to quickly build extendable knowledge base using classification rules. The classification rules are written using a set of features. Summary is generated using the identified rhetorical roles in the document. Experiments demonstrate that the RDR based Legal Document summarization approach outperforms the supervised and unsupervised machine learning models.

KEYWORDS- Ripple-Down-Rules, Rhetorical Roles, Legal Document Summarization.

I. INTRODUCTION

A new challenge for AI & ML experts is to bring Legal Intelligence with the help of digitized legal judgments in Indian Law System. Legal Intelligence is about applying the intelligence in storing or organizing, retrieving, and processing the legal judgment.

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. 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.

Complexity of Automatic text summarization is due to the NP-hard sentence selection task in summarization. Another major issue is how to present the summary to user. AI&ML experts have proposed Various Supervised and Unsupervised Machine Learning models for creating headnote-summary of legal judgments [1].

This paper is an extended version of the work presented in [22]. In this paper, we propose an approach to summarize the legal document by identifying the rhetorical role of a sentence using an incremental knowledge Base build using the RDR.

Supervised approach for learning of rhetorical rules has been proposed by Hachey et.al. [5], Saravanan et.al. [3] and Bhattacharya et.al. [4]. Legal text summarization using Ripple Down Rules has been proposed by Galgani et.al [15].

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 [5] which was based on work by Teufel et al. [7]. Teufel and Moens [7] proposed a supervised learning algorithm for summarization of scientific articles. Summary was generated using the extracted sentences with the rhetorical role.

SUM project by Hachey et al. [8] is a system for summarizing the legal judgments of the House of Lords(HOLJ) using rhetorical status classifier.

Hachey et.al. [5], Saravanan et.al. [3] and Bhattacharya et.al. [4] have proposed an approach for Legal document summarization using rhetorical roles of sentences in a legal case document.

Sarvanan et.al [3] have proposed a rhetorical annotation scheme with seven roles : *identifying the case, facts of the case, arguing the case, history of the case, Analysis or arguments, Ratio of the decision, final decision.*

Bhattachrya and et.al[4] have proposed following seven rhetorical roles of a sentence: *facts, ruling by lowe court, arguments, statute, Prior case documents, Ratio of the decision, Ruling by Present Court.*

The supervised machine learning algorithms used for learning of rhetorical roles are: C4.5 decision tree, Navie Bayes, Winnow algorithm, SVM [5], Conditional Random Fields(CRF)[3] and Neural Model : Hierarchical BiLSTM CRF classifier [4]. Unsupervised approach: DELSumm, proposed by Paheli Bhattachrya et.al [6] is based on Integer Linear Programming (ILP) based optimization.

For supervised learning of rhetorical roles of sentences in legal documents, a high-quality gold standard corpus with accurately identified rhetorical roles of sentences is required. It requires special manual annotation of sentences in legal document by skilled human or law professionals.

In this paper we have proposed an approach to identify the rhetorical roles of sentences in the legal documents using the incremental knowledge acquisition framework built using the ripple down rules. In the proposed 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.

RDR[2] 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 [10] and MCRDR [11], [12]. 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.

Galgani et al. have proposed LEXA, an approach for automatic legal citation classification [13] using knowledge acquisition methodology using RDR. They have designed a knowledge base of 72 RDR rules to recognize distinguished citations. Galgani et al. [14] have proposed an approach for legal case report categorization using RDR. Galgani et al. [15] have proposed a novel legal document summarization technique using RDR knowledge acquisition to combine different summarization techniques.

Kavila et al. [9] 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.

For legal document summarization, we propose to use the human expert knowledge in the form of RDR to identify the rhetorical role of each sentence.

Important contribution in the proposed work is to use ripple down rules to build incremental knowledge acquisition framework to identify the rhetorical roles of sentences. A structured summary is generated for the legal document by using the rhetorical roles of sentences.

II. METHODOLOGY

The proposed legal document summarization approach is carried out in five stages:

1. Pre-processing,
2. Feature Extraction, K
3. Knowledge Acquisition using RDR,
4. Rhetorical role identification and
5. Legal document summary generation.

In the preprocessing stage, the input legal document in PDF file is prepared for processing in further stages. In the second stage, semantic, syntactic and statistical features are extracted at n-gram level, sentence level and document level. In the third stage, human knowledge is represented in the form of RDRs. In the fourth stage, sentences are labelled with thirteen rhetorical roles listed in Table 2. In the fifth stage, summary is generated using the rhetorical role labelled sentences.

Table 1: Set of Features

	Feature	Description of Feature	Contribution
Keyword(Statistical)	KF: Keyword Frequency	Number of times the keyword occurs in the document	Word Relevance
	ISF: Keyword ISF –Inverse Sentence Frequency	Importance of keyword based on its frequency of occurrence in sentences of document.	Coverage and Diversity
Sentence (Statistical)	Sent_length: Sentence Length	Number of words in the sentence	Sentence Relevance
	Ratio_Stop: Ratio of Stop Words	Ratio of number of stop words to number of keywords in sentence.	
	Ratio_Cue: Ratio of Cue-phrases	Ratio of number of cue phrase words to number of keywords in sentence.	
	Ratio_Key: Ratio of Keywords	Ratio of number of keywords to number of words in sentence.	
	Ratio_Noun: Ratio of Proper Noun	Ratio of number of proper nouns to number of keywords in sentence.	
	Ratio_cap: Ratio of Capitalized Words	Ratio of number of capitalized words to number of keywords in sentence.	
	Ratio_Numeric: Ratio of Numerical Data	Ratio of number of Numerical Values in sentence to number of keywords in sentence.	Informative Data
Sentence (Semantic)	Sent_Sentment: Sentence Sentiment	Positive, Negative and Neutral Sentiment for sentence.	Semantic Relevance of Sentence
	Ratio_Quote: Ratio of Quoted Text	Ratio of number of Quoted text to number of keywords in sentence.	Informative Data
	Ratio_Date: Ratio of Date Values	Ratio of number of Dates in each sentence to number of keywords in sentence.	
Document (Statistical)	Sent_Pos: Sentence Position	Importance of sentence based on its position in document.	Sentence Relevance
Document Semantics Based	Sent_Sim: Sentence Similarity Score	Similarity of the sentence with other sentences in document.	

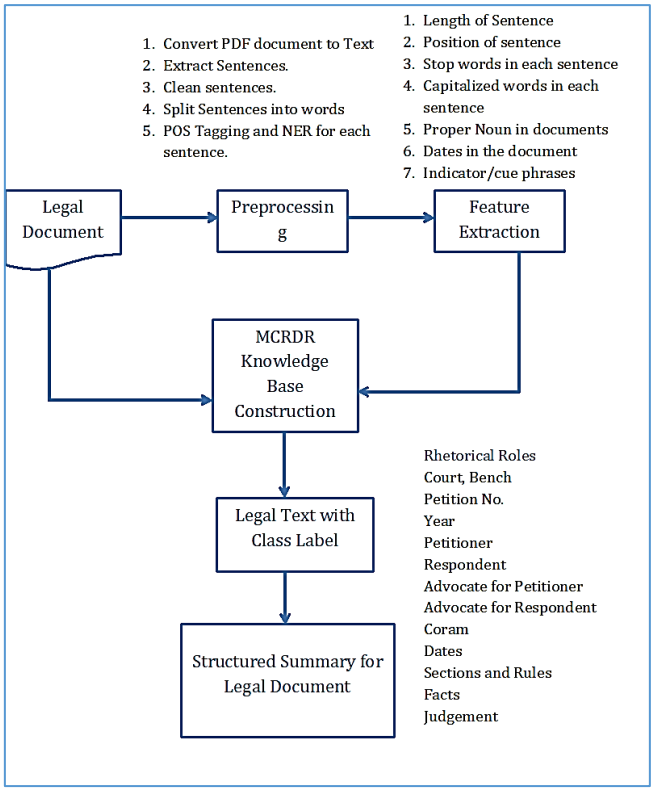

Figure 1: System Architecture

A. *Preprocessing*

The preprocessing involves, PDF to text conversion, tokenization, sentence splitting, Part-Of-Speech (POS) tagging and named entity recognition. Preprocessing of the input document is carried out using, Apache Tika, Spacy, Stanford Natural Language Processing Toolkit: CoreNLP. Tokenization process splits the document into units of different levels of granularity: unigram, bigram, trigram and sentence. Sentence extraction is carried out using Spacy. For n-gram keyword extraction, a Python implementation of Rapid Automatic Keyword Extraction (RAKE) [19] algorithm is used.

B. *Feature Extraction*

The details of different features extracted at word level, sentence level and document level are provided in table 1.

C. *Knowledge Acquisition Using RDR:*

The Ripple Down Rules [2] are created by the domain experts. The knowledge base is built without the knowledge Engineer. It is built from scratch with incremental refinements. The refinement or the new rule is recommended by the domain expert for the case which generated an error. This newly added rule in the Knowledge Base corrects the error. In this implementation, all the rules are made by the author with the help of legal expert.

Considering the structure of the Indian legal judgment, the problem of legal judgment summarization needs a different approach. Structure of Indian legal judgment is as stated in table 3.

For the legal document summarization problem, in addition to the sentence selection and sequencing, major concern is to extract informative sections. The approach proposed here is based on extractive summarization approach. The proposed approach for legal document summarization is based on building the knowledge base to annotate the legal document using different Rhetorical Roles. The Rhetorical Roles become the tool for sentence and content selection from source texts.

Table 2: 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

Table 3: 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

The knowledge base contains a set of ripple down rules 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.

The work proposed by Galgani et.al [13] 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.[19] makes use of attributes defined at sentence level and document level. In the ripple down rules, condition part is specified as conjunction of constraints defined using attributes.

In our proposed approach, ripple down rules are written by considering the keyword, sentence and document level features. For ripple down rules, condition part is either the

regular expression or conjunction of constraints defined using features. User defined rules for annotation using Rhetorical Roles are categorized into three groups. For the first group of rules condition part is stated as conjunction of constraints. Second group of rules use regular expression to specify the condition part of rule. Third group is combination of both the regular expression and conjunction of constraints.

While preprocessing the document in the first step, all the sentences up to the Judgment part are merged in single sentence. In this, the ripples down rules are applied to the first sentence of the document to obtain the specific information such as: Name of Court, Judicature, Bench, Appeal or Petition Number, Year, Name of petitioner, Name respondent, Name of Advocates and Name of Judges. The ripple down rules written for extracting this information make use of regular expression and conjunction of constraints.

Table 4: Example RDR implementation

Rhetorical Role	RDR Rule
Court	`Sent_Pos<=1 and Ratio_Cap>90 and (index=Sent.find("HIGH COURT|SUPREME COURT"))!=-1`
Judicature	`Sent_Pos<=1 and Ratio_Cap>90 and (index=Sent.find("JUDICATURE"))!=-1`
Bench	`Sent_Pos<=1 and Ratio_Cap>90 and (index=Sent.find("BENCH"))!=-1`
Writ Petition or Appeal or Interim Application Number	`Sent_Pos<3 and Ratio_Cap>90 and (index=Sent.find("APPEAL NO"))!=-1 OR index=Sent.find("WRIT PETITION"))!=-1 OR index=Sent.find("INTERIM APPLICATION NO"))!=-1) and Sent_words[index+1].isdigit()`
Year	`Sent_Pos<=1 and Ratio_Cap>90 and (index=Sent.find("OF"))!=-1 and Sent_words[index+1].isdigit()`
Petitioner	`Sent_Pos<4 and Ratio_Cap>90 and (index=Sent.find("PETITIONER"))!=-1 OR (index=Sent.find("APPELLANT"))!=-1`

D. Corpus/Dataset

The data set is collected through Legal Search of Manupatra Legal Search System (https://www.manupatra.com/).

Data set used in this research consists of 100 legal documents belonging to five different domains or subjects such as civil, banking, consumer, education and human rights in Bombay High Court and Supreme Court of India. These judgments in the dataset are 2020 onwards. Manupatra uses human legal experts to annotate court case documents.

Manupatra provides case summary or case note. The Case note is having information about: Petition No., Appellant, Respondent, Coram, Counsels, Subject, Catch Words, Mentioned in, Acts/Rules/Orders, Disposition, and Decision.

The gold standard case summary for each of these cases is obtained from Manupatra.

Table 5: Results

Criminal			
Method	**ROUGE1**	**ROUGE2**	**ROUGEL**
SUMY	0.429	0.283	0.311
LEXRANK	0.213	0.134	0.16
GENSIM	0.572	0.398	0.408
Our System	0.612	0.423	0.501
Education			
Method	**ROUGE1**	**ROUGE2**	**ROUGEL**
SUMY	0.512	0.312	0.34
LEXRANK	0.4	0.315	0.329
GENSIM	0.569	0.412	0.405
Our System	0.623	0.423	0.545
Consumer			
Method	**ROUGE1**	**ROUGE2**	**ROUGEL**
SUMY	0.495	0.392	0.405
LEXRANK	0.251	0.208	0.205
GENSIM	0.691	0.565	0.564
Our System	0.695	0.612	0.634
Commercial			
Method	**ROUGE1**	**ROUGE2**	**ROUGEL**
SUMY	0.36	0.229	0.237
LEXRANK	0.176	0.133	0.146
GENSIM	0.578	0.403	0.386
Our System	0.623	0.578	0.489

III. EXPERIMENTAL RESULTS

Implementation of this project is carried out in Python. Python libraries used in this implementation are:

1. Tika: python For content extraction from PDF file
2. Spacy NLP:NER,POS tagging,dependency parsing, word vectors
3. NLTK : For statistical language processing.
4. KeyBERT[16] : Keyword extraction technique using BERT embeddings
5. LexNLP [18]: Library for working with real, unstructured legal text
6. Rake[17] : Domain Independent Keyword Extraction Algorithm
7. Sklearn- TfidfVectorizer: Conversion of raw documents to a matrix of TF-IDF features

Summary generated by our system has two parts. The first part is the abstractive summary which has the information related to: Appeal Number, Year, Details of Court and Bench, Details of Petitioner and Respondent Coram, Sections and Rules, and Names of Persons.

The second part is the Summary of the Judgement. For the performance evaluation of first part of summary, we

have carried out User Survey. In this survey, human evaluators were presented with original copy of Judgement, gold standard case summary obtained from Manupatra and the First part of Summary. Results obtained for the first part of the summary are observed to be 100% accurate. Evaluation of Judgement Summary is carried out using the ROUGE [20] evaluation approach. We have used ROUGE-1, ROUGE-2 and ROUGEL in this implementation. For the baseline summaries we have used summaries generated by Python implementations of LexRank [21], SUMY and Gensim:Summarize. Figure 2 represents ROUGE scores for baseline methods and proposed system for the four types of Judgement documents.

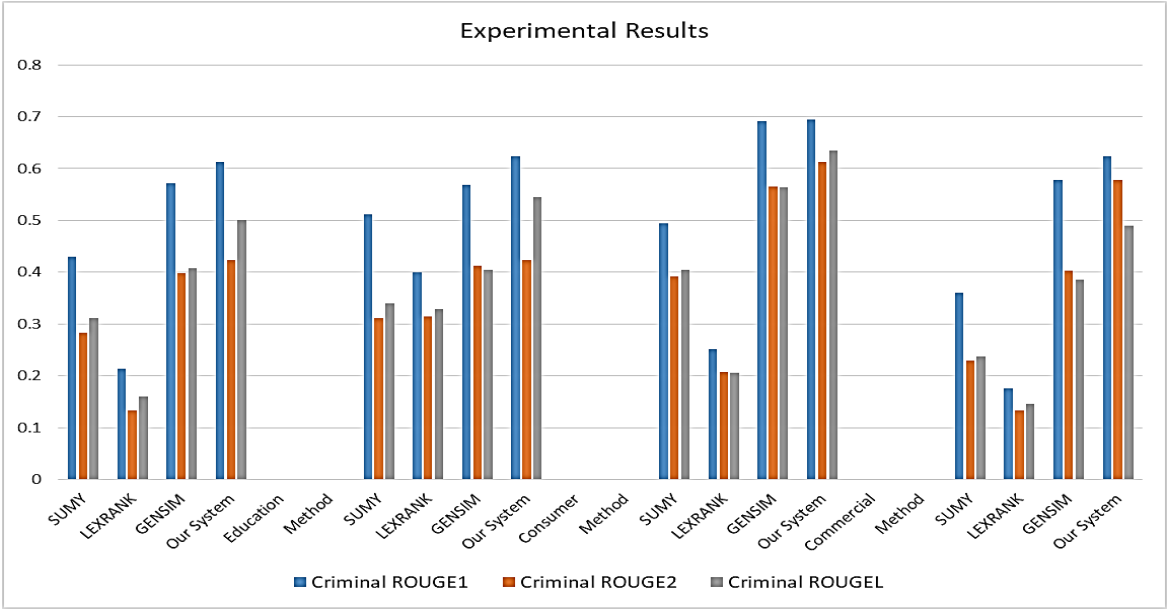

Figure 2: ROUGE Scores for Various Methods

IV. CONCLUSION

This paper presents an approach for summarization of Indian Legal Document using Ripple-Down-Rules and rhetorical roles. RDRs are used as a technique for rapidly building an intelligent system for role labelling of sentences in the document. Knowledge base of classification rules is build using word level, sentence level and document level syntactic, semantic and statistical features. Major contribution in the proposed work is generation of structured summary for Indian Legal judgement using 13 different rhetorical role labels assigned to sentences using 27 RDRs. Experiments carried out for performance evaluation of the proposed approach are using the data set collected through 'Legal Search' of Manupatra Legal Search System. We have used ROUGE-1, ROUGE-2 and ROUGE-L scores for comparing performance of our proposed system with baseline methods: LexRank, SUMMY, and Gensim Summarizer. Experiments demonstrate that our system outperforms as compared to the baseline methods.

REFERENCES

[1] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh, and S. Ghosh, "A comparative study of summarization algorithms applied to legal case judgments," in ECIR, 2019.

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

[3] 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

[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, vol. 322. IOS Press, 2019, pp. 3–12.

[5] 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

[6] P. Bhattacharya, S. Poddar, K. Rudra, K. Ghosh, and S. Ghosh, "Incorporating domain knowledge for extractive summarization of legal case documents," CoRR, vol. abs/2106.15876, 2021. [Online]. Available: https://arxiv.org/abs/2106.15876

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

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

[9] 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.

[10] P. Compton and R. Jansen, “A philosophical basis for knowledge acquisition,” Knowledge Acquisition, vol. 2, no. 3, p. 241–257, 1990.
[11] D. Richards, “Two decades of ripple down rules research,” The Knowledge Engineering Review, vol. 24, pp. 159–184, Jun. 2009.
[12] 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.
[13] 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.
[14] 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.
[15] 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
[16] M. Grootendorst, “Keybert: Minimal keyword extraction with bert.” 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265
[17] S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic keyword extraction from individual documents,” in Text Mining. Applications and Theory, M. W. Berry and J. Kogan, Eds. John Wiley and Sons, Ltd, 2010, pp. 1–20. [Online]. Available: http://dx.doi.org/10.1002/9780470689646.ch1
[18] M. J. B. II, D. M. Katz, and E. M. Detterman, “Lexnlp: Natural language processing and information extraction for legal and regulatory texts,” CoRR, vol. abs/1806.03688, 2018.
[19] F. Galgani, P. Compton, and A. G. Hoffmann, “Combining different summarization techniques for legal text,” in Proceedings of HYBRID12, 2012, p. 115–123.
[20] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013
[21] G. Erkan and D. Radev, “Lexrank: Graph-based lexical centrality as salience in text summarization,” Journal of Artificial Intelligence Research - JAIR, vol. 22, 09 2011.
[22] 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.

ABOUT THE AUTHOR

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.