Summarization of Indian Legal Judgement Documents via Ensembling of Contextual Embedding based MLP Models

Deepali Jain, Malaya Dutta Borah and Anupam Biswas

Department of Computer Science & Engineering, National Institute of Technology Silchar, India

Abstract

Automatic summarization of lengthy legal documents can provide great help to the involved legal practitioners, as well as all the other end users. In this work, an extractive summarization approach has been developed which represents legal document sentences in terms of domain specific pre-trained embeddings and performs subsequent multilayer perceptron based classification to find their summary worthiness. With this approach, we have participated in the summarization related shared tasks of AILA 2021 (Task 2(a) and 2(b)). The results on the test dataset for Task 2 show that our proposed approach is able to outperform most of the other competitors, achieving 2nd position in Task 2(a) and best ROUGE-F1 scores across all the ROUGE metrics for Task 2(b). While the proposed approach has produced impressive results for Task 2, the same approach could not do well for the rhetorical labeling task (Task 1) as per the results provided by the organizers on the test dataset.

Keywords

legal bert, legal judgement documents, extractive summarization, contextual embeddings

1. Introduction

Legal case judgement documents are usually very lengthy and unstructured, making it difficult for legal professionals to read them and understand the key information. Also, reading such lengthy legal documents is time-consuming where the text is around 4500 words on an average [1]. In this direction, if some shorter versions are available for these lengthy documents, it would be beneficial for lawyers, judges, lawmakers, and ordinary citizens. To deal with this, the organizers of the FIRE 2021 Artificial Intelligence for Legal Assistance (AILA) track have introduced the shared task known as Legal Document Summarization (Task 2) [2]. This task is further divided into two subtasks: (a) Identifying the summary-worthy sentences in legal judgements for creating a headnote or a summary. (b) Automatic generation of summaries from legal documents.

In the text summarization literature, it has been identified that there are mainly two types of automatic summarization approaches—Abstractive summarization and Extractive summarization. The techniques corresponding to abstractive summarization involves novel text generation based summary formation which is dependent on the understanding from the input documents.

Whereas, in extractive summarization, the main idea is to extract summary worthy sentences from the input document itself to form a summary. Several research works have attempted to summarize such lengthy legal documents [3, 4, 5, 1, 6, 7]. Some of the works have also made use of rhetorical roles for performing the downstream summarization task [8, 9, 10]. A detailed discussion on the various legal document summarization techniques along with several future research directions can be found in [11]. It is important to note here that both the subtasks of Task 2 directly correspond to the idea of extractive summarization, which is why in this work, we have primarily focused on performing summarization of legal judgement documents via efficient selection of summary-worthy sentences from the input documents. More specifically, we find domain specific vectorized representation of sentences in the input documents, followed by their summary worthiness classification, with a Multilayer Perceptron (MLP) model.

Following the introduction, the organization of the rest of the paper is as follows: Section 2 presents a detailed methodology for the automatic legal document summarization task. Results and analysis are given in Section 3. Finally, Section 4 concludes this paper with a summarization of the findings along with potential future research directions.

2. Data and Methods

2.1. Datasets

The organizers of the AILA track have provided the training dataset for Task 1 and Task 2. For Task 1, they provide 60 documents, where each sentence of a document is labeled with one of the seven rhetorical labels (Facts, Ruling by Lower Court, Argument, Statute, Precedent, Ratio of the decision, Ruling by Present Court) [12, 13]. For Task 2, the organizers have provided 500 document-summary pairs of judgments by the Supreme Court of India [14]. The organizers have provided the pre-processed and sentence tokenized versions for both judgments and summaries, along with summary-worthiness and rhetorical roles labels for each sentence in the documents. A description of the AILA track FIRE 2021 is given in [2] and the overview of the tasks organized in this track is presented in [15].

2.2. Methodology

In order to perform the summarization task, we have randomly split the training dataset into five folds consisting of 400 training samples and 100 validation samples. Now, for each fold, we have performed the steps as shown in Fig. 1(a). All the sentences are firstly converted into contextual embeddings using a pre-trained Legal-Bert model [16]. In this way, a 768-dimensional vector is obtained since Legal-Bert has a fixed hidden size of 768 dimensions. These sentence embeddings are then fed through the MLP model as shown in Fig. 1(b). For this summary-worthiness classification problem, it is to be noted here that the last dense layer consists of single node with sigmoid activation function and binary cross-entropy loss.

2.2.1. Summary-Worthy Sentence Identification Task (Task 2(a))

Specifically for Task 2(a), firstly, contextual embeddings of every sentence in a document are found using the Legal-Bert pre-trained model. Then these embeddings are fed through the MLP
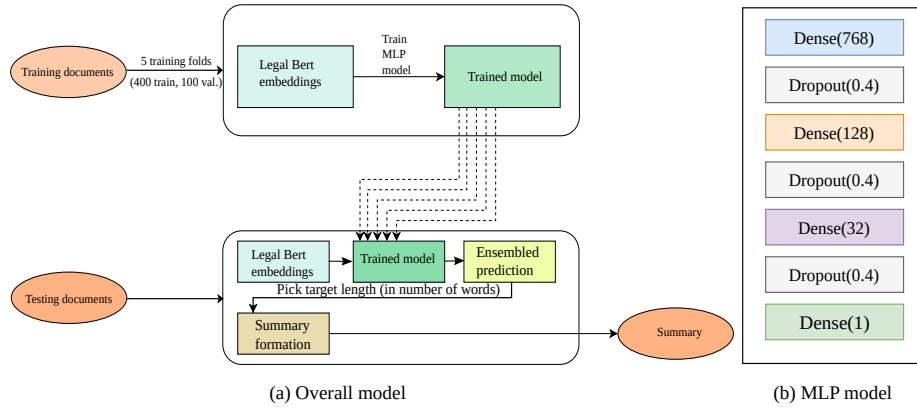
| (a) Overall model | (b) MLP model |

Figure 1: Overall Methodology of the proposed approach

model. Since it is a binary classification task, we have considered the dense layer of 1 node as the last layer with sigmoid activation and binary cross-entropy. This way, training is done for all five models (for five folds). We predict the summary-worthiness probabilities of each sentence of a test document using all the five models and take an average of all five probabilities to get an overall probability measure. We assign label '1' to the sentence if the predicted average probability is greater than or equal to the threshold of 0.4, otherwise the '0' label is assigned. This is the exact approach utilized for Run-1 submission of our team (nits_legal).

For Run-2 submission, instead of using a simple MLP model, we use a multi-task learning based MLP model. In this approach, we have used both rhetorical and summary-worthy labels and fed the training dataset into a multi-task learning MLP model. We have chosen to perform a multi-task learning based approach hoping that learning rhetorical labeling might be helpful for appropriately predicting the summary-specific relevance labels of sentences. The exact MLP model utilized for this is depicted in Fig. 2. The training of such a model again results in 5 different versions for five different folds, which enables averaging based model ensembling. At inference time, we predict in the same manner as that of Run-1 using all the individual multi-task learning models.

For Run-3 submission, we take the average of all the individual summary-worthiness probabilities resulting from each individual trained model of Run-1 and Run-2. If the average probability value is found to be greater than or equal to 0.4, we assign label '1' to that sentence, otherwise a '0' label is assigned.

2.2.2. Legal Document Summarization Task (Task 2(b))

Participating in the summarization task, we have performed a sentence classification based extractive summarization of legal judgement documents. We have made use of those models that have been saved for Task 2(a). We used those models corresponding to each Run during the testing phase and got the probabilities for each sentence of a document. We took average probabilities directly as scores for each sentence and ranked the sentences in decreasing order according to these probability scores. Sentences are then picked up according to desired
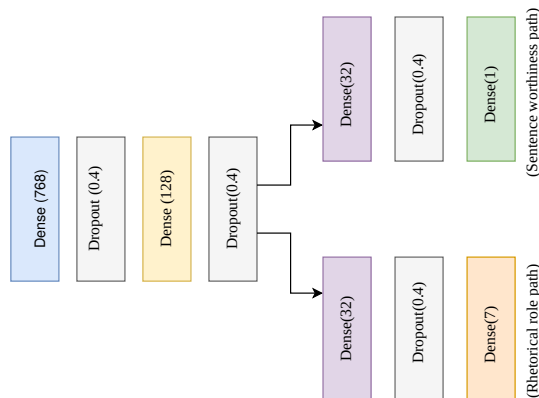
Figure 2: Multitasking based MLP model utilized for Task 2 (Run 2 submission)

summary length (in number of words) as given by the organizers. For Task 2(b), we have submitted three runs of our proposed approaches each of which directly utilizes the individual models saved during the Task 2(a) submissions for Run-1, Run-2 and Run-3.

3. Experimental results and analysis

3.1. Setup

For training an MLP model, we have chosen a batch size of 32 and adam optimizer with a learning rate of 0.001. We run the MLP model for 500 epochs with an early-stopping of 100 epochs. The classification task is evaluated using the standard metrics such as Precision, Recall, and F-score. In contrast, the summarization task (task 2(b)) is evaluated with the help of ROUGE metrics, which is a prevalent metric for evaluating automatically generated summaries. All the experiments have been performed with the help of a Linux based machine with RTX-2070 (8 GB GPU).

3.2. Results

The results on the test dataset for Task 2 has been shown in Tables 1,2,3 and 4. Our team *nits_legal*, has ranked second for the Task 2(a), whereas the Run id 1 for Task 2(b) has achieved the highest scores in terms of ROUGE-F1 scores for all variants of ROUGE as shown in Table 2. In terms of ROUGE-recall, our team has achieved the highest scores for ROUGE-3 and ROUGE-4 metrics as shown in Table 3. Whereas, for ROUGE-precision, our team has achieved the second-highest scores for all variants of ROUGE metrics except ROUGE-1 metric as shown in Table 4. Please note that the best scores for each measure is in bold in the tables.

One of the key observations drawn from the results on the test dataset in the case of Task 2(b), is that there is not much difference between the performances of our proposed approaches across all the ROUGE-recall, precision, and F1-scores. All three metrics are very well-balanced, which demonstrates that our proposed approach is able to produce very precise summaries, while having decent recall values. Usually, when the target summary lengths are not known,

Table 1
Obtained results for Summary-Worthiness Sentence Classification task on test dataset

Team Name	Run id	Precision	Recall	F-Score
Enigma	1	**0.64**	**0.58**	**0.59**
nits_legal (ours)	2	0.61	0.57	0.58
nits_legal (ours)	3	**0.64**	**0.58**	0.58
nits_legal (ours)	1	0.63	0.57	0.57
NeuralMind	1	0.58	0.54	0.54
NeuralMind	2	0.55	0.56	0.52
Chandigard_concordia	3	0.55	0.52	0.51
Chandigard_concordia	2	0.55	0.56	0.5
NeuralMind	3	0.55	0.57	0.49
Chandigard_concordia	1	0.54	0.55	0.46
nit_agartala_nlp_team	1	0.38	0.5	0.43

Table 2
Obtained results of Legal Document Summarization in terms of ROUGE-F1 metric on test dataset

Team Name	Run id	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
Chandigard_concordia	1	0.44233	0.22208	0.13352	0.09743
	2	0.61060	0.31111	0.18960	0.13924
	3	0.62825	0.33752	0.21448	0.16277
nits_legal (ours)	1	**0.64435**	**0.36228**	**0.24354**	**0.19069**
	2	0.63900	0.35339	0.23492	0.18243
	3	0.64059	0.35882	0.24140	0.18956
NeuralMind	1	0.62849	0.33197	0.20646	0.15288
	2	0.56431	0.27825	0.16242	0.11817
	3	0.59228	0.29723	0.17847	0.13057
Enigma	1	0.53006	0.30703	0.20731	0.16243
nit_agartala_nlp_team	1	0.57564	0.31892	0.20370	0.15433

top k words are taken during the summary formation step, where k depends on the average summary lengths in the training set. However, in this sub-task, the desired summary lengths for each test document were already given by the organizers, and in spite of such constraints, our model is able to achieve very well-balanced summarization.

We tried to apply the same MLP based classification approach for Rhetorical role labeling (Task 1) multi-class classification problem also. However, our approach could not perform as efficiently for this task as it did for the summarization specific tasks, as shown in Table 5. For this task, apart from the straightforward MLP based submission, we also submitted a minority class oversampling based run (Run id 2) which did improve the classification performances slightly. Interestingly, the best performing team's prediction accuracy scores for the minority class (Ruling by Lower Court) is 0 across all the metrics, which is much worse than our results for the minority class. However, even with this improvement, the results were not very encouraging.

Table 3
Obtained results of Legal Document Summarization in terms of ROUGE-Recall metric on test dataset

Team Name	Run id	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
Chandigard_concordia	1	0.38219	0.18972	0.11274	0.08133
	2	0.62619	0.31825	0.19346	0.14184
	3	0.63965	0.34298	0.21757	0.16493
nits_legal (ours)	1	0.64141	0.36055	**0.24230**	**0.18967**
	2	0.63590	0.35157	0.23364	0.18137
	3	0.63753	0.35701	0.24009	0.18848
NeuralMind	1	0.62662	0.33080	0.20562	0.15221
	2	0.57792	0.28499	0.16560	0.12024
	3	0.59581	0.29986	0.18039	0.13198
Enigma	1	0.49168	0.28433	0.19134	0.14920
nit_agartala_nlp_team	1	**0.66806**	**0.36985**	0.23436	0.17607

Table 4
Obtained results of Legal Document Summarization in terms of ROUGE-Precision metric on test dataset

Team Name	Run id	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
Chandigard_concordia	1	0.65839	0.33217	0.20082	0.14833
	2	0.59710	0.30484	0.18618	0.13692
	3	0.61860	0.33291	0.21191	0.16100
nits_legal (ours)	1	0.64741	0.36409	0.24483	0.19176
	2	0.64223	0.35528	0.23625	0.18353
	3	0.64376	0.36069	0.24275	0.19069
NeuralMind	1	0.63085	0.33337	0.20745	0.15365
	2	0.57403	0.28205	0.16437	0.11905
	3	0.60066	0.29891	0.17836	0.13025
Enigma	1	**0.68037**	**0.39362**	**0.26491**	**0.20849**
nit_agartala_nlp_team	1	0.54649	0.30286	0.19483	0.14894

Such reduced performance may be attributed to the fact that we have considered the rhetorical labeling task as a sentence level multi-class classification problem and not as a sequential sentence classification problem at the document level. In our proposed approach, although each sentence is represented by its contextual embedding, it still lacks the information on how these sentences contribute towards the overall document. Further explorations can be performed where the rhetorical labeling task is considered both at the sentence as well as at the document levels, in a hierarchical fashion. Please note that the highest overall precision, recall and F1-score is in bold in the case of Table 5.

Table 5
Obtained categorywise and overall results on test dataset for Rhetorical Roles

Run id	Metric	Argument	Facts	Precedent	Ratio of the decision	Ruling by Lower Court	Ruling by Present Court	Statute	Overall
Rustic Run 1	Precision	0.808	0.765	0.296	0.751	0.000	0.595	0.619	**0.548**
	Recall	0.539	0.695	0.746	0.620	0.000	0.846	0.867	0.616
	Fscore	0.646	0.728	0.424	0.679	0.000	0.698	0.722	**0.557**
Rustic Run 2	Precision	0.767	0.749	0.270	0.768	0.000	0.568	0.571	0.528
	Recall	0.590	0.749	0.612	0.620	0.000	0.962	0.800	**0.619**
	Fscore	0.667	0.749	0.374	0.686	0.000	0.714	0.667	0.551
Rustic Run 3	Precision	0.735	0.676	0.316	0.737	0.111	0.523	0.483	0.511
	Recall	0.641	0.724	0.448	0.627	0.133	0.885	0.933	0.627
	Fscore	0.685	0.699	0.370	0.677	0.121	0.657	0.636	0.549
MiniTrue Run 1	Precision	0.737	0.604	0.304	0.692	0.000	0.535	0.522	0.485
	Recall	0.718	0.690	0.254	0.657	0.000	0.885	0.800	0.572
	Fscore	0.727	0.645	0.276	0.674	0.000	0.667	0.632	0.517
Arguably Run 1	Precision	0.644	0.622	0.281	0.708	0.024	0.435	0.542	0.465
	Recall	0.744	0.565	0.582	0.545	0.067	0.769	0.867	0.591
	Fscore	0.691	0.592	0.379	0.616	0.036	0.556	0.667	0.505
MiniTrue Run 3	Precision	0.455	0.611	0.315	0.700	0.000	0.550	0.600	0.461
	Recall	0.769	0.678	0.254	0.645	0.000	0.846	0.800	0.570
	Fscore	0.571	0.643	0.281	0.671	0.000	0.667	0.686	0.503
MiniTrue Run 2	Precision	0.492	0.609	0.309	0.697	0.000	0.539	0.571	0.460
	Recall	0.769	0.678	0.254	0.648	0.000	0.808	0.800	0.565
	Fscore	0.600	0.642	0.279	0.671	0.000	0.646	0.667	0.501
SSN_NLP Run 2	Precision	0.5849	0.6122	0.3148	0.6852	0.06383	0.4222	0.4762	0.45133
	Recall	0.7949	0.5021	0.5075	0.5927	0.2	0.7308	0.6667	0.57067
	Fscore	0.6739	0.5517	0.3886	0.6356	0.09677	0.5352	0.5556	0.49105
Arguably Run 2	Precision	0.558	0.616	0.279	0.720	0.043	0.413	0.522	0.450
	Recall	0.744	0.565	0.612	0.517	0.133	0.731	0.800	0.586
	Fscore	0.637	0.590	0.383	0.602	0.065	0.528	0.632	0.491
SSN_NLP Run 3	Precision	0.5172	0.6243	0.32	0.6642	0.1034	0.2857	0.55	0.43783
	Recall	0.7692	0.4728	0.3582	0.6201	0.2	0.8462	0.7333	0.5714
	Fscore	0.6186	0.5381	0.338	0.6414	0.1364	0.4272	0.6286	0.47547
nits_legal Run 2 (ours)	Precision	0.618	0.579	0.206	0.661	0.038	0.500	0.571	0.453
	Recall	0.539	0.582	0.448	0.545	0.067	0.539	0.533	0.464
	Fscore	0.575	0.580	0.282	0.597	0.049	0.519	0.552	0.45
nits_legal Run 1 (ours)	Precision	0.583	0.570	0.235	0.630	0.065	0.478	0.529	0.441
	Recall	0.359	0.511	0.403	0.611	0.133	0.423	0.600	0.434
	Fscore	0.444	0.539	0.297	0.620	0.087	0.449	0.563	0.428
SSN_NLP Run 1	Precision	0.2153	0.599	0.3922	0.694	0.07407	0.1667	0.7333	0.41065
	Recall	0.7949	0.4812	0.2985	0.4462	0.1333	0.8846	0.7333	0.53886
	Fscore	0.3388	0.5336	0.339	0.5432	0.09524	0.2805	0.7333	0.40909
Legal AI 2021 Run 1	Precision	0.444	0.567	0.200	0.623	0.000	0.647	0.278	0.394
	Recall	0.205	0.603	0.373	0.590	0.000	0.423	0.333	0.361
	Fscore	0.281	0.584	0.260	0.606	0.000	0.512	0.303	0.364
UB_BW Run 3	Precision	0.4545	0.585	0.1943	0.6198	0	0.5	0	0.33623
	Recall	0.3846	0.4895	0.5075	0.5446	0	0.6538	0	0.36857
	Fscore	0.4167	0.533	0.281	0.5798	0	0.5667	0	0.3396
UB_BW Run 2	Precision	0.4571	0.597	0.1823	0.6212	0	0.4857	0	0.33476
	Recall	0.4103	0.5021	0.5224	0.5103	0	0.6538	0	0.37127
	Fscore	0.4324	0.5455	0.2703	0.5603	0	0.5574	0	0.33799
Chandigarh Concordia Run 3	Precision	0.316	0.532	0.276	0.684	0.000	0.167	0.246	0.317
	Recall	0.308	0.628	0.239	0.382	0.000	0.923	0.933	0.488
	Fscore	0.312	0.576	0.256	0.491	0.000	0.282	0.389	0.329
Chandigarh Concordia Run 2	Precision	0.324	0.532	0.268	0.682	0.000	0.161	0.250	0.317
	Recall	0.308	0.623	0.224	0.382	0.000	0.923	0.933	0.485
	Fscore	0.316	0.574	0.244	0.490	0.000	0.274	0.394	0.327
UB_BW Run 1	Precision	0.3878	0.5236	0.191	0.6331	0	0.3721	0	0.30109
	Recall	0.4872	0.4644	0.5075	0.4897	0	0.6154	0	0.36631
	Fscore	0.4318	0.4922	0.2776	0.5523	0	0.4638	0	0.31681
Chandigarh Concordia Run 1	Precision	0.161	0.591	0.207	0.682	0.000	0.244	0.146	0.290
	Recall	0.615	0.490	0.254	0.334	0.000	0.769	0.867	0.476
	Fscore	0.255	0.536	0.228	0.449	0.000	0.370	0.250	0.298
Legal NLP Run 3	Precision	0.108	0.397	0.102	0.554	0.125	0.208	0.077	0.225
	Recall	0.103	0.427	0.209	0.455	0.067	0.192	0.133	0.227
	Fscore	0.105	0.411	0.137	0.500	0.087	0.200	0.098	0.220
CEN NLP Run 2	Precision	0.081	0.414	0.118	0.550	0.000	1.000	0.000	0.309
	Recall	0.077	0.414	0.134	0.590	0.000	0.115	0.000	0.190
	Fscore	0.079	0.414	0.126	0.570	0.000	0.207	0.000	0.199
Legal NLP Run 1	Precision	0.070	0.384	0.096	0.582	0.017	0.163	0.071	0.197
	Recall	0.077	0.360	0.224	0.391	0.067	0.269	0.133	0.217
	Fscore	0.073	0.372	0.135	0.468	0.027	0.203	0.093	0.196
Legal NLP Run 2	Precision	0.068	0.440	0.085	0.549	0.014	0.152	0.074	0.198
	Recall	0.077	0.402	0.224	0.334	0.067	0.269	0.133	0.215
	Fscore	0.072	0.420	0.124	0.415	0.023	0.194	0.095	0.192
CEN NLP Run 1	Precision	0.066	0.384	0.140	0.536	0.000	0.130	0.000	0.179
	Recall	0.103	0.435	0.313	0.391	0.000	0.115	0.000	0.194
	Fscore	0.080	0.408	0.194	0.452	0.000	0.122	0.000	0.179
NIT Agartala Run 1	Precision	0.119	0.369	0.128	0.590	0.000	0.089	0.049	0.192
	Recall	0.180	0.444	0.299	0.224	0.000	0.192	0.200	0.220
	Fscore	0.143	0.403	0.179	0.325	0.000	0.122	0.079	0.179

4. Conclusion

Legal document summarization task becomes very important when there is unstructured and lengthy documents such as Indian case judgements documents. In this paper, we describe our methodology for summarization of legal documents, as part of the AILA shared task in FIRE 2021. We have explored the application of the Legal-Bert model to find effective sentence embeddings and then fed these embeddings as input to an MLP model for the purpose of extractive summarization. This kind of MLP based classification approach is found to be very effective at generating extractive summaries of legal judgement documents with very impressive ROUGE scores. Our proposed approach is able to obtain the best summarization scores among all the participants for most of the ROUGE metrics under consideration in Task 2(b). Moreover, for the sentence summary-worthiness prediction task (Task 2(a)) also, our proposed approach was able to attain the second position among all the participants. We found that even though this approach is very effective at summarization of legal documents, it is not as efficient at the task of rhetorical role labeling (Task 1).

In order to further improve the performances for both of the tasks, hierarchical representation of the documents should be taken into consideration, along with the exploration of some of the recent neural architectures such as Graph Neural Networks (GNN).

References

[1] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh, S. Ghosh, A comparative study of summarization algorithms applied to legal case judgments, in: European Conference on Information Retrieval, Springer, 2019, pp. 413–428.

[2] V. Parikh, U. Bhattacharya, P. Mehta, A. Bandyopadhyay, P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Fire 2021 aila track: Artificial intelligence for legal assistance, in: Proceedings of the 13th Forum for Information Retrieval Evaluation, 2021.

[3] P. Bhattacharya, S. Poddar, K. Rudra, K. Ghosh, S. Ghosh, Incorporating domain knowledge for extractive summarization of legal case documents, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 2021, pp. 22–31.

[4] D. Jain, M. D. Borah, A. Biswas, Fine-tuning textrank for legal document summarization: A bayesian optimization based approach, in: Forum for Information Retrieval Evaluation, 2020, pp. 41–48.

[5] D. Jain, M. D. Borah, A. Biswas, Automatic summarization of legal bills: A comparative analysis of classical extractive approaches, in: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, 2021, pp. 394–400.

[6] D. Anand, R. Wagh, Effective deep learning approaches for summarization of legal texts, Journal of King Saud University-Computer and Information Sciences (2019).

[7] A. Farzindar, G. Lapalme, Legal text summarization by exploration of the thematic structure and argumentative roles, in: Text Summarization Branches Out, 2004, pp. 27–34.

[8] M. Saravanan, B. Ravindran, S. Raman, Improving legal document summarization using graphical models, Frontiers in Artificial Intelligence and Applications 152 (2006) 51.

[9] A. Farzindar, G. Lapalme, Letsum, an automatic legal text summarizing system, Legal knowledge and information systems, JURIX (2004) 11–18.

[10] C. Grover, B. Hachey, I. Hughson, C. Korycinski, Automatic summarisation of legal documents, in: Proceedings of the 9th international conference on Artificial intelligence and law, 2003, pp. 243–251.

[11] D. Jain, M. D. Borah, A. Biswas, Summarization of legal documents: Where are we now and the way forward, Computer Science Review 40 (2021) 100388.

[12] P. Bhattacharya, P. Mehta, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Overview of the fire 2020 aila track: Artificial intelligence for legal assistance, in: FIRE (working notes), 2020.

[13] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, A. Wyner, Identification of rhetorical roles of sentences in indian legal judgments, in: Proc. International Conference on Legal Knowledge and Information Systems (JURIX), 2019.

[14] V. Parikh, V. Mathur, P. Mehta, N. Mittal, P. Majumder, Lawsum: A weakly supervised approach for indian legal document summarization, arXiv preprint arXiv:2110.01188v3 (2021).

[15] V. Parikh, U. Bhattacharya, P. Mehta, A. Bandyopadhyay, P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Overview of the third shared task on artificial intelligence for legal assistance at fire 2021, in: FIRE (Working Notes), 2021.

[16] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).