Improving Accessibility of SCOTUS Opinions: A Benchmark Study and a New Dataset for Generic Heading Prediction and Specific Heading Generation

Malek Yaich and Nicolas Hernandez
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
malek.yaich@etudiant-fst.utm.tn, nicolas.hernandez@univ-nantes.fr

Abstract

The opinions of the U.S. Supreme Court (SCOTUS) are known for their extensive length, complex legal language, and lack of titled sections, which pose significant challenges for accessibility and comprehension. This paper defines the task of automatic section titling by proposing both generic and specific headings for each section. Given the scarcity of sections with headings in SCOTUS, we study the possibility of using data from lower courts for training models. A dataset of sections with generic or specific headings covering three courts (SCOTUS and two lower courts) was compiled. A supplementary SCOTUS set was manually annotated with these two types of titles. In order to establish a benchmark, we provide the performance of different systems trained for each subtask: For generic heading prediction, we compare the performance of fine-tuning non-contextual, general and domain-oriented pretrained language models. Transformer-based sequence-to-sequence models are considered for specific heading generation. Our results show that a fine-tuned LegalBERT can achieve a F1 score of about 0.90 % in predicting generic headings. They also show that BART and T5 have similar performance in generating specific headings and that, although this performance is good, there is still room for improvement. In addition, we provide a human assessment to support the generation experiment and show a quasi-linear correlation between human degrees of agreement and the results of conventional measures such as ROUGE and BERTScore.

1 Introduction

As the highest court in the United States, dealing with constitutional issues and federal law, the Supreme Court (SCOTUS) defines a model of society whose global impact extends beyond the borders of the United States (see the recent decisions limiting the EPA's regulation of carbon emissions, for example). It is crucial that international legal practitioners who do not speak American as their first language are able to access and understand these decisions. Nevertheless SCOTUS opinions are notoriously long and use specialised language, making them laborious to read and understand.

Various researchers have shown that the segmentation of text (Florax and Ploetzner, 2010; Lemarié et al., 2008; Weiss, 1983) and the titling of passages (Wiley and Rayner, 2000) improves the comprehension and memorability of text. Unfortunately, few SCOTUS opinions are divided into sections, and even fewer into titled sections, despite their length, which can run to several pages (See Section 3).

Based on (Goźdź-Roszkowski, 2024), we consider that the sections of SCOTUS'opinions can be described by two types of headings: generic headings and specific headings. Specific headings are unique to each section and describe the precise and detailed content of the section in question. Generic headings, on the other hand, are common to several sections and describe the type of information or main theme of the section without going into specific details. We define the task of section titling as involving two sub-tasks: 1) *predicting a generic heading* and 2) *generating a specific heading*. We reserve for future work the task of deciding which of these titles best describes a section.

Heading prediction is a text classification task focused on identifying high-level argumentative labels within the text. This task bears resemblance to the process of identifying moves in Swales' genre analysis (Swales, 2004), particularly in the legal domain, where it facilitates the categorization and structuring of complex legal documents. Heading generation can be seen as a form of text title generation (Omidvar and An, 2023; Bouchekif et al., 2015; Iwama and Kano, 2019), but at the section level (Field et al., 2020).

In this paper we propose to address the following questions: How do conventional classification approaches perform in predicting generic headings

Proceedings of the 31st International Conference on Computational Linguistics, pages 6827–6839
January 19–24, 2025.

when the target domain is aligned with the source domain (i.e. when the test data comes from the same source as the training data)? What are the results when the source domain is different? Does mixing training sources have an impact on model performance? Which approach gives the best results? What specific performance can be observed on SCOTUS? Can the models be trained on lower jurisdictions for the titling of SCOTUS sections? The same questions apply to the task of generating specific headings. In addition, to complete these questions, what is the human agreement on headings generated by a model trained on a source domain not aligned with the target domain? Are the evaluation scores equivalent between human-human and human-machine generations?

Following are the key contributions of our work:

1. We defined a new task of titling sections either with a generic heading or with a specific one;
2. We propose a study on the exploitation of lower courts to compile training data and learn models to address the task on the highest legal court in US (aka SCOTUS);
3. We establish a benchmark of well-known systems both for generic heading prediction and specific heading generation;
4. We release under an open source licence a new dataset, `scotus-heading`[1], which compiles sections with generic or specific headings covering three courts (SCOTUS and two lower courts); with a SCOTUS part manually annotated with these two types of titles.

2 Related work

Structure of SCOTUS opinions Integrating research methods from linguistics with contemporary legal argumentation theory, (Goźdź-Roszkowski, 2024) compares "The structure of US opinion" with "The structure of Polish judgment," offering insights into the top-level sections and potential headings found in SCOTUS opinions. A court opinion typically begins by presenting the legal issue and questions for consideration, outlining the differences in opinion between litigants, and summarizing their arguments. The main justification sections are marked with Roman numerals (I, II, III), and each part is further divided using capital letters (A, B). Part I usually details the case's facts, Part II interprets relevant laws, and Part III analyzes the arguments presented. SCOTUS opinions may have more than three parts and exhibit flexible and variable organization.

Automatic titling As with abstractive summarisation, the majority of recent studies on title generation propose using a sequence-to-sequence architecture (Iwama and Kano, 2019; Omidvar and An, 2023) to generate titles from a source text. These approaches require the availability of training data to supervise the learning of the models. This also applies to the generation of section headings. (Field et al., 2020), for example, explored the generation of abstract titles for Wikipedia sections by evaluating the use of transformer encoders paired with various decoders. They observed that "Wikipedia section titles contain a mix of short abstractive headings like "History" and longer extractive headings like song titles, where many of the words in the section title also appear in the section text." Although they have not used this distinction to train their models, they have observed distinct generation capacities in their test data, depending on the decoder architectures used. The types of titles were distinguished using a heuristic based on their length. This observation underlines the importance of learning to distinguish between these titles in models. The types of section headings in our data also differ in size, but while our 'generic heading' category may be close to Field et al.'s 'abstract' category, our 'specific' category does not share the characteristic of being constructed by extracting words from the section. In our study we decided to study the production of these two types of titles separately, leaving their joint processing for later.

Concerning the task of section heading prediction, (Zhang et al., 2022) proposed a multi-task BERT model to identify and classify sections in clinical notes based on a predefined list of section markers, considered as section headings. This model operates on the hypothesis that section content is similar across distributions and can be used to generate a robust section classifier. Similarly, in this article, we investigate the potential of reusing headings from lower courts to title the sections of SCOTUS opinions, providing evidence that section content is consistent across different courts.

Similarly to (Field et al., 2020) and (Zhang et al., 2022), we take as input the content of a given section. Like (Field et al., 2020), we observe that this approach does not allow us to take advantage of the

[1] https://huggingface.co/datasets/taln-ls2n/scotus-heading

hierarchical contextualisation of the section (which also can be considered as a future direction to explore). Furthermore, this simplification avoids the limitation suffered by all transformer-based models (Vaswani et al., 2017), namely the processing of long input sequences. As a result, we did not need to use architectures dedicated to summarizing long documents, such as the Longformer-Encoder-Decoder (LED) based on sparse attention (Beltagy et al., 2020) or LOCOST based on state-space models (Le Bronnec et al., 2024).

Although the use of LLMs and in-context learning for generation tasks is a current avenue of research (Xu and Ashley, 2023), this approach remains costly in terms of energy and money, and is hampered by the problem of input data size. So we decided not to use it for this current study either.

Data scarcity Another challenge we have encountered is the lack of sufficient SCOTUS data to evaluate our model. In response to similar data scarcity issues, (Chen and Chen, 2019) propose adversarial domain adaptation using artificial titles for abstractive title generation. In this approach, the first line of each section is used as an artificial title to capture the style of the unlabelled target, thereby bridging the gap between the source and target domains.

Evaluation of automatic summarization systems "is problematic: there is no natural upper bound on the quality of summarisation systems, and even humans are excluded from performing optimal summarisation" (Schluter, 2017). It is important to keep in mind what aspect a given metric is intended to measure and most of all, how the metric correlates with a human preference (Fu et al., 2024). In a general framework, a candidate heading is compared to a reference heading (Field et al., 2020; Iwama and Kano, 2019; Omidvar and An, 2023; Chen and Chen, 2019). ROUGE-N (Lin, 2004) measures n-grams overlap to evaluate informativeness. BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), offer further insights into the semantic quality of generated headings. ROUGE-L evaluates the longest common subsequence to gauge fluency while perplexity (Jelinek et al., 2005) can lead to a more robust assessment. The FEQA metric (Durmus et al., 2020) with the "Question answering - question generation" proposes to evaluate factual consistency by asking a generative model to generate questions from a summary and compare its answers with the answers from the text source. More recently (Zheng et al., 2023) proposed to exploit Large Language Model as a judge or to use e the probability of a text being generated as an indicator of the quality (Fu et al., 2024). Although these measures hold promise as a replacement for human assessments, they require powerful computing capabilities. Qualitative human evaluations (van der Lee et al., 2019) remain the most valuable to assess any aspect utilizing techniques such as Likert scales and intrusion tests and measuring correlations.

3 Data analysis

Our dataset was sourced from CourtListener's bulk data, retrieved on February 28, 2023. **CourtListener**[2] is a free legal research platform offering access to millions of opinions from US courts.

3.1 SCOTUS statistics

The results of extracting sections and headings from SCOTUS opinions (See Table 1) indicate that the majority of SCOTUS sections (96,12%) lack headings[3]. With only 1,302 headings in total, the number is exceedingly low and insufficient for training any model to automate section titling.

As illustrated in Figure 1, the majority of SCOTUS sections exhibit a depth of 1, with the maximum depth observed reaching 4. It is noteworthy that 97.08% of depth-1 sections are devoid of headings. As the section depth increases, there is a discernible trend where judges increasingly incorporate headings. Of the 111 judges examined, 41 (36.9%) have contributed at least one title. This distribution underscores the diversity of the dataset, mitigating any potential bias stemming from the input of individual judges.

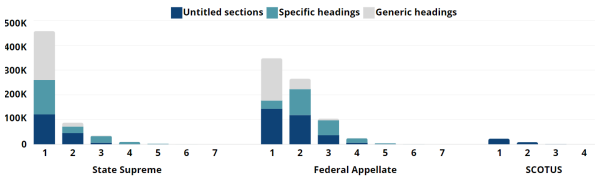

Figure 1: Distribution of specific headings, generic headings, and untitled sections by section depth across the three courts

[2] https://www.courtlistener.com

[3] The need for section titling is nonetheless real. Indeed, the average size of SCOTUS opinions with section headings is 12,113 NLTK words, or 24 pages, taking into account the rule of thumb of 500 words for a single-spaced page. As a matter of fact, SCOTUS lists 27,694 opinions of more than 500 words (i.e. approximately one page) with an average of 4,578 words, i.e. more than 4.5 theoretical pages.

To address the challenge of data scarcity, we considered reusing headings from lower courts to label SCOTUS sections, assuming that: 1) Lower courts consistently have section headings; 2) Courts share similar distributions of section headings; And 3) section contents are similar across different courts.

By leveraging these assumptions, we aim to ensure an adequate amount of training data to automate section titling, based on established practices in lower courts' opinions. We focus particularly on federal appellate and state supreme courts, given their proximity to SCOTUS within the American judicial hierarchy (Kim, 2015). SCOTUS, as the highest judicial authority in the United States, hears appeals from federal appellate courts, which review decisions from district courts and administrative agencies, as well as from state supreme courts, which handle appeals from lower state courts.

3.2 Federal and state courts statistics

Our final dataset comprises opinions from SCOTUS, federal appellate, and state supreme courts. We extracted sections and section headings from each court (See Table 1).

	SCOTUS	Federal Appellate	State Supreme
Opinions	493,203	360k	470k
Opinions with sections	5,289	230k	370k
Total sections	33,602	771,777	671,202
Total headings	1,302	457,531	444,568
Distinct headings	1,061	138,967	169,161
Unique headings	77.06%	83.47%	89.27%
>1,000 occurrences	-	18	21
% >1,000 occurrences	-	48.48%	45.76%
Avg. words per title	6.62	4.88	7.49

Table 1: Overview of SCOTUS data, compared to State Supreme and Federal Appellate corpus. "Distinct headings" refers to the total number of headings with duplicates removed. "Unique headings" refers to the number of headings that occur exactly 1 time. ">1,000 occurrences" refers to the number of headings exceeding 1000 occurrences.

Although the presence of headings in SCOTUS opinions is rare, the majority of sections in federal appellate (59%) and state supreme (66%) opinions contain headings. This confirms our hypothesis that these courts consistently utilize section headings. The headings from the three courts share similar characteristics, with comparable average lengths and percentages of unique headings, which lends support to our hypothesis of a shared distribution.

Analyzing the occurrence of headings, we found that some are repeated more than 1,000 times, accounting for 48.43% of the data in federal appellate court and 45.59% in state supreme court. Among these frequently occurring headings, 15 are common between federal appellate and state supreme courts, and 9 are shared across all three courts. These headings are primarily found at the first hierarchical level (See Figure 1) and tend to be short.

Based on the description of the structure of SCOTUS opinions in (Goźdź-Roszkowski, 2024), we consolidated these headings into 8 general labels. This consolidation involved grouping similar headings under unified labels, simplifying the data's organization while ensuring that all relevant headings were effectively represented. These labels are: Introduction, Background, Analysis, Jurisdiction, Issues, Standard of Review, Sufficiency of the Evidence, and Conclusion. The complete selection process is described in Appendix A.

Based on this analysis, we categorized section headings into two distinct types:

- **Specific headings** are unique to each section and describe the precise and detailed content of the section in question.
- **Generic headings**, on the other hand, are common to several sections and describe the general context or main theme of the section, without going into specific details, such as "Facts" or "Analysis".

4 Model

As we have identified two types of section headings, the process of section titling is divided into two distinct tasks: the prediction of a generic heading and the generation of a specific heading for each section.

4.1 Predicting generic section headings

The goal of this task is to predict appropriate generic headings from a predefined set of categories. To achieve this, we performed experiments using three models: FastText (Joulin et al., 2017), BERT (Devlin et al., 2019), and Legal-BERT (Chalkidis et al., 2020), fine-tuning the latter two specifically for the task of title classification.

FastText[4]: Serving as our baseline model, FastText employs average word or n-gram embeddings

[4] https://fasttext.cc/docs/en/supervised-tutorial.html

to represent documents. Despite its simplicity, FastText is known for its efficiency and effectiveness in text classification tasks, leveraging semantic information embedded in word representations.

BERT[5] (Bidirectional Encoder Representations from Transformers) is a state-of-the-art model for various NLP tasks. For our experiments, we used the base version and fine-tuned it for the specific task of title classification. The process was conducted over 5 epochs, with a batch size of 16 and a weight decay of 0.01 to improve generalization.

LegalBERT[6] is a specialized version of BERT pre-trained on legal corpora. For our study, we fine-tuned the base version under the same conditions as our fine-tuned BERT.

4.2 Generating specific section headings

In this task, we aim to generate specific headings that accurately describe the content of each section. We fine-tuned two pre-trained models, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), specifically for the heading generation task.

BART[7] is a denoising autoencoder coupled with a sequence-to-sequence modeling. It combines the benefits of bidirectional context from models like BERT with the autoregressive generation capabilities of models like GPT (Radford et al., 2018).

T5[8] is a Transformed-based model designed to treat any NLP problem as a text-to-text problem.

For fine-tuning both models, we determined the optimal hyperparameters through experimentation. The batch size was set to 8 for BART and 3 for T5, balancing computational resource constraints with model performance. We used a learning rate of 5e-5 and trained each model for 3 epochs.

5 Experimental protocol

To evaluate the distribution similarity between courts and the content similarity of sections across different courts, we conducted several experiments with various data configurations. For each of our tasks, we aimed to assess different systems across the two lower courts and SCOTUS.

We first trained each model using data from a single court and then assessed its performance on data from another court. Next, we explored the effect of combined training by using data from both lower courts to analyze potential performance variations. This approach allows us to understand how well the models generalize from one court to another and to determine if combining data sources improves the robustness of the results.

Table 2 details the different data configurations used for heading prediction and generation tasks across the three types of courts.

For lower courts, our dataset for heading generation consists only of unique headings—headings that appear only once. This strategy aims to capture a broader diversity of specific heading types while minimizing potential bias from repeated headings. For heading prediction, we selected 1,000 examples for each generic heading in our list (See Appendix A). This selection is based on the fact that each generic heading appears at least 1,000 times, ensuring sufficient representation in the dataset. We divided each dataset into training, validation, and testing sets in the ratio of 8:1:1.

For SCOTUS, due to the limited number of available headings, the dataset has been reserved exclusively for testing. For specific headings, we selected only the unique headings, while for generic headings, we included all the generic headings found in the corpus.

Finally, we considered a configuration where data from the two lower courts are combined. In this configuration, the training and validation data for each type of heading consist of a balanced mix from both jurisdictions. The test sets for each court remain independent and are identical to those used for testing each court individually, allowing us to observe the impact of using multiple data sources on model performance.

	Task	Train	Validation	Test
Federal appellate	Prediction	6,400	800	800
	Generation	82,924	10,366	10,365
State Supreme	Prediction	6,400	800	800
	Generation	99,991	12,499	12,498
SCOTUS	Prediction	-	-	76
	Generation	-	-	818
Mixed courts	Prediction	6,400	800	-
	Generation	91,457	11,432	-

Table 2: Split of datasets into train, validation, and test sets for headings prediction and generation tasks across the three courts. Numbers denote the number of examples in each set.

To evaluate our models, we utilized both clas-

[5] https://huggingface.co/google-bert/bert-base-uncased

[6] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[7] https://huggingface.co/facebook/bart-base

[8] https://huggingface.co/google-t5/t5-base

sification metrics, such as Precision, Recall and F1-score, and text generation metrics, such as ROUGE (Lin, 2004) and **BERTScore** (Zhang et al., 2020). In particular, we utilized ROUGE-1 which corresponds to a unigram recall score and ROUGE-L which computes a F1 similarity score in terms of the Longest Common Subsequence. In comparison to ROUGE which uses lexical overlap as a proxy for measuring content similarity, BERTScore is intended to capture a deeper semantic similarity between headings.

6 Results and discussion

6.1 Predicting generic section headings

The evaluation results for predicting generic headings (See Table 3) demonstrate that, when the training and testing domains are aligned (a configuration not applicable to SCOTUS), the models demonstrate the highest performance, with F-scores consistently between 0.91 and 0.92, regardless of the target domain. LegalBERT exhibits the best results, while BERT achieves comparable outcomes, albeit slightly lower by a few hundredths. In contrast, FastText performs less effectively, with scores lower by several tenths.

When training and testing the models on different, non-overlapping domains, a comparable ranking of the models and a consistent trend in performance were observed. The precision score of the best system exhibited a slight decline of 5 to 6 hundredths, yet remained robust. The recall score followed a similar pattern, demonstrating values comparable to precision, except when SCOTUS was the target domain, where recall experienced a notable decline of approximately two-tenths.

Upon analyzing classification errors on SCOTUS, a notable confusion between the "Background" class and the "Introduction" class was observed, which explains the drop in recall. For instance, LegalBERT trained on state supreme court data, shows 10 errors out of 24 examples in the "Background" class, 7 of which are incorrectly classified as "Introduction". This confusion could be explained by the fact that "Background" sections often begin by introducing the case, making them similar to "Introduction" sections. However, due to the limited size of the SCOTUS dataset, this observation is specific to this particular dataset and cannot be generalized to all SCOTUS opinions.

In the configuration where training data consists of a balanced mix of both lower courts, the performance of the best model is either equivalent to or slightly lower than those trained on data fully aligned with the target domain.

The highest F-score on SCOTUS remains with a LegalBERT model trained on the State Supreme source. However, the small size of the SCOTUS test dataset, consisting of only 76 instances, limits the ability to draw definitive conclusions about the performance differences observed.

6.2 Generating specific section headings

The evaluation results for generating specific headings (See Table 4) show that the ROUGE scores obtained for BART and T5 are quite similar, differing by only half a point across training sources when tested on the Federal Appellate and State Supreme datasets, with BART achieving the highest scores. However, this trend is reversed for the SCOTUS test set, where T5 achieves ROUGE scores nearly one point higher. Given the small differences between these scores, it is irrelevant to try to distinguish between them. Regarding BERTScore, the difference between BART and T5 is negligible across all target corpora, with scores varying from 86.31 to 87.31.

Notably, when the domain of the test data is aligned with the domain of the training data, the highest ROUGE scores are observed. However, there is no significant variation in BERTScore regardless of the domain.

ROUGE scores decline as the divergence between the source and target domains increases. When trained on the State Supreme dataset and tested on the Federal Appellate dataset, BART experiences a reduction of approximately 1 point, while T5 exhibits a slightly smaller decline. In contrast, both models exhibit a more significant drop of 4 points when the reverse configuration is applied. The most substantial reductions occur when the target domain is SCOTUS, with decreases ranging from 3 to 6 points. Notably, a model trained on the Federal Appellate data achieves ROUGE scores approximately one point higher than a model trained on State Supreme data. Both training sources exhibit comparable semantic similarity to SCOTUS, though Federal Appellate demonstrates a slight advantage in lexical alignment.

When trained on the mixed dataset, the ROUGE scores are marginally lower, by less than one point, than those obtained with models trained on data aligned with the target domain. On the SCOTUS test set, the mixed-data models produce ROUGE

			Test Court								
			Federal Appellate			State Supreme			SCOTUS		
			P	R	F1	P	R	F1	P	R	F1
Train Court	**Federal Appellate**	FastText	0.55	0.61	0.56	0.42	0.45	0.42	0.67	0.39	0.39
		BERT	0.90	0.90	0.90	0.80	0.80	0.80	**0.89**	0.63	0.73
		LegalBERT	**0.91**	**0.91**	**0.91**	0.85	0.85	0.85	0.84	0.59	0.68
	State Supreme	FastText	0.43	0.51	0.46	0.40	0.49	0.43	0.581	0.33	0.36
		BERT	0.85	0.84	0.84	0.91	0.91	0.91	0.82	0.63	0.70
		LegalBERT	0.87	0.86	0.86	**0.92**	**0.92**	**0.92**	0.87	**0.71**	**0.76**
	Mixed courts	FastText	0.51	0.58	0.53	0.46	0.54	0.49	0.65	0.47	0.5
		BERT	0.90	0.90	0.90	0.88	0.87	0.87	**0.89**	0.63	0.73
		LegalBERT	**0.91**	**0.91**	**0.91**	0.90	0.89	0.89	0.86	0.64	0.73

Table 3: Performance of different models in predicting generic headings across courts. The rows indicate the court used for training, while the columns represent the court used for testing. Each cell displays the performance score for a given model under a specific training and testing court combination. Values in bold highlight the best performance for each test court configuration.

			Test Court								
			Federal Appellate			State Supreme			SCOTUS		
			R1	RL	BERT	R1	RL	BERT	R1	RL	BERT
Train Court	**Federal Appellate**	BART	36.98	35.65	87.31	32.98	31.45	86.44	30.32	29.18	86.52
		T5	36.35	35.01	87.16	32.41	30.90	86.33	31.41	30.53	86.57
	State Supreme	BART	34.09	32.84	86.75	35.12	33.46	86.97	29.38	28.18	86.41
		T5	33.54	32.19	86.63	33.89	32.34	86.63	30.02	29.05	86.31
	Mixed Courts	BART	36.12	34.78	87.14	34.63	33.07	86.89	30.17	29.00	86.59
		T5	35.50	34.17	87.03	33.55	32.04	86.52	31.32	30.27	86.48

Table 4: Performance of different models in generating specific headings across courts. The rows indicate the court used for training, while the columns represent the court used for testing. Each cell displays the performance score for a given model under a specific training and testing court combination.

scores comparable to those of the best models trained exclusively on Federal Appellate, with a difference of less than two-tenths of a point. This suggests that training on a mixed corpus does not improve model robustness in this context.

Globally, the T5 model trained on Federal Appellate data appears to be the best solution for the task of generating titles in the SCOTUS corpus, without making a big difference to the BART model.

7 Human evaluation

To further evaluate our generic heading prediction and specific heading generation approaches, we built a new corpus by asking two human experts to annotate sections of a selection of opinions extracted from SCOTUS. These opinions were selected in a previous work (Lavissière and Bonnard, 2024) to represent the authors and the distribution of subjects in the SCOTUS corpus from 1945 to 2020. 51 opinions were fully annotated, comprising a total of 283 sections. The experts were law students who had completed their third year of undergraduate studies. They had the advantage of being familiar with the documents, having spent two months and half annotating their rhetorical structure at the sentence level. They received no training, apart from the definition of the various generic titles and the meaning of the labels in our rating scale for specific titles.

The annotators were tasked with selecting a generic title for each section and evaluating two specific title proposals generated by the BART and T5 models, both fine-tuned on a mixed dataset. The specific titles were rated on a 4-point scale. The annotators were also asked to propose an alternative specific title, independent of the assigned ratings. In terms of generic headings, this new dataset is almost 4 times larger than the one we were able to build with the native SCOTUS headings.

The Cohen's kappa score (Cohen, 1960) for annotating the sections with generic headings was 0.81, indicating substantial agreement between the evaluators. For the rating of the specific headings, the intra-class correlation coefficient (ICC) (Weir, 2005) was 0.467 for those generated by BART and 0.548 for those generated by T5, reflecting a moderate level of agreement between the evaluators.

7.1 Evaluation of generic headings

LegalBERT trained on mixed court data achieved an F1 score of 0.72 on this new SCOTUS data against 0.73 on the native SCOTUS dataset, highlighting the effectiveness of our approach in this context. The most frequently used classes by the annotators were "Analysis," "Conclusion," and "Background," accounting for 97.53% of the annotations, whereas other classes were rarely used in annotating SCOTUS sections.

7.2 Evaluation of specific headings

The human evaluation of specific headings (See Figure 2) shows that annotators are in complete agreement with the proposed headings in 43% of the cases for both BART and T5, indicating that these headings accurately summarize the content of the sections. In 27% of cases for BART and 29% for T5, the headings are considered reasonably appropriate, covering most relevant aspects of the text but still requiring improvements. Conversely, in 19% of cases for BART and 16% for T5, annotators find that the headings do not align well with the content, though they contain some relevant elements. Finally, in 11% of the cases for both models, annotators disagree entirely with the proposed headings, suggesting that these headings do not reflect the section content at all. These results indicate that BART and T5 exhibit similar performance in generating specific headings. They also confirm that using lower court data for training the models has effectively contributed to generating relevant headings for SCOTUS sections.

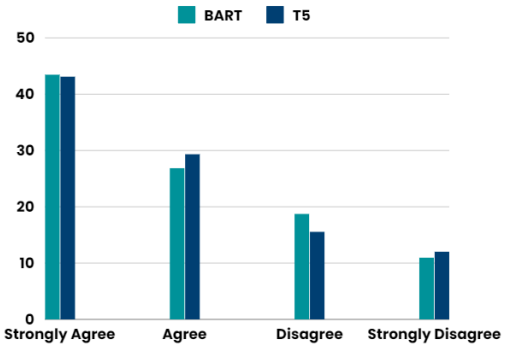

Figure 2: Distribution of human evaluations for specific headings generated by BART and T5.

7.3 Human and automatic evaluation

We calculated ROUGE-1, ROUGE-L, and BERTScore for the generated headings and those proposed by the evaluators to analyze the variation in automatic metrics based on human ratings. The results (See Figure 3) reveal that automatic scores are significantly higher when evaluators agree with the headings. For instance, headings rated as fully adequate achieve higher ROUGE-1 and ROUGE-L scores, reaching 45.67 and 43.17 for BART, and 48.22 and 46.07 for T5, respectively. BERTScore is also higher, with values of 87.13 for BART and 87.84 for T5 in cases of total agreement. In contrast, headings rated as somewhat or entirely disagreeable show much lower ROUGE and BERTScore, with minimum values of 8.16 for ROUGE-1 and 17.03 for BERT in cases of complete disagreement. These results highlight the importance of ROUGE and BERT metrics as indicators of quality for generated headings.

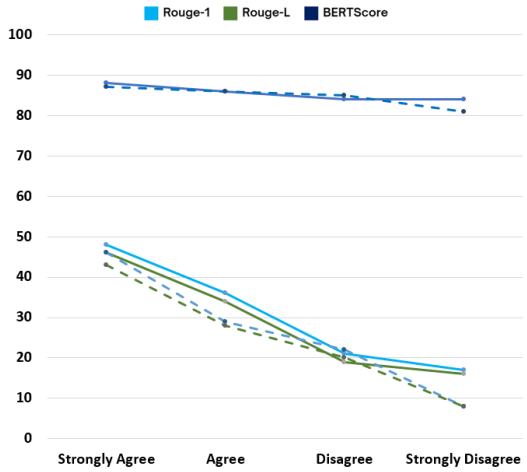

Figure 3: Variation in ROUGE-1, ROUGE-L, and BERTScore between headings generated by BART (dashed lines) and T5 (solid lines) and headings proposed by annotators, based on annotator ratings.

8 Conclusion and future works

Our research demonstrates the viability of leveraging lower court data to predict and generate section headings for SCOTUS opinions. Despite the limited availability of SCOTUS-specific training data, models trained on federal appellate and state supreme court data generalize well, achieving strong results in both generic heading prediction and specific heading generation. LegalBERT produced the best results for predicting generic headings, while BART and T5 delivered a comparable performance in generating specific headings. Future work will focus on determining whether sections are better titled with generic or specific headings and addressing the challenge of automatic

document segmentation to further enhance legal document structuring while considering the task of paragraph heading.

9 Limitations

This work was specifically aimed at the automatic titling of SCOTUS sections because of the economic, political and societal stakes involved in understanding these documents in a context of globalisation. The transfer and generalisation of the approach to other documents was not our primary objective. Nevertheless, the present study has enabled us to make progress in our knowledge of the linguistic characteristics of several American courts of justice and will provide a better understanding of their processing. The full CourtListener database contains nearly 10 million[9] legal opinions from federal, state, and specialty courts and the legal domain in many countries is characterised by the length of its documents, whether they be patents, financial or court rulings (Sharma et al., 2019; Kornilova and Eidelman, 2019; Shukla et al., 2022).

The study presented here is a first step: the automatic hierarchical structuring of sections (or the segmentation of a text into homogeneous chunks) and the decision to prefer a generic title to a specific title to describe a section are future steps that will be necessary if our work is to be used effectively.

With this in mind, a future human evaluation will have to consider productivity gains as an indicator, rather than the quality of the summaries generated.

As a matter of fact, we could have used the latest generative models to carry out the task of generating specific headings. When choosing which generative models to use, it was important for us to weigh up the carbon cost of the models against the acceptability of the results to the experts. We share our position in Section 2. To clarify our position, in the first phase of our research, we wanted to determine whether the results produced by a "simple" finely-tuned model with a few hundred million parameters were acceptable by an expert. It was with this in mind that we formulated the points of our rating scale in our human evaluation. Instead of a numerical value, we chose values that expressed degrees of acceptability and that led us to take a position, namely strongly agree, tend to agree, tend to disagree, strongly disagree. We show in Section 7.2 that annotators agree or strongly agree with about 71% (43% + 28%) of the generated headings. This score shows indeed that there is room for improvement. In order to obtain the best results, we should compare the use of a few-shot and the fine-tuning of different recent LLMs. We are currently considered this point but we do not see it as a research question.

[9] https://www.courtlistener.com/faq/#explain-neutral-citations

10 Ethics

BERT, LegalBERT, BERTScore, BART, T5 required the use of GPUs. Experiments were run on a single GPU RTX 2080 Ti with 12G VRAM. These models include only a few hundred million parameters, but we have not accurately estimated the carbon impact of our research.

All published US Legal case law are freely available to the public and can be used to support the academic research and the development of legal technology.

The persons involved in the manual annotation task were remunerated at the going rate for their professional status.

Acknowledgments

This research was funded, in whole or in part, by the Agence Nationale de la Recherche (ANR), grant ANR-22-CE38-0004. A CC BY license is applied to the AAM arising from this submission, in accordance with the grant's open access conditions.

References

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

Abdessalam Bouchekif, Géraldine Damnati, Nathalie Camelin, Yannick Estève, and Delphine Charlet. 2015. Segmentation et titrage automatique de journaux télévisés. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 221–227, Caen, France. ATALA.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Francine Chen and Yan-Ying Chen. 2019. Adversarial domain adaptation using artificial titles for abstractive title generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics, pages 2197–2203, Florence, Italy. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Anjalie Field, Sascha Rothe, Simon Baumgartner, Cong Yu, and Abe Ittycheriah. 2020. A generative approach to titling and clustering Wikipedia sections. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 79–87, Online. Association for Computational Linguistics.

Mareike Florax and Rolf Ploetzner. 2010. What contributes to the split-attention effect? the role of text segmentation, picture labelling, and spatial proximity. *Learning and Instruction*, 20(3):216–224.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

S. Goźdź-Roszkowski. 2024. *Language and Legal Judgments: Evaluation and Argument in Judicial Discourse*. ISSN. Taylor & Francis.

Kango Iwama and Yoshinobu Kano. 2019. Multiple news headlines generation using page metadata. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 101–105, Tokyo, Japan. Association for Computational Linguistics.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

J. Kim. 2015. *American Law 101: An Easy Primer on the U.S. Legal System*. American Bar Association.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Mary C Lavissière and Warren Bonnard. 2024. Who's really got the right moves? Analyzing recommendations for writing American judicial opinions. *Languages*, 9(4):119.

Florian Le Bronnec, Song Duong, Mathieu Ravaut, Alexandre Allauzen, Nancy Chen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. 2024. LOCOST: State-space models for long document abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1144–1159, St. Julian's, Malta. Association for Computational Linguistics.

Julie Lemarié, Hélène Eyrolle, and Jean-Marie Cellier. 2008. The segmented presentation of visually structured texts: Effects on text comprehension. *Computers in Human Behavior*, 24(3):888–902. Instructional Support for Enhancing Students' Information Problem Solving Ability.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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, Barcelona, Spain. Association for Computational Linguistics.

Amin Omidvar and Aijun An. 2023. Learning to generate popular headlines. *IEEE Access*, 11:60904–60914.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

John M Swales. 2004. *Research genres: Explorations and applications*. Cambridge University Press.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Joseph P Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240.

David S. Weiss. 1983. The effects of text segmentation on children's reading comprehension. *Discourse Processes*, 6(1):77–89.

Jennifer Wiley and Keith Rayner. 2000. Effects of titles on the processing of text and lexically ambiguous words: Evidence from eye movements. *Memory & Cognition*, 28(6):1011–1021.

Huihui Xu and Kevin Ashley. 2023. Argumentative segmentation enhancement for legal summarization. In *Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023)*, pages 141–150, Braga, Portugal. CEUR Workshop Proceedings.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Preprint*, arXiv:2106.11520.

Fan Zhang, Itay Laish, Ayelet Benjamini, and Amir Feder. 2022. Section classification in clinical notes with multi-task transformers. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 54–59, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Generic headings

Table 5 presents the list of common frequent headings between federal appellate and state supreme courts ("Headings") and the final list of generic headings we considered (See the "Final Label" column). The selection of headings was based on those extracted from SCOTUS decisions and the literature (Goźdź-Roszkowski, 2024) which provided us with a basis of expertise.

According to the literature, Part I of the opinions typically outlines the facts of the case and the historical background. To represent this description, we selected headings such as "Fact," "Background," "Facts and Procedural History," "Factual and Procedural Background," and "Factual Background." Since these headings are semantically very similar, we grouped them under a single generic label, "Background," to avoid any confusion for the model, which might struggle to differentiate between such closely related headings. Similarly, Part III, dedicated to the analysis of the case, corresponds to the headings "Analysis" and "Discussion." Following the same reasoning, we grouped these two headings under the label "Analysis." We did not select the heading "Procedural History," as the literature indicates that this section is included in the "Syllabus." The other headings were selected due to their presence in SCOTUS opinions.

It is important to note that, according to the literature, Part II is generally dedicated to the interpretation of relevant laws. However, we found no

Heading	In SCOTUS	Final Label	Justification
Introduction	✓	Introduction	Found in SCOTUS
Facts	✓	Background	*Found in SCOTUS "Part I usually recounts facts of the case constituting the historical dimension of a case. " (Goźdź-Roszkowski, 2024)*
Background	✓		
Facts and procedural history			
Factual and procedural background			
Factual background			
Decision			
Discussion		Analysis	*"Part III analyzes argumentation advanced by the parties"*
Analysis			
Jurisdiction	✓	Jurisdiction	Found in SCOTUS
Issues	✓	Issues	Found in SCOTUS
Standard of review	✓	Standard of review	Found in SCOTUS
Procedural history	✓		*It's part of the syllabus: "The first two parts combined seem to correspond to what is known as 'procedural history' in common law jurisdictions and which is placed in the syllabus of a US Supreme Court opinion." (Goźdź-Roszkowski, 2024)*
Sufficiency of the evidence	✓	Sufficiency of the evidence	Found in SCOTUS
Conclusion	✓	Conclusion	Found in SCOTUS

Table 5: Generic Headings Covered in the Paper.

recurring headings in our list or in the general corpus that correspond to this description. Therefore, we assumed that this section is part of the "Analysis" and could be detailed with a specific heading depending on its content.

B Document structuring and section extraction

The SCOTUS dataset comprises partially structured documents, wherein the identification of sections and headings is frequently challenging. To address this, we developed a rule-based section extractor that identifies and extracts section markers present in the text. This extractor unifies different document types (PDF, plain text, and HTML) into a well-structured HTML file, ensuring all existing headings are systematically captured.

A section marker can take various forms: it can be a numbered bullet (Roman numerals, Arabic numerals, alphabetical numbering, etc.), a bullet followed by a heading (e.g. "I. Background"), or simply a stand-alone heading (e.g. "Background").

The extractor first identifies the beginning and end of each opinion within a document, whether it is the majority opinion, a dissenting opinion, or a concurring opinion. These opinions are typically marked with a judicial title that indicates the beginning of the opinion, such as "Justice X delivered the opinion of the Court", "X, Chief Justice," or "MR. JUSTICE X, dissenting," where X is the justice's name. This segmentation makes it easier to identify the different sections within each opinion.

Next, the extractor scans each opinion line by line, regardless of document type, looking for lines that begin with bullet points. For each bullet type found, it assigns an appropriate hierarchical level using HTML heading tags according to their appearance in the document, thereby preserving the sequential integrity of the numbering and ensuring that the logical order of numbers is maintained, so that, for example, a "(c)" cannot occur without "(a)" and "(b)".

Once all section markers are detected, the extractor collects the headings present in all documents and creates a whitelist of headings to identify those that do not begin with bullets, but still mark the beginning of a section or subsection. This whitelist improves the accuracy of the headings extraction, ensuring that documents are properly structured and all sections are identified.

Table 6 shows the Recall, Precision, and F1-Score results from evaluating our section extractor on a manually annotated test set of 100 opinions from three different US courts, amounting to a total

of 300 opinions. The evaluation was conducted at two levels: the recognition of different opinions within a document (referred to here as "Opinions") and the recognition of sections and subsections within these opinions (referred to here as "Sections"). In the context of Section recognition, the detection of headings is also implied, given that headings are included in section markers.

		Opinions	**Sections**
SCOTUS	Precision	0.93	0.99
	Recall	0.96	0.98
	F-1 Score	0.94	0.98
Federal Appellate	Precision	0.92	0.98
	Recall	0.9	0.96
	F-1 Score	0.91	0.97
State Supreme	Precision	0.84	0.98
	Recall	0.79	0.88
	F-1 Score	0.81	0.93

Table 6: Evaluation Results of the Extractor."Opinions" refers to the evaluation of recognizing different opinions within a document, while "sections" refers to the evaluation of the recognition of sections and subsections within the opinions.

These results demonstrate that our extractor effectively structured the documents by accurately identifying the different opinions within each document and recognizing the sections and subsections of SCOTUS opinions as well as other opinions that are similar to those of SCOTUS.

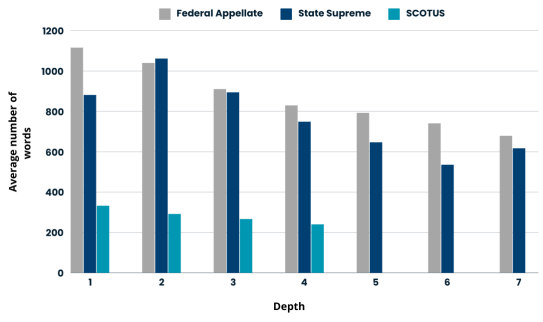

Figure 4: Average word count distribution per section depth across the three selected courts.