Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking

Mohamed Elaraby, Yang Zhong, Diane Litman
University of Pittsburgh
Pittsburgh, PA, USA
{mse30,yaz118,dlitman}@pitt.edu

Abstract

We propose a simple approach for the abstractive summarization of long legal opinions that considers the argument structure of the document. Legal opinions often contain complex and nuanced argumentation, making it challenging to generate a concise summary that accurately captures the main points of the legal opinion. Our approach involves using argument role information to generate multiple candidate summaries, then reranking these candidates based on alignment with the document's argument structure. We demonstrate the effectiveness of our approach on a dataset of long legal opinions and show that it outperforms several strong baselines.

1 Introduction

Legal opinions contain implicit argument structure spreading across long texts. Existing summarization models often struggle to accurately capture the main arguments of such documents, leading to summaries that are suboptimal (Xu et al., 2021; Elaraby and Litman, 2022). We propose an approach for the abstractive summarization of long legal opinions that leverages argument structure.

Legal opinions often follow a specific argumentative structure, with the main points of the argument being presented clearly and logically (Xu et al., 2021; Habernal et al., 2022; Xu and Ashley, 2022). Prior work has shown that by considering this structure during summarization, it is possible to generate extractive and abstractive summaries that more accurately reflect the original argumentation in the document (Elaraby and Litman, 2022; Zhong and Litman, 2022; Agarwal et al., 2022). In this paper, we present a framework for abstractive summarization of long legal opinions that extends this literature by *leveraging argument structure during summary reranking* to both generate and score candidates. Our method involves utilizing the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) model to generate multiple candidate summaries by training it on various input formats. This allows for the consideration of different argument representations in the summary generation process. Additionally, we use beam search to further diversify the output. Finally, we rank the candidate summaries by measuring their lexical similarity to the input's main arguments.

We evaluate our approach on a dataset of long legal opinions obtained from the Canadian Legal Information Institute (CanLII)[1] and demonstrate that our method outperforms competitive baselines. Our results with ROUGE and BERTScore (Lin, 2004; Zhang et al., 2019) suggest that considering the argumentative coverage of the original opinions can lead to a more effective selection of summaries.

Our contributions are: **(1)** We propose a simple reranking approach that takes into account the argumentative structure of legal opinions to improve over the standard finetuning of generation models. **(2)** We demonstrate through empirical results and ablation analysis reasons for the effectiveness of our approach for summarizing long legal opinions. Our code can be accessed through this repository: https://github.com/EngSalem/legalSummReranking

2 Related Work

Long Legal Document Summarization Legal documents have a distinct format, with a hierarchical structure and specialized vocabulary that differs from that of other domains (Kanapala et al., 2019). They also tend to be longer in length (Kan et al., 2021; Huang et al., 2020; Moro and Ragazzi, 2022), which has led to the use of transformer models with sparse attention mechanisms (Michalopoulos et al., 2022; Guo et al., 2022; Beltagy et al., 2020) to reduce the complexity of encoding lengthy text. Legal *opinions*, in particular, have a complex argu-

[1]Data was obtained through an agreement with CanLII (https://www.canlii.org/en/).

Findings of the Association for Computational Linguistics: ACL 2023, pages 7601–7612

mentative structure that spans across the text, making it crucial to address in summaries (Xu et al., 2021; Xu and Ashley, 2022; Elaraby and Litman, 2022). *We use prior legal opinion summarization methods as evaluation baselines.*

Summarization and Argument Mining Using a dialogue summarization dataset with argument information, Fabbri et al. (2021b) converted an argument graph into a textual format to train a summarizer. For legal documents, Agarwal et al. (2022) used argument role labeling to improve extractive summarization using multitask learning. Elaraby and Litman (2022) blended argument role labeling and abstractive summarization using special markers, generating summaries that better aligned with legal argumentation. *We incorporate the models of Elaraby and Litman (2022) into summary reranking and further improve performance.*

Second Stage Reranking Generating multiple outputs and reranking them according to certain criteria has been successfully applied in NLP downstream applications including abstractive summarization. Some methods use different input formats to generate multiple outputs. Oved and Levy (2021) perturbed input multi-opinion reviews to generate multiple candidate summaries, then ranked them using coherency. Ravaut et al. (2022) used a multitask mixture of experts to directly model the probability that a summary candidate is the best one. Liu and Liu (2021) ranked candidate summaries generated from 16 diverse beam searches to improve news summarization in terms of ROUGE score. Liu et al. (2022) presented a novel technique for summary reranking that involves a non-deterministic training objective. Their approach enables the model to directly rank the summaries that are probable from beam-search decoding according to their quality. *We rely on distinct argument-aware input formats in addition to diverse beam decoding to develop our argument-aware reranking method.*

3 Annotated Dataset

We employ the annotated subset (Xu et al., 2021; Elaraby and Litman, 2022) of the **CanLII** dataset (Zhong and Litman, 2022) used in prior summarization research of legal opinions. This subset contains 1049 opinion/summary pairs annotated with sentence-level argument role labels for both input documents and reference summaries. The input opinions have mean/max lengths of 4375/62786 words, motivating us to use models for long text.

Recent work has proposed argument role taxonomies aligned with structures commonly found in legal text (Habernal et al., 2022; Xu et al., 2021). The CanLII data was **annotated for argument roles** using the **IRC scheme** for legal opinions (Xu et al., 2021), which divides argument roles into **Issues** (legal questions which a court addressed in the document), **Reasons** (pieces of text which indicate why the court reached the specific conclusions), and **Conclusions** (court's decisions for the corresponding issues). We use these 3 fine-grained IRC labels, as well as collapse them into a single argumentative label, to incorporate argument structure into our models. An IRC-annotated opinion and summary pair can be found in Appendix A.

4 Model and Methods

Our proposed method follows the generate and ranking paradigm and can be split into two parts. First, we explore techniques to utilize an argumentation augmented LED model to generate multiple candidate summaries $\mathbb{S}$. Second, we propose a function μ that scores a summary S where $S \in \mathbb{S}$ based on its argumentative alignment with the input document. The best candidate S^* is selected such that $S^* = \arg\max_{S_i \in \mathbb{S}}\{\mu(S_1), \mu(S_2), .., \mu(S_n)\}$. Figure 1 shows an overview of our approach.

4.1 Generating Candidates: Argument-Aware Training + Diverse Decoding

Diverse decoding techniques such as beam-search can help diversify the *summary output*; however, it's only limited to the underlying language model used in the decoder and is completely isolated from the input format. Alternatively, we propose to complement the beam search via finetuning LED on three *different input formats*. We refer to this model as $M_{arg-augmented}$ such that the model parameter $\theta^*_{arg-augmented}$ is selected such that

$$\theta^*_{arg-augmented} = \arg\max_{\theta} P(S|\mathbb{X})$$

During finetuning, S is the reference summary, θ represents the trainable model parameters, and $\mathbb{X}$ is a set of inputs $\mathbb{X} = \{X_{raw}, X_{arg\_binary}, X_{arg\_finegrained}\}$, where X_{raw} is the input without the argument markers, X_{arg_binary} is the input document with binary argument markers added to highlight argument role sentences, and $X_{arg\_finegrained}$ is the input document with the fine-grained argumentative markers added to also delineate

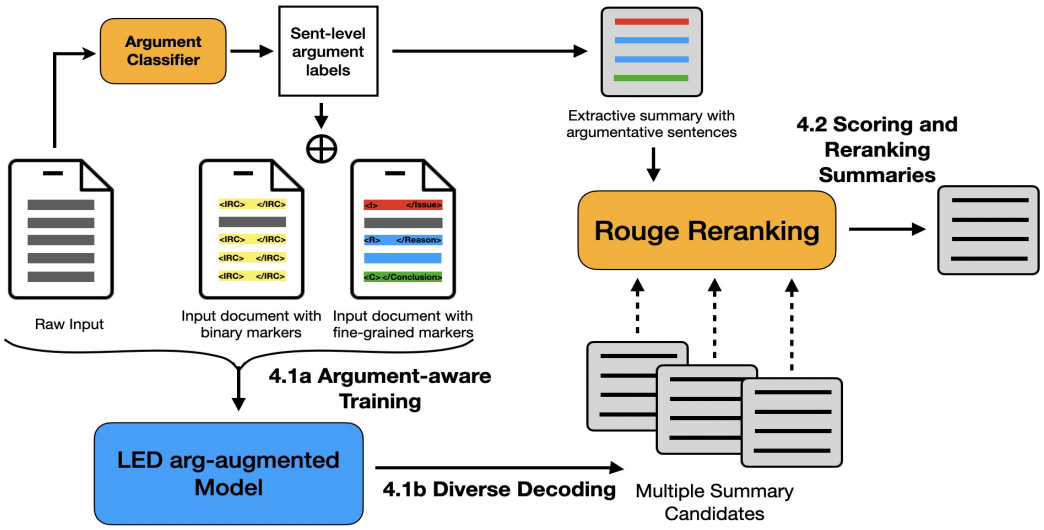

Figure 1: Illustration of basic components of our approach. For input documents with fine-grained markers, colored sticks are sentences with argument role labels of Issue, Reason, and Conclusion. We used one IRC label for the binary version. In our real dataset, marked sentences are surrounded with special markers (Appendix C).

the roles (i.e., Issue, Reason, Conclusion). These three representations of the input share the same reference summary, meaning that we augmented the training data three times. Table 1 shows an example of the distinct representations of our new training data. At inference time, we use the predicted markers by adopting the argument mining code[2] from Elaraby and Litman (2022) instead of the manually labeled ones to construct $\hat{X}_{arg\_binary}, \hat{X}_{arg\_finegrained}$ of $\hat{\mathbb{X}}$ where $\hat{\mathbb{X}} = \{X_{raw}, \hat{X}_{arg\_binary}, \hat{X}_{arg\_finegrained}\}$. Our incentive is that different formats of the input would yield different generated summaries that take into account different representations of the argumentative structure in the input.

Input format	Example
X_{raw}	$S_1\|S_2\|...\|$ Issue Sentence | Reason Sentence |...
X_{arg_binary}	$S_1\|S_2\|...\|$ <IRC> Issue Sentence </IRC> | <IRC> Reason Sentence </IRC> |...
$X_{arg\_finegrained}$	$S_1\|S_2\|...\|$ <Issue> Issue Sentence </Issue> | <Reason> Reason Sentence </Reason> |...

Table 1: An example of $\mathbb{X}$, which consists of three data points in different formats that share the same reference summary. In the table, $S1$ refers to the first sentence of the text, $S2$ to the second sentence, and so on. <IRC>, <Issue>, and <Reason> are the argumentative marker tokens described in Appendix C.

4.2 Scoring and Reranking Summaries

We propose a scoring method to rank the candidate summaries based on their capability to capture the main argument points in the input. First, we employ a sentence-level argument role classifier to extract sentences with argument roles $\hat{X}_{args}$. The predicted sentences are used to construct an extractive summary. Then, we measure the lexical overlap between a generated candidate summary $\hat{S}$ and the constructed extracted one using *ROUGE-1 F1-score*[3], to compute a score to each candidate summary that represents its alignment with the legal opinion argument content. Our scoring function μ can be written as $\mu = ROUGE1(\hat{X}_{args}, \hat{S})$.

5 Experiments

All models use *LED-base* checkpoint as a base model. *LED-base* encodes up to $16k$ tokens, which fits our long inputs. All experiments use 5-fold cross-validation, with the 4-fold documents split into 90% training and 10% validation; the validation split is used to select the best checkpoint.[4]

We compare all rank-based methods (baseline and proposed) to **abstractive baselines** previously explored in legal opinion summarization: *finetune LED-base* (which refers to vanilla model finetuning

[2]We retrain the model (details in Appendix B), yielding a macro-average of 0.706 F1 on the four-way classification (Issues, Reasons, Conclusions, Non-argumentative).

[3]Using R-2 or R-L makes little difference (Appendix E).

[4]Full experimental details can be found in Appendix B.

Experiments	ID	Model	R-1	R-2	R-L	BS	src. marker
Abstractive baselines	1	finetune LED-base	47.33	22.80	44.12	86.43	-
	2	arg-LED-base (binary markers)	48.85	24.74	45.82	86.79	predicted
	3	arg-LED-base (fine-grained markers)	49.02	24.92	45.92	86.86	
	4	arg-LED-base (binary markers)	50.64	26.62	47.48	86.90	oracle
	5	arg-LED-base (fine-grained markers)	51.07	27.06	48.01	86.92	
Ranking baselines	6	baseline ranking	49.79	25.13	46.63	86.87	predicted
	7	arg-LED-base (fgrain) + diverse beams	50.92	26.06	47.74	86.87	
	8	baseline ranking	51.85	27.31	48.61	87.26	oracle
	9	arg-LED-base (fgrain) + diverse beams	52.74	**27.93**	49.50	**87.46**	
Our framework	10	arg-augmented-LED	50.52	24.82	47.19	86.85	predicted
	11*	**arg-augmented-LED + diverse beams**	*54.13*	*27.02*	*50.14*	*87.38*	
	12	arg-augmented-LED	51.96	25.69	48.56	87.03	oracle
	13	**arg-augmented-LED + diverse beams**	**54.30**	27.00	**50.80**	87.35	

Table 2: Summarization ROUGE (R1, R2, RL) and BertScore (BS) cross-validation results. Best results in each column are **bolded** when obtained with the oracle markers and *italicized* with predicted markers. For full framework (rows 11/13), * indicates results are statistically significant in all scores over best argument-aware baseline (row-3).

using our dataset), and *arg-LED-base* (Elaraby and Litman, 2022) (which finetunes LED on the *dataset blended with argument markers* that mark the start and the end of each argument role in the input).[5]

We also compare our proposed rank-based approach from Section 4 with **ranking baselines** that use different input formats or diverse decoding alone. Specifically, we have employed ranking on top of the output of the three LED models outlined in Elaraby and Litman (2022) which are trained on distinct argument aware input formats (we refer to this model as "baseline ranking"). Additionally, for diverse decoding, we have employed different beam widths within the range of 1 and 5 [6] on top of the model trained on the input with fine-grained markers (arg-LED-fine-grained), which achieved the best abstractive baseline ROUGE results.

All models utilizing argument markers employed both *oracle* and *predicted* conditions during inference time, using human annotations or argument mining respectively, to produce the markers.

6 Results and Discussion

Table 2 shows our results in terms of *ROUGE-score* (Lin, 2004) and *BERTScore* (Zhang et al., 2019), computed using *SummEval* (Fabbri et al., 2021a)[7].

Utility of any Ranking The ranking-based methods (rows 6-13) consistently outperform the abstractive baselines[8] (rows 1-5) in both predicted and oracle conditions. Also, abstractive baseline results (rows 1-5) align with those of Elaraby and Litman (2022), where leveraging fine-grained markers in the input yields the highest scores.

Utiliy of Proposed Ranking Framework and its Components In the predicted case, our proposed arg-augmented-LED (row 10) improves over the abstractive baselines (rows 1-3) with ranges $1.5 - 3.19$ and $1.27 - 3.07$ in ROUGE-1 and ROUGE-L respectively, while maintaining a limited drop of 0.1 and 0.01 in terms of ROUGE-2 and BS respectively. Similarly, compared to our ranking baselines, our proposed model improves over ROUGE-1 and ROUGE-L scores obtained by baseline ranking with ranges $0.56 - 0.73$ while dropping in ROUGE-2 and BS by 0.31 and 0.02 points respectively. This indicates that incorporating argument information into the source inputs can lead to the generation of effective summary candidates. Our best predicted results were achieved by combining our proposed model with diverse beam decoding (row 11), which combines the strengths of various input formats and multiple beam decoding, resulting in statistically significant improvements over the previously proposed argument-aware abstractive baseline (row 3).

Inference with Predicted versus Oracle Argument Roles For the same model, predicted markers can impact the summarization results. In prior baselines (rows 3 and 5), we observe a drop in ROUGE score with ranges $2.05 - 2.14$, and 0.06 in terms of BS when switching from oracle to predicted markers. This observation is consistent among row

[5] Argument marker details can be found in Appendix C.
[6] We ran out of memory with BeamWidth > 5.
[7] https://github.com/Yale-LILY/SummEval
[8] See Appendix D for extractive baseline results.

6 and 8; and row 10 and 12. With our proposed arg-augmented-LED and diverse beam decoding, this performance gap is mitigated and reduced to $-0.02 - 0.66$ and -0.03 in ROUGE and BS, respectively (rows 11 and 13). We believe this is due to the combination of distinct argumentative formats and diverse decoding, allowing more diverse candidates to be considered in the ranking and enhancing robustness to noisy predictions during inference.

7 Conclusion and Future Work

We proposed a framework for improving the summarization of long legal opinions by combining distinct argument formats of the input with diverse decoding to generate candidate summaries. Our framework selects the summary with the highest lexical overlap with the input's argumentative content. Our results indicate that ranking alone can improve over abstractive baselines. Moreover, combining ranking with our proposed candidate generation method improves results while maintaining robustness to noisy predictions. In future research, we plan to incorporate human expert evaluations to compare automatic metrics with human ratings. Also, we aim to explore the impact of using noisier argument roles during training on a larger corpus by using the predicted markers obtained from our smaller dataset to experiment with the remaining unannotated portion of the CanLII dataset.

Limitations

The primary constraints encountered in our research result from our dependence on a single dataset for experimentation and computing resource limitations. Despite these, we postulate that our ranking-based methodology can be utilized for any summarization task that necessitates robust correspondence with a specific structure within the input. To validate this hypothesis, further experimentation is required to assess the generalizability of our technique to alternative datasets and domains. In addition, our limited computational resources prevented us from experimenting with other long document encoder-decoder models such as BigBird and LongT5 (Michalopoulos et al., 2022; Guo et al., 2022) as well as using higher beam widths during decoding. Furthermore, the cost and complexity of procuring expert evaluators within the legal domain resulted in using automatic metrics alone.

Ethical Considerations

The usage of the generated summary results from legal opinions remains important. Abstractive summarization models have been found to contain hallucinated artifacts that do not come from the source texts (Kryscinski et al., 2019; Zhao et al., 2020; Kryscinski et al., 2020). While our model incorporated the argument structure of the source article, the generation results may still carry certain levels of non-factual information and need to be utilized with extra care. Similarly, as mentioned in the prior line of works using CanLII (Elaraby and Litman, 2022; Zhong and Litman, 2022), CanLII has taken measures to limit the disclosure of defendants' identities (such as blocking search indexing). Abstractive approaches may cause user information leakage. Thus using the dataset needs to be cautious to avoid impacting those efforts.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2040490 and by Amazon. We would like to thank the members of both the Pitt AI Fairness and Law Project and the Pitt PETAL group, as well as the anonymous reviewers, for valuable comments in improving this work.

References

Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. 2022. Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1857–1872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. Ms^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander Richard Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021b. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Christoph Burchard, et al. 2022. Mining legal arguments in court decisions. *arXiv preprint arXiv:2208.06178*.

Yuxin Huang, Zhengtao Yu, Junjun Guo, Zhiqiang Yu, and Yantuan Xian. 2020. Legal public opinion news abstractive summarization by incorporating topic information. *International Journal of Machine Learning and Cybernetics*, 11(9):2039–2050.

Tai-Jung Kan, Chia-Hui Chang, and Hsiu-Min Chuang. 2021. Home appliance review research via adversarial reptile. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 183–191, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen Mckeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. 2022. ICDBigBird: A contextual embedding model for ICD code classification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 330–336, Dublin, Ireland. Association for Computational Linguistics.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long legal documents in low-resource regimes. In *Proceedings of the Thirty-Six AAAI Conference on Artificial Intelligence, Virtual*, volume 22.

Nadav Oved and Ran Levy. 2021. Pass: Perturb-and-select summarizer for product reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 351–365.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. Summareranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Huihui Xu and Kevin D. Ashley. 2022. Multi-granularity argument mining in legal texts. In *International Conference on Legal Knowledge and Information Systems*.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery.

Yang Zhong and Diane Litman. 2022. Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 322–337, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Argument Role Labeling in CanLII Cases

The concept of argument roles, specifically issues, reasons, and conclusions, is of paramount importance in legal case summarization. An illustration, presented in Figure 2, demonstrates the annotation of these roles in the input text of a legal opinion and its associated summary. This example shows that the issues, reasons, and conclusions can effectively encapsulate the critical points of discussion within the court, the ultimate decision reached, and the rationale for said decision.

B Experimental Setup and Hyper-parameters

LED experiments For all of our LED-base experiments, we use the LED-base implementation by the *HuggingFace Library* (Wolf et al., 2020). We finetune the LED-base model for 10 epochs. We select our best model based on the $ROUGE-2$ score on the validation set. We rely on the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of $2e-5$ to update the LED-base weights. We also employ an early stopping with 3 epoch patience to avoid overfitting during training.

Argument Role Classification Our argument role classifier leverages a finetuned *legalBERT* (Zheng et al., 2021) model due to its superiority to other contextualized embeddings-based models like BERT (Devlin et al., 2019) and ROBERTa (Liu et al., 2019) as shown in Elaraby and Litman (2022); Xu et al. (2021). We utilized the same training setting and hyperparameters described in Elaraby and Litman (2022) to train the *5-fold* cross-validation sentence level argument classifiers used in our experiments. [9]

C Argumentative Markers

In abstractive summarization, special markers can indicate the most important parts of a text that

[9] Classifier code is available at https://github.com/EngSalem/arglegalsumm/tree/master/src/argument_classification

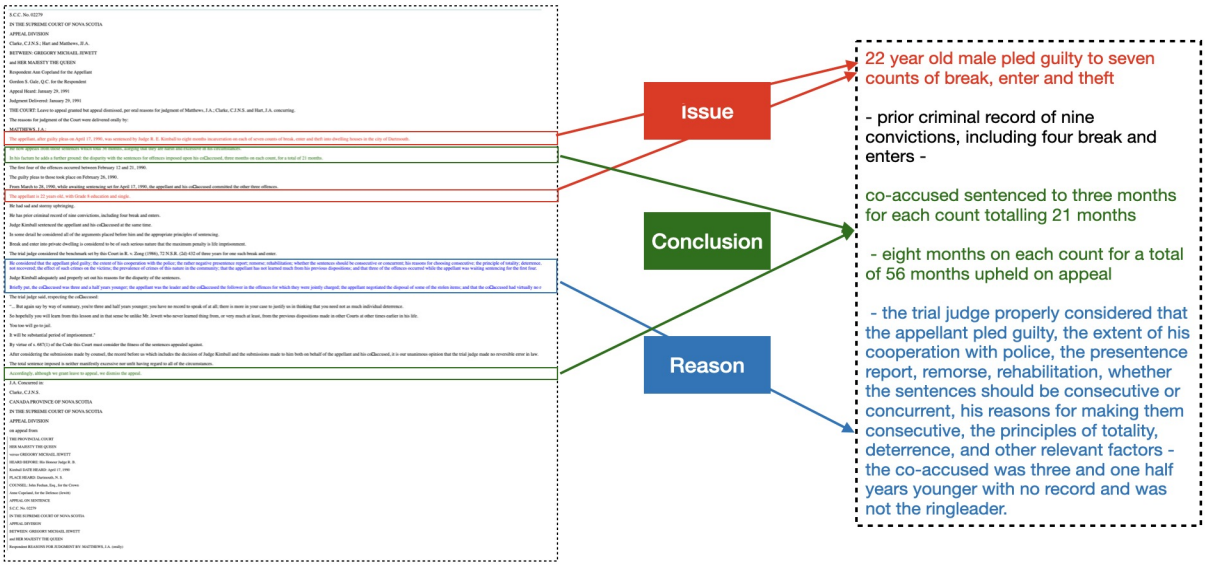

Figure 2: An example of the annotated Issue, Reason, and Conclusion sentences in the CanLII dataset's legal opinion and summary pair (ID: a_1991canlii2497).

	Example of using argument markers
Original	The plaintiful should have taken more appropriate measures to avoid the accident.
Binary Markers	**<IRC>** The plaintiful should have taken more appropriate measures to avoid the accident. **</IRC>** .
Fine-grained Markers	**<Reason>** The plaintiful should have taken more appropriate measures to avoid the accident. **</Reason>.**

Table 3: Example of using argumentative marker tokens

ground the summary (Khalifa et al., 2021; DeYoung et al., 2021). These markers can be added to the text by a human annotator, or they can be generated automatically by a model. These markers can take many forms, such as highlighting certain words or phrases or adding special tags to certain sentences. A summarization model can use them to identify the key parts of the text that should be included in the summary while also considering the overall structure and coherence of the text. This can help to improve the accuracy and effectiveness of the summarization process, especially when the text is long or complex. In this work, we use marker sets proposed by Elaraby and Litman (2022) to distinguish between argumentative and non-argumentative sentences.

Binary markers The binary markers aim to distinguish argumentative and non-argumentative sentences regardless of the type of the argument role (i.e, issues, reasons, or conclusions). In our work, we used the markers `<IRC>,</IRC>` to highlight the start and end of each argumentative sentence.

Fine-grained markers We also used the markers designated to distinguish between each argument role type by using the markers `<Issue>,</Issue>`, `<Reason>`, `</Reason>`, `<Conclusion>`, `</Conclusion>`.

Table 3 shows an example of using different argumentative markers to highlight the start and end of a "Reason" sentence.

D Extractive Baselines

In addition to the abstractive baselines, we compare our methods to graph-based unsupervised extractive baselines built on top of *HipoRank* (Dong et al., 2021) and extractive baselines based on *Extractive-BERT* (Zheng and Lapata, 2019), which were leveraged before on the same dataset (Zhong and Litman, 2022). Table 4 shows our abstractive summarization results compared to the extractive baselines in cross-validation settings. Our ranking-based methods show consistent improvement over both the extractive and the abstractive baselines.

E ROUGE based ranking results

Table 5 shows a comparison between the usage of ROUGE-1, ROUGE-2, and ROUGE-L as potential ranking criteria to select the summary that aligns with the predicted argumentative content outlined in the input legal opinion. While there is no substantial differences between results with each ROUGE metric, ROUGE-L seems to have marginally lower scores compared to ROUGE-1, and ROUGE-2.

Experiments	ID	Model	R-1	R-2	R-L	BS	src. marker
Extractive baselines	1	sentence-level legalBERT	49.66	28.42	46.72	86.54	-
	2	HipoRank	41.24	17.19	38.54	81.67	-
	3	HipoRank rewighted	42.88	18.03	39.99	84.11	-
	4	Extractive BERT	43.053	17.75	39.99	84.15	-
Abstractive baselines	5	finetune LED-base	47.33	22.80	44.12	86.43	-
	6	*arg-LED-base (binary markers)*	*48.85*	*24.74*	*45.82*	*86.79*	*predicted*
	7	*arg-LED-base (fine-grained markers)*	*49.02*	*24.92*	*45.92*	*86.86*	
	8	arg-LED-base (binary markers)	50.64	26.62	47.48	86.90	oracle
	9	arg-LED-base (fine-grained markers)	51.07	27.06	48.01	86.92	
Ranking baselines	10	*baseline ranking*	*49.79*	*25.13*	*46.63*	*86.87*	*predicted*
	11	*arg-LED-base + diverse beams*	*50.92*	*26.06*	*47.74*	*86.87*	
	12	baseline ranking	51.85	27.31	48.61	87.26	oracle
	13	arg-LED-base + diverse beams	52.74	**27.93**	49.50	**87.46**	
Our framework	14	*arg-augmented-LED*	*50.52*	*24.82*	*47.19*	*86.85*	*predicted*
	15	***arg-augmented-LED + diverse beams***	***54.13***	***27.02***	***50.14***	***87.38***	
	16	arg-augmented-LED	51.96	25.69	48.56	87.03	oracle
	17	**arg-augmented-LED + diverse beams**	**54.30**	27.00	**50.80**	87.35	

Table 4: Full Extractive and Abstractive Results

Ranking metric	Model	R-1	R-2	R-L	BS
ROUGE-1	baseline ranking	49.79	25.14	46.63	86.87
	arg-LED + diverse beams	50.92	26.06	47.74	86.87
	arg-augmented-LED	50.52	24.82	47.19	86.85
	arg-augmented-LED + diverse beams	54.13	27.02	50.14	87.38
ROUGE-2	baseline ranking	49.39	25.27	46.30	86.88
	arg-LED + diverse beams	50.35	26.24	47.23	87.15
	arg-augmented-LED	49.46	24.16	46.19	86.71
	arg-augmented-LED + diverse beams	54.00	27.82	50.08	87.42
ROUGE-L	baseline ranking	49.12	24.97	46.01	86.84
	arg-LED + diverse beams	49.84	25.66	46.72	87.07
	arg-augmented-LED	49.16	23.87	45.87	86.67
	arg-augmented-LED + diverse beams	53.34	26.88	49.98	87.39

Table 5: R1, R2, RL ranking scores with predicted argumentative markers

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
see section 8 (Limitations) after conclusion

☑ A2. Did you discuss any potential risks of your work?
see section 9, Ethical Consideration after limitation section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, see Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Grammarly is used to help in checking grammar and writing style.

B ☑ Did you use or create scientific artifacts?

Yes, please see section 4 models.

☑ B1. Did you cite the creators of artifacts you used?
Please see section 3 and 4 dataset and models.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Please see footnote in section 1 on the license of the dataset used

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, please refer to sections 1, 3, 4, and 5 discussing previous artifacts and how we use them.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Doesn't apply to our dataset owe used

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 3 and the appendix A.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Please refer to section 3 datasets for details

C ☑ Did you run computational experiments?

section 4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We briefly discusses the computational infrastructure in the limitation and appendices.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental details in the appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
results can be found in section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4, and 5 and appendix B

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.