Text Summarization from Judicial Records using Deep Neural Machines

Ayesha Sarwar[1], Seemab Latif[1], Rabia Irfan[1], Adnan Ul-Hasan[2], and Faisal Shafait[1,2]
[1] School of Electrical Engineering and Computer Sciences (SSECS),
National University of Science and Technology (NUST),
Islamabad, Pakistan
[2] Deep Learning Laboratory, National Center of Artificial Intelligence (NCAI),
Islamabad, Pakistan
{asarwar.mscs18seecs, seemab.latif, rabia.irfan, adnan.ulhassan, faisal.shafait}@seecs.edu.pk

***Abstract*—Courts are generating a large amount of data as legal proceedings. In Pakistan, the ratio of cases that are registered every year and the judgments made is very high mainly due to the time it takes to prepare for a trial. Text Summarization is one of the applications of Natural Language Processing (NLP) that can be used to provide a brief overview of the judgment to both the lawyers and the judges which will help save a lot of their precious time, and hence speedy justice can be provided to the people. Transformer-based models in NLP are a benchmark in solving sequence-to-sequence modeling problems. A downside of these powerful machines is that training a model demands high computation power. We have shown that fine-tuning a pre-trained legal Longformer Encoder-Decoder (LED) transformer model on a downstream task provides better accuracy scores on Australian judgments and our prepared datasets from the Supreme Court of Pakistan (SCP) and Islamabad High Court of Pakistan (IHCP). ROUGE, a commonly used metric in sequence modeling, is used to evaluate the trained model. For the Australian judgments, our approach exhibited a significant improvement for ROUGE-1 and ROUGE-2 scores of 37.97% and 20.04%. For our prepared dataset, our approach produced a ROUGE-1 score of 53.11%, a ROUGE-2 score of 32.12%, and a ROUGE-L score of 34.09%.**

***Index Terms*—Legal Text Summarization, Transfer Learning, Natural Language Processing, Supreme Court of Pakistan**

I. Introduction

Every day, large amounts of unstructured text is generated by legal systems all over the world. In Pakistan alone, lawyers, judges, and case workers process and evaluate millions of cases every year. These files can be very lengthy, with hundreds of pages of dense legal material. A well structured summary of a judgment can provide the same insight and understanding as reading the long judgment. As the volume of legal information continues to grow, appropriate efforts are required in the areas of automated processing and access to relevant forensic information. This will also save considerable time.

Significant research has been focused on many forms of summaries, as well as the methods for creating and evaluating them. In [1], authors divided summarization operations into three categories. Based on the input factors, summarization can be performed on a single document or a collection of documents, and is referred to as single-document summarization or multi-document summarization [2]. Based on the output factors, Text Summarization (TS) utilize two broad categories of approaches i.e, extractive and abstractive. Extractive summarization is a type of summary in which the sentences of summary being extracted are phrases or words taken from the original text, whereas abstractive summaries construct new sentences, sometimes known as paraphrases [2]. Based on the purpose, summarization could be generic, domain-specific, and query-based.

Legal text is different from other types of text. For example, general documents of the news genre have little or no structure. The hierarchy of the structure, on the other hand, is critical in legal texts. The characteristics of legal text pose different challenges for both the approaches of TS. In case of abstractive summarization, the generated summary could use the synonym words, which can have different meanings and implications in the given context. However, if trained on a larger dataset, the model can learn to adopt to the domain-specific vocabulary. In case of extractive summarization, the fluency and flow of the generated summary is a major concern, since it selects the top-ranked sentences from the source document to generate a summary. However, abstractive summarization is better than the extractive summarization in a way that it is an approximate representation of the original document with human-generated language [3]. The main challenge for legal text summarization is the domain knowledge to prepare gold summaries by human operators [4].

Considering the challenges in legal text summarization, our work has two main key contributions i.e. dataset preparation and transfer-learning based transformer models for abstractive summarization. Our research is focused on developing a system to summarize long legal documents with the im-

978-1-6654-7095-7/22/$31.00

plementation of transformer-based models while considering both the legal text summarization challenges and the existing limitations of the system. The scope defined for this research pertains to our prepared dataset from the judgments of Supreme Court of Pakistan (SCP) and Islamabad High Court of Pakistan (IHCP).

Subsequent part of the document is organized in the following sections. Section II serves as a window into the notable work that has been done on text summarization generally and in the legal domain over the period of last decades. Section III discusses our proposed approach of text summarization for legal text in detail. Section IV explains the process of how we prepared the dataset, the experimental setup implemented, and the performance metrics used for evaluation. Section V presents the experiments and the analysis of the results in detail. The last section provides the conclusive remarks and sheds light upon the future direction for the research community.

II. Literature Review

In order to better understand and get clear picture of existing work done in literature, we break down the literature review into approaches for both general text and legal documents.

After the invention of the Transformer models in deep learning, the rate of progress in text summarization accelerated. Most of the summarization models [5] and [6] are based on the architecture of transformers. PEGASUS [7] is one of the more recent efforts on abstractive summarization. Abstractive summarization is harder than the extractive text summarization since it is an approximate representation of the original document with human-generated language [3] and hence requires real-word knowledge, and semantic and contextual analysis. However, abstractive summarization has now became a reality in the era of deep learning.

The work in the domain of legal text summarization has started earlier but the progress is not at a very significant pace. LETSUM, a legal text summarizer system [8], generates a table style summary based on four themes, i.e., introduction, context, juridical analysis, and conclusion. The system was trained on judgments of Canadian federal courts. Human based evaluations are also performed for validation purposes. Kanapala et al. [9] developed a new summarization algorithm based on the gravitational search algorithm (GSA) to binary classify a sentence as whether to include it in the final summary or not on the FIRE-2014 dataset. The solution is optimized using GSA that can be trapped into local optimum.

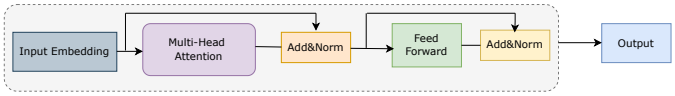
Apart from heuristic based approaches, the models for legal text summarization cover a range of architectures, from simple multilayer networks [10] to complex neural network architectures [6] and [5]. Fuzzy Analytic Hierarchical Process (FAHP) weighting for features is presented in [11] as a novel technique for producing an effective and efficient legal judgment summary. These summaries are subsequently evaluated by experts and are found to be more accurate than summaries produced by traditional approaches. This paper [4] prepared a labelled dataset from the judgments of Supreme Court of India and implemented Feed Forward Neural Network (FFNN) and Long Short Term Memory (LSTM) models to obtain extractive document summary.

Modern NLP is driven by transformer-based models for solving sequence-to-sequence modeling problems, but there seems to be very little amount of research when discussion is set for abstractive summarization in the legal specific domain. The authors in [12] attributed this lack of demonstration on the legal corpora to the difficulty of obtaining large legal datasets due to their confidential nature. In Pakistan, there has also been no work done in the field of legal text summarization. Nasar et al. [13] have only discussed text summarization in general, and then discussed the tools available for legal text summarization. We believe that our work will provide a baseline in the field of legal text abstractive summarization using transformer models.

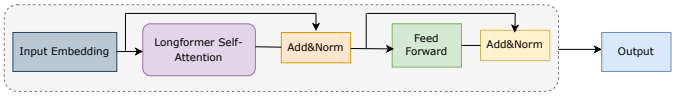
III. Design and Methodology

A. Architectural Analysis

Transformer-based models, such as BERT [14] and BART [5] outperform the LSTMs for neural machine translation tasks with the self-attention mechanism [15]. Although transformers are very powerful neural machines, they are not feasible for long sequences because the memory and computational requirements increase quadratically with the length of the document. To address this issue, a modified transformer design called *Longformer* [6] is proposed whose self-attention grows linearly with length sequence that can be used to process long documents. This property makes it useful for tasks of natural language: classification of long documents, co-reference resolution, question answering, etc. Reformer [16] is an alternate transformer-based architecture that addresses the issue of long input sequences. We do not use this strategy because the authors found that the benefits appear only when inputs get very long, which is outside the range we experiment with. In the Longformer, the maximum input token size is $16,384$. We will be utilizing Longformer Encoder Decoder (LED), a variant of Longformer, for supporting long document generative sequence-to-sequence tasks. A schematic comparison between BART transformer model and LED is presented in Figure 1.



(a) Architecture of the BART transformer model.



(b) Architecture of the LED transformer model.

Fig. 1. Schematic comparison of two transformer-based models i.e., BART and LED architectures, BART transformer model utilizes multi-head attention requiring the quadratic memory while Longformer replaces the attention mechanism in LED with the self-attention which grows linearly with length sequence that can be used to process long documents.

B. Proposed Methodology

Our proposed methodology illustrates how we can handle the Out-of-Vocabulary (OOV) words for the legal text summarization and perform a downstream task for input sequence length.

1) Common Vocabulary: We will utilize the pre-trained model, *legal-led-base-16384* [17], to fine-tune it on our dataset and avoid the OOV words. The sec-litigation-releases dataset [18] that contains around 2700 litigation releases and complaints from year 1995 to 2021, was used to train the *legal-led-base-16384* model. These litigation releases detail the federal court civil lawsuits by the SEC (U.S. Securities and Exchange Commission).

2) Fine-tune for Downstream Task: The input dataset has a median token length of $1,933$ with the $98\% - ile$ token length being $6,101$. The output data has a median token length of 374 with the $90\% - ile$ token length being 385. The legal Longformer Encoder Decoder (legal-led) base model with 16K tokens is fine-tuned on a downstream task for our prepared legal dataset with $8,192$ input tokens and 512 output tokens according to our data statistics. For summarizing, we follow the recommendations of [6] and only apply global attention to the very first token. Figure 2 shows the architecture of our proposed methodology.

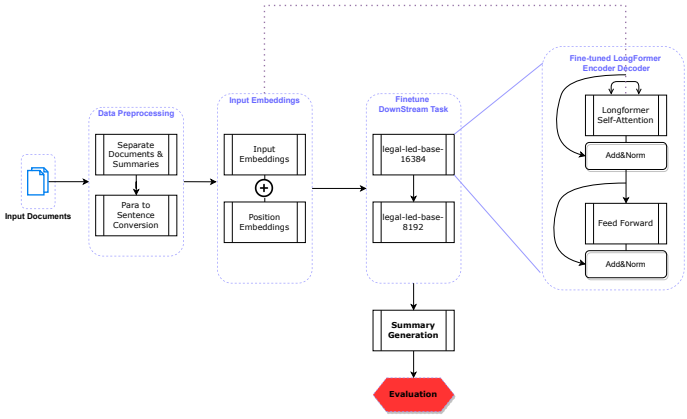


Fig. 2. Architecture of the proposed methodology, the pre-trained legal-led-base-16384 model on sec-litigation-releases dataset is fine-tuned on a downstream task for our dataset to avoid the Out-of-Vocabulary (OOV) words. The model is evaluated on ROUGE, the standard metric used in automatic evaluation of machine translation. The same metric is utilized for evaluation during the training phase to improve the model performance.

IV. Experimental Setup

This section describes the details of the process for dataset preparation, hyper-parameters selection, and the performance evaluation metrics utilized.

A. Dataset Preparation

The preprocessed dataset comprises a total of 429 judgments [1]. All the judgments and their corresponding headnotes are downloaded manually from SCP and IHCP from the year 1991-Present. The process was kept manual to avoid duplicates as one judgment could be included in various journals. The headnotes from the judgments need to be removed to separate the judgment text. Since it is a sequence problem, the input is full single sentence. The University of Malta provides an online service, Maltese Language Software Services [19], to convert the documents into paragraphs and sentences. Afterwards, regular expressions have been used for domain specific abbreviations and sentence segmentation. It also involved manual effort to cross-check for any kind of errors. The distribution of the documents from both courts into training, validation, and testing splits is considered according the 80-20% distribution to avoid over-fitting.

B. Hyper-Parameters Selection

Based on the statistics of the input dataset, we defined an input length of $8,192$ and an output length of 512 to make sure that the model can handle most inputs and can generate enough outputs. The minimum output length is set to 100, and maximum to 512 to make sure that the output length is within the specified range. Tokenizing data samples is carried out up to their respective maximum lengths of 8192 and 512. To prevent out-of-memory errors, we trained on batch size of 2. To save memory, we used beam search with only two beams. Larger beam value results in the improved performance of the model, but at the cost of the speed at the decoding step. Therefore, an appropriate value should be selected accordingly.

A number of other parameters have been set in order to improve the summary generation. According to the GPU RAM specifications, we converted gradient accumulation to a batch size of 8, by setting gradient accumulation steps to 4. Since the batch size is 2 and the gradient accumulation steps are 4, the gradient accumulation batch size becomes 8. Besides the usual attention mask, LED can make use of the global attention mask to define which input tokens are being handled globally and which are being handled locally. In summary, we follow the recommendations of the paper [6] and only apply global attention to the very first token.

We have performed multiple experiments to analyze the behaviour of our model through loss and accuracy graphs after each iteration of optimization. The accuracy is increased and loss is decreased with the number of epochs and it was stable and constant in the last iterations, making 5 to be a suitable and optimal fine-tuned hyper-parameter value for the number of epochs.

C. Performance Evaluation

ROUGE scores are the de facto standard automatic evaluation metric used to measure the accuracy of the sequence length problems [20]. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation.

There are different variants of the rouge scores to evaluate the quality of the generated summaries. In a generated summary of n-grams, a count match (gram n) shows how many times an n-gram appears in the generated summary as well as gold/human generated summaries. N-grams are simply groups

[1] https://drive.google.com/drive/u/1/folders/1zlP7oJ50FH-WmZBDSFoiuXk7uQz5qLHG

of tokens. A uni-gram (or one-gram) consists of one word and is used by ROUGE-1. Two consecutive words make up a bi-gram (2-gram) and is used by ROUGE-2. The ROUGE-L measure measures overlap based on the longest common subsequences (LCS) in the summaries.

V. Results and Analysis

We have performed experiments to fine-tune two models on a downstream task for our SCP and IHCP dataset. One of the models, *led-base-16384*, has been pre-trained on the CNN/DM news dataset and we downstreamed it to *led-base-8192*. The other model, *legal-led-base-16384*, has been pre-trained on the sec-litigation-releases dataset [18] that contains around $2,700$ litigation releases and complaints from year 1995 to 2021, and we downstreamed it to *legal-led-base-8192*.

The comparison of ROUGE scores on our SCP and IHCP dataset for the base pre-trained model (*led-base-8192*) on CNN/DM dataset and *legal-led-base-8192* model (further fine-tuned on the $2,700$ legal sec-litigation-releases) is given in the Table I. We have also performed experiments to compare the results on the AustL-II judgments [21] for *legal-led-base-8192* and the proposed model in [22]. The comparison of the ROUGE scores on the AustL-II judgments for the model *legal-led-base-8192* and the proposed model in [22] are given in the Table II.

TABLE I
Comparison of ROUGE scores on our SCP and IHCP dataset for the base pre-trained model on CNN/DM dataset and the model fine-tuned on the 2700 legal sec-litigation-releases.

Model	led-base-8192	legal-led-base-8192
ROUGE-1	48.95	**53.11**
ROUGE-1-Recall	43.87	**48.25**
ROUGE-1-Precision	64.48	64.22
ROUGE-2	25.27	**32.12**
ROUGE-2-Recall	40.13	**28.96**
ROUGE-2-Precision	40.13	39.67
ROUGE-L	31.22	**34.09**
ROUGE-L-Recall	27.59	**30.79**
ROUGE-L-Precision	43.17	41.91

The comparison in Table I shows that the model performance has been improved with transfer learning if a model is fine-tuned on a dataset with similar vocabulary or domain [23]. In Transfer Learning (TL), the problem occurs with the use of uncommon vocabulary. The technique used in transfer learning to avoid the unnecessary OOV words is to use a common vocabulary between two datasets [23]. Although word2vec and FastText are trained using Wikipedia or other online corpora, the vocabulary that is used in these systems is finite. When training, words that aren't frequently used are often omitted. It is possible, therefore, that legal words specific to law aren't supported in the dictionary.

With pre-trained word embeddings, the OOV words are usually replaced with the UNK (UNKnown) token to represent the unknown words. A corpus that is domain-specific is highly inefficient, as domain-specific words often have significant meaning. Considering that UNK tokens can replace most (meaning-carrying) words consequently, the model will be unable to learn much. Since, the size of our prepared dataset is small, training the model from scratch will cause the issue of over-fitting. Moreover, fine-tuning a model with a dataset from different domain will lead to the issue of OOV words. Therefore, the approach of transfer learning with a model pre-trained on a dataset with a similar domain will overcome this issue. Figures 3(a) and 3(b) show the system generated and human written headnote of the same judgment. We can determine the quality of the summary generated by the *legal-led-base-model-8192* for the SCP dataset from the comparison of Figure 3(a) and Figure 3(b).

The comparison in Table II shows that the accuracy has improved significantly for the judgments from AustL-II. The reason in the difference of the ROUGE-N scores is because of the reason that the methodology proposed in [22] are the sentences from the judgments, whereas the headnotes are generated in an abstractive way by the professional lawyers. The reason in the the difference of the ROUGE-L scores is because of the reason that the longest common sub-sequence with the abstractive summarization is shorter than the extractive summarization. But, the increase in ROUGE-2 shows that the *legal-led-base-8192* model is using the same words in the summary generation task.

TABLE II
Results comparison of our fine-tuned legal-led-base-8192 and proposed model in [22] for AustL-II judgments.

Model	Methodology [22]	legal-led-base-8192
ROUGE-1	27.88	**37.97**
ROUGE-1-Recall	28.16	**28.61**
ROUGE-1-Precision	27.62	**73.75**
ROUGE-2	5.83	**20.04**
ROUGE-2-Recall	5.88	**14.86**
ROUGE-2-Precision	5.77	**41.33**
ROUGE-L	**33.5**	23.49
ROUGE-L-Recall	**33.78**	17.48
ROUGE-L-Precision	33.24	**48.59**

With the added advantages of our proposed approach, it also has some limitations, which we intend to take care of in our future work. Since our model takes the output length of the summary as one of the parameters, and hence even if a sentence is not complete, it is bound to cut off any extra tokens generated outside the provided token length. One way to handle is to generate the output length greater than required and excluding the last sentence to avoid incomplete sentences at the cost of added resources.

ROUGE counts the number of word or words between the gold standard summary and candidate summary. Ideally, a ROUGE score of 100% can only be achieved if both the summaries are identical. Whereas, in case of abstractive summarization, the generated summary could use the synonym words and different sentence structure. This means ROUGE scores are not a true representative of the quality of generated summaries [24]. Table III shows the lowest five ROUGE scores (1,2, and L) for the fine-tuned model, *legal-led-base-8192*, on the SCP and IHCP dataset calculated on individual documents.

GOVERNMENT OF KHYBER PAKHTUNKHWA through Capital City Police Officer Peshawar and others ---Appellants Versus SHAHID ---Respondent

Civil Appeal No. 58 of 2020, decided on 2nd April, 2020

(Against judgment dated 20.11.2017 of Khyber Pakhtunkhwa Service Tribunal, Peshawar, passed in Service Appeal Nos. 734 and 734 of 2014)

Khyber Pakhunkhawa Service Tribunal Ordinance (X of 2014), S.734

Penal of dismissal from service

Trial Court by interfering with penalty imposed by department had exceeded from its jurisdiction more so when the Respondent was employed in a disciplined force where he could not have remained absent from duty

High Court declined to interfere in the impugned order passed by the Service Tribunal

Appellant was dismissed from service on allegation of wilful absence from duty for a period of six months and three days, vide office order

Validity

Punishment of removal from service was converted to withholding of two increments for two years, since, the penalty imposed upon the respondent was harsh

Impugned Order appeared to be harsh one and not commensurate with the lapse/guilt on the part of the appellant and as such the punishment of removal of service of the respondent had been converted to withhold of two-year penalty

Intra-judgment order imposed by the service tribunal was harsh and could not be made basis by the tribunal to modify the penal imposed by it

In such circumstances, the punishment was harsh, as such, punishment of removing from service of appellant was converted from withholding of 2 increments for 2 years

No law in law had been cited by him

Where was the parameters of imposition of major and minor penalties, under what circumstances such penalties were to be imposed and what law governed the imposition of such penalties, the Tribunal had not taken trouble of examining the same or making any observations in the judgment

Just whimsically stating that the punishment is harsh could not make basis by Tribunal to change the penalty imposing by the competent authority to that of withholding two increments

The penalty of removal from the service was reinstated.

(a) System generated headnote of the Civil Appeal Judgment from SCP with the fine-tuned legal-led-base-8192 model.

GOVERNMENT OF KHYBER PAKHTUNKHWA through Capital City Police Officer Peshawar and others ---Appellants Versus SHAHID ---Respondent

Civil Appeal No. 58 of 2020, decided on 2nd April, 2020.

(Against judgment dated 20.11.2017 of Khyber Pakhtunkhwa Service Tribunal, Peshawar, passed in Service Appeal No. 734 of 2014)

Civil service

Police official

Dismissal from service

Wilful absence from duty for a period of six months and three days

Service Tribunal considering penalty of dismissal from service imposed upon respondent to be too harsh a penalty modified the same to withholding of two increments for a period of two years and absence period was treated as leave of kind due

Legality

Tribunal had not taken trouble of examining or making any observations regarding the parameters of imposition of major and minor penalties, and circumstances under which such penalties were to be imposed and what law governed the imposition of such penalties

Whimsically stating that the punishment was harsh could not be made basis by the Tribunal to modify the penalty imposed by the competent authority

Tribunal while modifying the penalty had not acted in accordance with law, in that, no law in such regard whatsoever was cited by him

Tribunal by interfering with the penalty imposed by the department had exceeded from its jurisdiction more so when the respondent was employed in a disciplined force where he could not have remained absent from duty for a long period of 06 months and 03 days

Impugned judgment passed by the Tribunal suffered from illegality and was unsustainable in the eyes of law, therefore the same was set aside, and the penalty of dismissal from service imposed upon the respondent was restored

Appeal was allowed.

Barrister Qasim Wadood, Additional AG, Khyber Pakhtunkhwa for Appellants.

Muhammad Asif, Advocate Supreme Court for Respondent.

(b) Human written reference headnote of the Civil Appeal Judgment from SCP.

Fig. 3. System-generated vs human-written headnote of one of the Civil Appeal Judgment from SCP. The summary produced with our model is following the required format, and also fluent in its language.

Figures 4(a) and 4(b) depict a scenario in which the machine-generated summary is factually correct, but the ROUGE score (2.87 percent) indicated that the summary is mediocre and inaccurate. ROUGE is considered as an intrinsic evaluation, whereas extrinsic evaluation is also as much necessary which involves the human judgment. If the summary is well-written, and covers all the important facts of the source judgment, and required information, the user will be able to answer all the related questions. In this case, a set of related questions need to be prepared. In another scenario, if the legal expert is satisfied with the produced summary, we can consider it as a true one. But involving humans is always an expensive task.

TABLE III
RESULTS OF LOWEST FIVE ROUGE F-SCORES FOR THE FINE-TUNED LEGAL-LED-BASE-8192 MODEL ON THE SCP AND IHCP DATASET CALCULATED FOR INDIVIDUAL DOCUMENTS.

ROUGE-1	ROUGE-2	ROUGE-L
2.36	1.17	2.25
2.77	1.61	2.29
2.87	3.47	2.87
5.63	3.87	4.99
7.57	4.50	6.49

Before Miangul Hassan Aurangzeb, J

RAZI Versus OIL AND GAS DEVELOPMENT COMPANY LTD. and others

Writ Petition No.3964 of 2016, decided on 23rd November, 2016.

Industrial Relations Act (X of 2012)

S. 85 Constitution of Pakistan, Art. 199

Civil service

Change of date of birth

Scope

Employee who joined service as Security Guard after his entry into service was required to declare his correct date of the birth at the time of his entry in service

Validity

Plaintiff could not seek correction of his year of birth, as the rule that a government employee could not make an application for a change in his date of his birth to suit their career and to lengthen their service career

Stale and belated applications for alteration of date could not be entertained

High Court declined to interfere in the judgment passed by the National Industrial Relations Commission, Islamabad

Mushtaq Hussain Bhatti, Advocate, for Petitioner

Date of birth recorded in the records of a government servant was to be treated as final and no amendment would be allowed in it at any stage

Reference was dismissed in circumstances

(a) System generated headnote of the judgment from IHCP with the fine-tuned legal-led-base-8192 model.

Before Miangul Hassan Aurangzeb, J

RAZI Versus OIL AND GAS DEVELOPMENT COMPANY LTD. and others

Writ Petition No.3964 of 2016, decided on 23rd November, 2016.

Industrial Relations Act (X of 2012)

Ss. 33 & 85(1)

Constitution of Pakistan, Art.199

Constitutional petition

Industrial dispute

Date of birth

Correction

Petitioner was "workman" and was aggrieved of decision passed by National Industrial Relations Commission, declining to allow him to amend his date of birth

Validity

Petitioner superannuated on 30-6-2016 and steps taken by him to have his year of birth entered in records of employer changed about a year prior to his retirement had made petitioner's case bereft of bona fides

Stale and belated applications for alteration of date of birth could not be entertained

High Court observed that change of date of birth was a very important responsibility to be discharged since there had been a general tendency amongst employees to lower their age and change their date of birth to suit their career and to lengthen their service career

High Court declined to interfere in the appellate order passed by National Industrial Relations Commission

Petition was dismissed in circumstances.

(b) Human written reference headnote of the judgment from IHCP.

Fig. 4. System-generated vs human-written headnote of one of the judgment from IHCP. The summary produced with our model is factually correct and fluent in its language, but the ROUGE scores obtained for this document were low.

VI. CONCLUSIONS

In this research, we have employed transfer learning based transformer model for legal text summarization. Transformer based models have been introduced recently to deal with the long input sequence lengths, but at the cost of higher processing power. Moreover, the available judgments are not enough to train a model from scratch. Whereas, deep neural network models are data hungry. Therefore, we proposed fine-tuning transformer models on a downstream task for legal text summarization. The results obtained through evaluation of this approach on the prepared dataset have shown an improved and satisfactory performance. The results have been verified on the judgments from AustL-II using the standard evaluation metric in text summarization. This is the first time that deep neural networks have been used to summarize the legal documents of Pakistan. This work provides a baseline for future research involving our dataset, making it our second contribution.

Our future work has two directions. Firstly, the future efforts can concentrate on overcoming the output word token limit for the incomplete sentences, since it seems to be a limitation of our system. Secondly, different variants of the trained model can be explored in further research to exploit the full potential of this approach. Although, we have not specified any

criteria to select sentences for the generated summary from separated different portions of the document depending upon its classification for it will make sure to select sentences from all parts of the document and decrease the input document length. But, for such kind of thematic segmentation, we require help from legal experts to provide such a baseline.

REFERENCES

[1] L. Abualigah, M. Q. Bashabsheh, H. Alabool, and M. Shehab, "Text summarization: A brief review," *Recent Advances in NLP: the case of Arabic language*, pp. 1–15, 2020 (cit. on p. 1).

[2] X. Mao, S. Huang, L. Shen, R. Li, and H. Yang, "Single document summarization using the information from documents with the same topic," *Knowledge-Based Systems*, p. 107 265, 2021 (cit. on p. 1).

[3] D. Suleiman and A. Awajan, "Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges," *Mathematical Problems in Engineering*, vol. 2020, 2020 (cit. on pp. 1, 2).

[4] D. Anand and R. Wagh, "Effective deep learning approaches for summarization of legal texts," *Journal of King Saud University-Computer and Information Sciences*, 2019 (cit. on pp. 1, 2).

[5] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019 (cit. on p. 2).

[6] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020 (cit. on pp. 2, 3).

[7] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, PMLR, 2020, pp. 11 328–11 339 (cit. on p. 2).

[8] A. Farzindar, "Letsum, an automatic legal text summarizing system," in *Legal Knowledge and Information Systems: JURIX 2004, the Seventeenth Annual Conference*, IOS Press, vol. 120, 2004, p. 11 (cit. on p. 2).

[9] A. Kanapala, S. Jannu, and R. Pamula, "Summarization of legal judgments using gravitational search algorithm," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8631–8639, 2019 (cit. on p. 2).

[10] K. Merchant and Y. Pande, "Nlp based latent semantic analysis for legal text summarization," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2018, pp. 1803–1807 (cit. on p. 2).

[11] N. Bansal, A. Sharma, and R. Singh, "Fuzzy ahp approach for legal judgement summarization," *Journal of Management Analytics*, vol. 6, no. 3, pp. 323–340, 2019 (cit. on p. 2).

[12] E. Elwany, D. Moore, and G. Oberoi, "Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding," *arXiv preprint arXiv:1911.00473*, 2019 (cit. on p. 2).

[13] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual keyword extraction and summarization: State-of-the-art," *Information Processing & Management*, vol. 56, no. 6, p. 102 088, 2019 (cit. on p. 2).

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018 (cit. on p. 2).

[15] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008 (cit. on p. 2).

[16] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020 (cit. on p. 2).

[17] *LED for legal summarization of documents*, https://huggingface.co/nsi319/legal-led-base-16384, [Online; accessed 9-Oct-2021] (cit. on p. 3).

[18] *U.S. Securities and Exchange Commission*, https://www.sec.gov/litigation/litreleases.htm, [Online; accessed 9-Oct-2021] (cit. on pp. 3, 4).

[19] *Sentence Splitter - Maltese Language Software Services*, http://metanet4u.research.um.edu.mt/SentenceSplitter.jsp, [Online; accessed 9-Oct-2021] (cit. on p. 3).

[20] K. Ganesan, "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks," *arXiv preprint arXiv:1803.01937*, 2018 (cit. on p. 3).

[21] *Cases & Legislation*, https://www.austlii.edu.au/au/other/AUPrivCS/, [Online; accessed 13-Sept-2021] (cit. on p. 4).

[22] V. Pandya, "Automatic text summarization of legal cases: A hybrid approach," *arXiv preprint arXiv:1908.09119*, 2019 (cit. on p. 4).

[23] Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Deep transfer reinforcement learning for text summarization," in *Proceedings of the 2019 SIAM International Conference on Data Mining*, SIAM, 2019, pp. 675–683 (cit. on p. 4).

[24] W. Tay, A. Joshi, X. J. Zhang, S. Karimi, and S. Wan, "Red-faced rouge: Examining the suitability of rouge for opinion summary evaluation," in *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, 2019, pp. 52–60 (cit. on p. 4).