Argumentative Segmentation Enhancement for Legal Summarization

Huihui Xu[1,3,*], Kevin Ashley[1,2,3]

[1]*Intelligent Systems Program, University of Pittsburgh, PA, USA*

[2]*School of Law, University of Pittsburgh, PA, USA*

[3]*Learning Research and Development Center, University of Pittsburgh, PA, USA*

Abstract

We use the combination of argumentative zoning [1] and a legal argumentative scheme to create legal argumentative segments. Based on the argumentative segmentation, we propose a novel task of classifying argumentative segments of legal case decisions. GPT-3.5 is used to generate summaries based on argumentative segments. In terms of automatic evaluation metrics, our method generates higher quality argumentative summaries while leaving out less relevant context as compared to GPT-4 and non-GPT models.

Keywords

legal summarization, natural language processing, neural networks, argument mining

1. Introduction

Automatic text summarization is a process of automatically generating shorter texts that convey important information in the original documents [2]. There are in general two different approaches for automatic text summarization: extractive summarization and abstractive summarization [3]. Extractive summarization can be conceptualized as a sentence classification task, where the algorithm selects important sentences from the original document directly [4]. Abstractive summarization can be a more natural way of summarizing in terms of novel words and expressions [5]. Authors of [6, 7] have experimented with several extractive summarization methods in domains like law and science.

Abstractive summarization is flourishing in recent years because of the rise of large pre-trained language models, like BART [8], T5 [9], and Longformer [10]. However, those models still require sizable training datasets to tackle a new task. For example, a language model trained on a Wikipedia text corpus requires fine-tuning on a legal dataset. In addition, unlike news articles, legal case decisions are longer and contain argumentative structures [11]. While some summarization approaches are beginning to take the argumentative structure of a legal case decision into account (e.g., [11]), none do so in a zero-shot setting.

In this paper, we conduct a study of summarizing argumentative segments extracted from a legal document using the latest GPT-3.5 model (text-davinci-003) and

✉ huihui.xu@pitt.edu (H. Xu); ashley@pitt.edu (K. Ashley)

GPT-4 [12] model. The new GPT-3.5 version is based on InstructGPT [13], which is also trained with reinforcement learning from human feedback (RLHF). Despite its potential for generating high quality summaries, the GPT-3.5 model has a 4,097-token input limitation. This is a disadvantage for summarizing long legal documents. Our work employs a method of cutting the long documents into shorter segments while still preserving argumentative components. GPT-4 is also trained with RLHF like GPT-3.5 but with more capability. For example, GPT-4 can handle 8,192[1] tokens as input, which has doubled GPT-3.5's input length. Even though GPT-4 can handle longer documents, there are still some legal documents that exceed the input limitation. Besides, we believe that legal argument mining and argumentative zoning can extract argumentative segments that can help models to generate better legal summaries.

In order to extract argumentative segments from legal decisions, we propose a novel task for automatically classifying segments as argumentative or non-argumentative segments. This task stems from Argumentative Zoning (AZ) addressed in [14, 1]. Teufel et al. define the task of AZ as a sentence level classification with mutually exclusive categories given an annotation scheme. AZ divides a paper into zones on the basis of the content knowledge claim in the corresponding segment [1]. We adopt the reasoning behind AZ and divide textual segments into argumentative or non-argumentative segments by examining if any argumentative sentences exist in the corresponding segment. The identified argumentative segments are then fed into the model for generating summaries.

Figure 1 illustrates the summarization pipeline of our approach. The pipeline comprises three stages. First, the

---

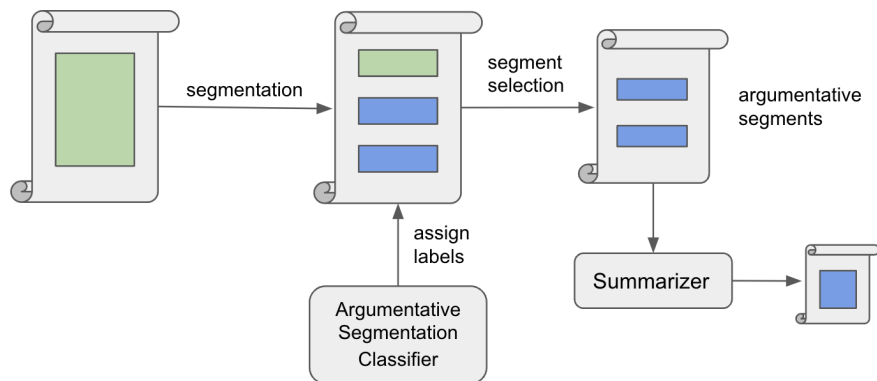[1]There is another version of the model that supports 32,768 tokens.

Figure 1: Pipeline of generating argumentative segmentation enhanced legal summarization.

document, a full-text legal opinion is segmented into several parts. Then, every segment is assigned a label based on the existence of argumentative sentences using a classifier. Finally, the predicted argumentative segments are fed into the model. The model will summarize each segment and concatenate them as the final summary for the decision.

Our contributions in this work are: (1) We propose a novel task of predicting argumentative segments in the legal context. (2) We show that our approach for using argumentative segments to guide summarizing is effective. (3) We overcome the token limitation of GPT-3.5 when applied to long document summarization and show a promising result in a zero-shot setting.

2. Related Work

2.1. Argumentative Zoning

Teufel, et al. [14] first proposed and defined the task of AZ as a sentence level classification with mutually exclusively categories given a certain annotation scheme for scientific papers. The earliest scheme includes seven categories of zones, such as Aim and Background. The annotation scheme is based on the rhetoric roles employed by authors. For example, one can identify sections that cover the background of the scientific research in a technical paper among other sections. Later, [1] made attempts toward discipline-independent argumentative zoning in two different domains. The idea of AZ is seeking to extract the structure of research components that follows authors' knowledge claims. As a result, there are different AZ schemes for different domains, such as [15] for

chemistry research articles, [16] for physical sciences and engineering and life and health sciences. AZ was later adopted for legal documents in [17, 18]. Since AZ classified sentences into different categories, it is helpful for generating summaries for long documents. [19] proposed a tool for AZ annotation and summarization. However, AZ annotation for legal documents can be expensive. We propose to leverage our sentence level annotation for AZ in the context of argumentative segmentation classification.

2.2. Legal Argument Mining

Legal Argument mining aims to extract legal argumentative components from legal documents. Most argument mining work consists of three sub-tasks: identifying argumentative units, classifying the roles of the argumentative units, and detecting the relationship between the argumentative units. [20] explored the argumentative characteristics of legal documents.[21, 22] identified rhetorical roles that sentences play in a legal context. Early work in legal argument mining rely on word patterns and syntactic features [23, 24]. Recently, contextual embedding has been used for legal argument mining [25, 26], like Sentence-BERT [27] and LegalBERT [28] embedding. [25, 26] have proposed a legal argument triples scheme to classify sentences for summarizing legal opinions in terms of Issues, Reasons, and Conclusions.

2.3. Summarization Methods and GPTs

As noted, the automatic summarization methods can be categorized as extractive or abstractive. Most ML ap-

proaches for learning to extract sentences for summarizing documents are unsupervised [29, 30]. They are based on learning sentence importance scores for selecting sentences to form summaries. The development of better sentence representations, like Sentence-BERT, has lead to improvements in generating better summaries [31].

Recent research applying sequence-to-sequence neural models to summarization is gaining more attention. [32] proposed a pointer generator architecture for generating higher quality abstractive summaries. Transformer-based sequence-to-sequence models, like BART (Bidirectional and Auto-Regressive Transformer), T5 (Text-to-Text Transfer Transformer) and Longformer, have been used in generating abstractive summaries. [11] incorporate legal argumentative structures into sequence-to-sequence model to further enhance the quality of summaries. In this work, Longformer Encoder-Decoder (LED), T5 and BART serve as the baseline for our experiments.

The mainstream transformer-based models, however, require a curated training set to adapt to a new domain. The success of prompt-based models provides a new way of solving the domain adaption problem by learning from a large unlabelled dataset. GPT-3.5 and GPT-4, developed by OpenAI, are the examples of prompt-based models. [33] investigated how zero-shot learning with GPT-3 compares with fine-tuned models on news summarization task. Their results show that GPT-3 summaries are preferred by humans. Our work focuses instead on legal summarization and takes argumentative structure into account. The results show a higher performance in terms of automatic evaluation metrics by taking account of argument structures. We further experimented with GPT-4 on legal summarization, since it has a larger context window compared to GPT-3.5. Our findings demonstrate that considering the argumentative structure leads to improved summaries.

3. Legal Decision Summarization Dataset

We use the legal decision summarization dataset provided by the Canadian Legal Information Institute (CanLII).[2] The summaries are prepared by attorneys, members of legal societies, or law students. The basic statistics of the annotated dataset are listed in Table 1. The court decisions involve a wide variety of legal claims. The average length of the court decisions is 4,382 tokens. It exceeds the token limitation of GPT-3.5 (4,097 tokens). This motivates us to explore argumentative segmentation to reduce the input document length.

In prior work, researchers conceptualized a type sys-

tem for annotating sentences in legal case decisions and summaries, which includes: **Issue** – Legal question which a court addressed in the case; **Conclusion** – Court's decision for the corresponding issue; **Reason** – Sentences that elaborate on why the court reached the Conclusion [34]. Those sentences are referred to as IRC triples. We have accumulated 1,049 annotated legal case decision and summary pairs. [11, 6] use the same dataset for legal summarization tasks. [11] use the IRC annotations as markers to inform models with argumentative information. [6] explored the structure of legal decisions and used the annotated dataset as the basis for domain-specific evaluation of summaries.

In this work, we use the idea of argumentative zoning to further expand the use of IRC triples. The documents in the dataset have already been split at a sentence level. They have not yet been split into paragraphs or annotated in terms of explicit rhetorical zones. We adopt C99 [35], a domain-independent linear text segmentation algorithm, to further segment the legal case decisions on a higher level. This algorithm measures the similarity between all sentence pairs to generate a similarity matrix. The similarity between a pair of sentences x, y is calculated using cosine similarity. Sentence-BERT is used for representing all sentences in the same space before computing the similarity scores. Then we cluster the neighboring sentences into groups based on the similarity scores.

Here, we propose a novel task – argumentative segmentation classification. For each group of sentences, we assign an "argumentative segment (1)" if there exists one or more IRC sentences, or a "non-argumentative segment (0)" otherwise. This combines the idea of argumentative zoning with semantic segmentation. Table 2 shows an example of an argumentative segment. As the example shows, segment no. 9 is labeled as an argumentative segment because of the existence of a conclusion sentence.

We split our data into 80% training, 10% validation and 10% test datasets.

4. Experiments and Results

4.1. Argumentative Segment Classification

Every legal case decision in our dataset has been split into segments using the C99 algorithm. Table 3 shows the results of C99 segmentation. From the table, the average number of argumentative segments is 6 in a legal decision while the number of non-argumentative segments is 59. Thus, the number of argumentative segments is much less than non-argumentative segments in legal decisions. We performed a segment-level classification using the mentioned data split. We conducted experiments with different transformer models, BERT [36] and

Table 1

Statistics of CanLII annotated court decision and summary pairs.

Type	Avg. # of tokens	Max. # of tokens	Min. # of tokens
Court decision	4382	62785	122
Human-written summary	273	2072	17

Table 2

An example of segmented legal case decision from the annotated dataset. The text with gray background is an annotated conclusion. Since it appears in segment no.9, the segment is labelled as argumentative.

..	..
segment no.8 (non-argumentative segment)	III As matter of public policy, the Crown is not required to disclose the name of the confidential informer. If the Information discloses too much information about the informer and his means of knowledge, the identity of the informer will become apparent. As result, the Crown has to take refuge in the kind of language employed in this Information.
segment no.9 (argumentative segment)	note the type of language used by the peace officer has been accepted, as compliance with the section, in other cases: see Re Lubell and The Queen (1973), 1973 CanLII 1488 (ON SC), 11 C.C.C. (2d) 188 (Ont. H.C.); Re Dodge and The Queen (1985), 1984 CanLII 59 (NL SC), 16 C.C.C. (3d) 385 (Nfl. S.C.). Perhaps more information could have been provided, however, there was information upon which the respondent, acting judicially, could be satisfied that search warrant should issue. Courts should not be too technical when scrutinizing the Information in support of search warrant; substantial compliance with s. 443 is sufficient.

LegalBERT[28]. We use those models to predict the argumentativeness of segments (i.e., argumentative segment, or non-argumentative segment). Figure 2 shows the results of the binary classification. The figure shows Legal-BERT achieved a better classification result compared to BERT. LegalBERT achieved 80.14% $F1$ score while BERT has 78.24%. As a result, we chose to use LegalBERT's predictions to select input segments for the following summarization task.
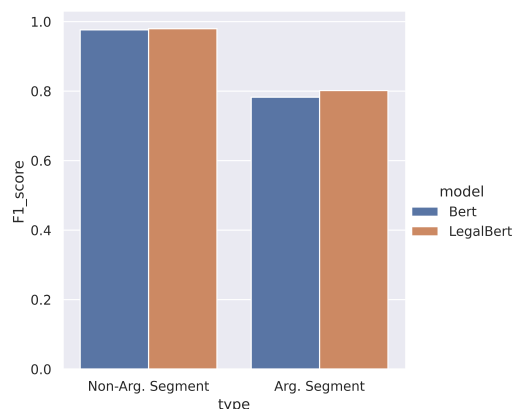
Figure 2: Argumentative segmentation prediction results.

4.2. Baselines

We use two different types of baselines for our proposed argumentative segmentation enhanced summarization method. One is non-GPT abstractive summarzation model, like LED, T5, and BART. The other one is vanilla GPT-3.5 and GPT-4. They are both developed by OpenAI.

The GPT-3.5 model is an auto-regressive language model. This model can generate high quality news summaries in a zero-shot setting according to [33]. We used the latest version, text-davinci-003, in our work just released in November 2022. There is little or no work, however, measuring how well the model performs on legal documents. GPT-4 is a multi-modal large language model, which is more capable than GPT-3.5. GPT-4 was released in March 2023, and it is by far the most advanced large language model in the field.

4.3. Prompting for GPT-3.5 and GPT-4

As mentioned, GPT-3.5 and GPT-4 are both prompt-based model. In order to use GPT-3.5 and GPT-4 to summarize a chunk of text, we have to inform the model of the type of task to perform. In our experiment, we add a short text "TL;DR" immediately after the input text. "TL;DR" is an abbreviation for "Too Long; Don't Read", and \n is the change of a new line. "TL;DR" instructs GPT-3.5 and GPT-4 to summarize the text in a fewer number of words. The example prompt is listed below:

$$\{\{Text\}\} + \backslash nTL;DR \tag{1}$$

Table 3

Statistics of text segmentation on court decisions using C99 with Sentence-BERT('bert-base-nli-stsb-mean-tokens') embedding.

Type	Avg. # of segments	Max. # of segments	Min. # of segments
Argumentative segmentation	6	30	1
Non-argumentative segmentation	59	732	0
Total	65	737	2

We only need to control the max output tokens and temperature without fine-tuning on our dataset. This is a zero-shot setting because the model does not see any human-written summaries before generating summaries. We noticed that the lengths of generated summaries are consistent. The average lengths of model-generated summaries are reported in Table 4, Table 5 and Table 6.

For the baseline GPT-3.5 model, we chunk the original document into lengths which the model accepts. We tried different lengths, and finally settled on 2,500 tokens to avoid an "over token request limitation error." The argumentative segmentation enhanced GPT-3.5 model does not have this problem because the argumentative segments are shorter than GPT-3.5's token limitation. It also helps GPT-3.5 to focus on the chunks of text that have important argument-related information. Even though GPT-4 has much longer context length, it still falls short for dealing with some long documents. We set 7,500 tokens as the limit of prompt length to avoid "over token request limitation error."

4.4. Results

Rouge-1, Rouge-2, Rouge-L, BLEU, METEOR, and BERTScore are used to measure the performance. Rouge stands for Recall-oriented Understudy for Gisting Evaluation [37]. Rouge-based evaluation metrics examine lexical overlap between generated and reference summaries. BLEU stands for Bilingual Evaluation Understudy [38]; it measures word overlap taking order into account. It is often used to measure the quality of machine translation. METEOR [39] computes the similarity between generated and reference sentences by mapping unigrams. BERTScore [40] uses contextual token embedding to compute similarity scores between generated and reference summaries on a token level.

Table 4 shows the test set results of different summarization models in different experimental settings. We first experimented with those non-GPT models in a zero-shot setting, and the results are shown in parentheses. Since zero-shot performance is not good, we further fine-tuned those models on the training set. We adopt some of the training hyperparameters from [11]: initializing LED and BART with learning rate of $2e^{-5}$, T5 with learning rate of $1e^{-4}$; and training both models for 10 epochs; set maximum input length is 6144 words for LED and T5

and 1024 for BART; maximum output length is set to 512 tokens for all the models. LED, T5 and BART outperform baseline GPT-3.5 and GPT-4 in term of automatic evaluation metrics. We also find that LED, T5 and BART produce longer summaries than GPT-3.5 and GPT-4 on average, which might directly contribute to the higher scores across some of the metrics.

Table 5 shows different combinations of two important control parameters in GPT-3.5: temperature and max_tokens. According to the official website,[3] temperature ranges between 0 and 1 and controls the randomness of generated text. With a 0 temperature, GPT-3.5 will select the most deterministic response, while a 1 temperature is the most random. max_tokens parameter controls the number of generated tokens. We found that the model generally performs better at a lower temperature. For example, when the max_tokens parameter is fixed at 128, the Rouge and BLEU scores decrease when the temperature rises from 0 to 0.8. We also notice that the max_tokens also affect the performance: when the temperature is set to 0, the model with 128 max_tokens achieves the best scores across all metrics except the BERTScore. We control GPT-4 with the same parameters, and the results are presented in Table 5.

Table 7 shows the comparison between a reference summary and GPT generated summary when the input does not exceed either the GPT-3.5 and GPT-4 token limitations. We observe that the generated summaries provide similar information regarding the case facts. However, the argumentative segmentation enhanced GPT-3 generated summary provides additional information about the judge's considerations.

Since GPT-3.5 imposes the token request limitation, any input text longer than the limit should be chunked before submitting to the model. In our test dataset, almost half of the cases exceed the token limitation. For these longer opinions, segmenting them using our implementation of argument zoning would seem to be a reasonable step, possibly increasing the likelihood that GPT-3.5's summaries would include useful argument-related information. Table 8 shows an example of generated summaries when the original case decision substantially exceeds GPT-3.5's token limit. As a result, we need to shorten the document first before feeding it to the model.

---

[3]https://platform.openai.com/playground/p/default-tldr-summary?model=text-davinci-003

Table 4

Summarization results on the test set. The numbers within parentheses represent zero-shot outcomes of the models; the numbers with no parentheses are results obtained after fine-tuning on the training set. Bolded numbers are the best results. We use GPT-3.5 ("text-davinci-003") for the experiments. More detailed experiment results are shown in Table 5. "*" represents the best result from Table 6. "**" represents the best result from the sets of experiments from Table 5.

Experiment	Model	Avg. length	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	BERTScore
Baseline	LED	393(387)	46.14(37.08)	21.84(12.33)	42.99(33.36)	14.28(5.95)	0.38(0.28)	85.98(81.49)
Baseline	T5	167(65)	45.30(21.39)	22.51(5.10)	42.11(19.47)	14.91(1.83)	0.28(0.10)	86.84(82.84)
Baseline	BART-large	365(286)	44.26(27.56)	18.52(8.89)	37.53(24.94)	10.82(3.80)	0.32(0.18)	84.85(76.55)
No Arg Seg.	GPT-3.5	126	40.88	18.90	37.63	10.47	0.23	86.56
No Arg Seg.	GPT-4*	144	43.46	19.84	30.17	11.13	0.26	86.51
Arg Seg.	GPT-3.5**	205	**49.42**	**23.98**	**46.07**	**17.54**	**0.32**	**87.30**

Table 5

GPT-3.5 model with different combinations of temperatures and max output tokens (temperature, max tokens).

Parameters	Avg. summary length	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	BERTScore
(0, 32)	128	41.60	19.05	38.76	10.88	0.23	86.02
(0.3, 32)	129	41.79	19.07	38.58	10.87	0.23	86.05
(0.5, 32)	129	41.55	18.94	38.47	10.38	0.23	86.06
(0.8, 32)	128	40.79	17.85	37.58	9.34	0.22	85.92
(0, 64)	184	47.78	22.95	44.51	15.77	0.30	87.03
(0.3, 64)	186	47.79	22.92	44.47	15.80	0.30	86.98
(0.5, 64)	183	46.93	21.80	43.50	14.55	0.29	87.00
(0.8, 64)	183	46.63	21.29	43.03	13.79	0.29	86.93
(0, 128)	205	49.42	23.98	46.07	17.54	0.32	87.30
(0.3, 128)	203	49.32	23.72	45.84	16.77	0.32	87.32
(0.5, 128)	200	48.84	23.29	45.23	16.22	0.32	87.25
(0.8, 128)	197	47.22	21.11	43.39	14.30	0.30	87.05
	Average		45.81	21.33	42.45	13.85	0.28	86.74

Meanwhile, GPT-4 can handle the length of the original case decision. We noticed that the baseline GPT-4 summary lacks some necessary details as compared to the argumentative segmentation enhanced approach. The latter included a more detailed presentation of the issue and conclusion and more of the reasons. The result was expected, since the input was shortened for the baseline. Despite the richness of information that a GPT-3.5 summary provides, GPT-4 generates smoother summaries. The main reason is that GPT-4 has a longer context span

Table 6

GPT-4 model with different combinations of temperatures and max output tokens (temperature, max tokens).

Parameters	Avg. summary length	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	BERTScore
(0, 128)	106	39.21	17.52	27.33	8.33	0.22	86.18
(0.3, 128)	109	40.57	18.41	28.38	8.66	0.23	86.30
(0.5, 128)	106	39.16	17.80	27.45	8.71	0.22	86.26
(0.8, 128)	110	39.00	16.75	27.25	8.07	0.22	86.18
(0, 256)	132	42.76	19.63	29.80	10.69	0.26	86.45
(0.3, 256)	134	42.64	19.21	29.36	10.10	0.26	86.43
(0.5, 256)	134	43.42	19.91	30.35	10.62	0.26	86.54
(0.8, 256)	135	42.58	18.84	29.06	9.57	0.26	86.39
(0, 512)	144	43.46	19.84	30.17	11.13	0.26	86.51
(0.3, 512)	137	43.07	19.58	29.96	10.43	0.26	86.55
(0.5, 512)	142	43.19	19.48	30.18	10.12	0.26	86.49
(0.8, 512)	140	42.15	18.50	29.38	9.99	0.25	86.45
	Average		41.77	18.79	29.06	9.70	0.25	86.39

Table 7

Argumentative segmentation enhanced GPT-3.5 and baseline GPT-4 generated summaries versus reference summary. `<Issue> </Issue>` , `<Reason> </Reason>` and `<Conclusion> </Conclusion>` are markers for respective type of sentence.

Reference summary (case name: 1994canlii3880(2,349 tokens in the original case decision))	Baseline GPT-4	Arg. Segmentation Enhanced GPT-3
The accused, a young offender, pleaded guilty to 16 counts of auto theft. `<Issue>` He was sentenced to 6 months open custody and 2 years probation. `</Issue>` Prior to being sentenced, he spent 13 days on remand and 2 months under house arrest under s.7.1 of the Young Offenders Act. `<Issue>` The Crown appealed the sentence. `</Issue>` `<Conclusion>` HELD: Appeal allowed. `</Conclusion>` `<Reason>` 1)The accused was 15 years of age and had no record. `</Reason>` `<Reason>` He assisted the police once apprehended. `</Reason>` `<Reason>` 2)Although another young offender involved in these offences had received a sentence of 9 months open custody, uniformity of sentence was not a factor to be given great weight in comparing dispositions of young offenders. `</Reason>` `<Conclusion>` 3)The sentence was increased to 10 months open custody. `</Conclusion>`	This case involves a young offender, J.M., who pleaded guilty to multiple counts of auto theft, possession of stolen auto, and dangerous driving. The court considered several factors, including the extent of J.M.'s co-operation with the police, his father's involvement in his life, and the need for uniformity in sentencing among young offenders involved in similar crimes. Ultimately, the court decided that a fit disposition for J.M. would be four months open custody for the offenses prior to June 21, 1994, and six months open custody for the offenses committed on June 21, 1994, resulting in a total of ten months open custody and 18 months probation.	J.M. pled guilty to 16 counts of auto theft, one count of possession of stolen auto and one count of dangerous driving. He was sentenced to six months' open custody and two years' probation. The judge considered the police chase, deliberate damage to vehicles, and the large number of charges as aggravating factors. He also took into account J.M.'s age, lack of criminal record, cooperation with the police, and his father's responsible approach throughout the event. No, it is not possible to equate the two. Uniformity of sentencing is not a factor to be given great weight when comparing dispositions of young offenders, but an exception must be made in the case of youths who form part of the same enterprise and who are of the same age and circumstances. Probation does not sufficiently recognize the multiple nature of the offences and the malicious damage caused, but imposing the same disposition as [T.D.] would not be appropriate for several reasons. The appeal is allowed and the dispositions for the auto offences prior to June 21, 1994 will be set aside and replaced with four months open custody and 18 months probation, running consecutively for a total of ten months open custody. The terms of the probation order remain the same.

than GPT-3.5.

In terms of cost, we consider the current pricing scheme for both GPT-3.5 and GPT-4 based on the number of tokens submitted to and generated by the model. The pricing of GPT-3.5 is set to $0.02 per 1,000 tokens in both prompt and completion, while the pricing for GPT-4 is set to $0.03 per 1,000 tokens in prompt and $0.06 per 1,000 tokens in completion. The cost of using GPT-3.5 with argumentative segmentation to generate a summary is approximately $0.19 on average. In comparison, the average cost for using GPT-4 is about $1.31. This means that GPT-4 is approximately 10 times more expensive than GPT-3.5 for the summarization task.

We also examined some of the summaries generated by the non-GPT models. The quality of summaries is clearly lower than GPT generated summaries. One pos-

sible reason is that large language models are trained on a much larger corpus and have more extensive model architectures, which makes them better few-shot or even zero-shot learners [41].

5. Limitations

In this study, we focus on the effect of using argumentative segmentation on legal summarization. While we observed improvements in the model performance of legal summarization with argumentative segmentation, we also some coherency issues in the generated summaries. For example, "Yes, I agree with Mr. Stobie" interrupt the information flow of the summary from Table 8. Thus, a systematic human evaluation of generated summaries is

Table 8

Argumentative segmentation enhanced GPT-3.5 generated summaries and baseline GPT-4 generated summaries versus the reference summary.

Reference summary (case name: 1995canlii4465 (5,221 tokens in the original case decision))	Baseline GPT-4	Arg. Segmentation Enhanced GPT-3.5
The plaintiffs statement of claim was served on the defendant in Ontario two days after it was issued.Two weeks later the prothonotary issued default judgment for $8,000, plus $2,880 in interest and $404.42 in costs. `<Issue>` The defendant applied, pursuant to Civil Procedure Rule 12.06, to set aside the default judgment, on the ground that the plaintiff did not meet the requirements for obtaining a default judgment from the prothonotary or, in the alternative, on the merits that it had, by affidavit evidence, raised triable issues and provided a reasonable excuse for its failure to file a defence. `</Issue>` `<Conclusion>` Dismissing the application with costs of $1,000 to the plaintiff, `</Conclusion>` `<Reason>` that although the plaintiff ought to have sought the court's leave to take default judgment such a procedural ommission did not, in the circumstances, prejudice the defendant. `</Reason>` `<Reason>` Further, while there were triable issues between the parties, the defendant's failure to file a defence was neither a mistake nor an oversight, but rather a deliberate decision so as not to attorn to Nova Scotia jurisdiction, and therefore was not a reasonable excuse. `</Reason>`	The defendant applied to set aside a default judgment obtained by the plaintiff, arguing procedural irregularities and that there were legitimate triable issues between the parties. The court dismissed the defendant's application, finding that any irregularities did not materially affect the defendant and that the cases relied upon by the defendant were not applicable to the circumstances of this case. The court also found that the defendant had failed to establish any reasonable excuse for its failure to file a defense, as it was a deliberate decision not to file a defense rather than a mistake or oversight.	The defendant argued three points in support of setting aside the default judgment: 1) procedural, 2) irregularity as of right, and 3) legitimate triable issues with reasonable excuse for failure to file a defence. The defendant's application to set aside the default judgment was dismissed because any irregularities that may have prompted the issuance of the Prothonotary's default judgment were not enough to make the Order a nullity, and the procedural omission did not affect the defendant. The appellant must show by affidavit facts which would indicate clearly that he had good defence to the action on the merits, and why his defence was not filed and delivered within the time limited by the Rules. Yes, I agree with Mr. Stobie that this was not a mistake or an oversight, but rather a deliberate decision not to file a defence. Ms. Roberts was aware that not filing a defence would result in default, and she had not provided reasonable excuse for the delay. The defendant's refusal to commit to defending in Nova Scotia or bring a motion to stay on the basis of forum non conveniens is not a reasonable excuse, and therefore the court will not exercise its discretion in favour of the defendant. The defendant's application is dismissed with costs to the plaintiff in the amount of One Thousand Dollars ($1,000.00).

needed to further examine the performance of the models and address these coherency issues.

Furthermore, reproducing our results may be challenging due to the proprietary nature of the OpenAI GPT models used in our experiments. Especially, we employed different combinations of control parameters in the experiment will further decrease the possibility of reproduction. Additionally, any updates or changes to the GPT models by OpenAI may result in changes to performance and results. So it is crucial to develop methods to increase the reproducibility of the results.

6. Conclusion and Future Work

We have proposed a novel task of extracting argumentative segments that include the main points of legal case decisions. We further proposed to utilize these argumentative segments to guide a summarizer. Our experiments with GPT-3.5, GPT-4 and other models showed that the argumentative segmentation enhanced method can improve the automatic evaluation scores of generated summaries. This method also overcomes the request token limitation imposed by GPT-3.5. Our findings reveal a boost in performance across all types of automatic evaluations scores using the predicted argumentative segments. Additionally, we observed that GPT-4 tends to produce more coherent summaries compared to GPT-3.5.

For future work, we will further explore methods to ensure more reliable performance of the proprietary models. Furthermore, we plan to investigate alternative prompt engineering techniques for the summarization task. Due to the nature of generative models, a systematic human evaluation on the generated summaries are much needed in the future.

Acknowledgments

References

[1] S. Teufel, A. Siddharthan, C. Batchelor, Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics, in: Proceedings of the 2009 conference on empirical methods in natural language processing, 2009, pp. 1493–1502.

[2] D. Radev, E. Hovy, K. McKeown, Introduction to the special issue on summarization, Computational linguistics 28 (2002) 399–408.

[3] W. S. El-Kassas, C. R. Salama, A. A. Rafea, H. K. Mohamed, Automatic text summarization: A comprehensive survey, Expert systems with applications 165 (2021) 113679.

[4] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, et al., Review of automatic text summarization techniques & methods, Journal of King Saud University-Computer and Information Sciences (2020).

[5] S. Gupta, S. K. Gupta, Abstractive summarization: An overview of the state of the art, Expert Systems with Applications 121 (2019) 49–65.

[6] Y. Zhong, D. Litman, Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions, in: Proceedings of the Natural Legal Language Processing Workshop 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 322–337. URL: https://aclanthology.org/2022.nllp-1.30.

[7] Y. Dong, A. Mircea, J. C. K. Cheung, 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, 2021, pp. 1089–1102.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, 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, 2020, pp. 7871–7880.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[10] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[11] M. Elaraby, D. Litman, Arglegalsumm: Improving abstractive summarization of legal documents with argument mining, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6187–6194.

[12] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155 (2022).

[14] S. Teufel, et al., Argumentative zoning: Information extraction from scientific text, Ph.D. thesis, Citeseer, 1999.

[15] M. Liakata, S. Teufel, A. Siddharthan, C. Batchelor, Corpora for the conceptualisation and zoning of scientific papers, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.

[16] C. Dayrell, A. Candido Jr, G. Lima, D. Machado Jr, A. Copestake, V. D. Feltrim, S. Tagnin, S. Aluísio, Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 1604–1609.

[17] B. Hachey, C. Grover, Automatic legal text summarisation: experiments with summary structuring, in: Proceedings of the 10th International Conference on Artificial intelligence and Law, 2005, pp. 75–84.

[18] M.-F. Moens, C. Uyttendaele, Automatic text structuring and categorization as a first step in summarizing legal cases, Information Processing &

Management 33 (1997) 727–737.

[19] A. El-Ebshihy, A. M. Ningtyas, L. Andersson, F. Piroi, A. Rauber, A platform for argumentative zoning annotation and scientific summarization, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 4843–4847.

[20] R. Mochales, M.-F. Moens, Study on the structure of argumentation in case law, in: Proceedings of the 2008 conference on legal knowledge and information systems, 2008, pp. 11–20.

[21] M. Saravanan, B. Ravindran, Identification of rhetorical roles for segmentation and summarization of a legal judgment, Artificial Intelligence and Law 18 (2010) 45–76.

[22] V. W. Feng, G. Hirst, Classifying arguments by scheme, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011, pp. 987–996.

[23] R. Mochales, M.-F. Moens, Argumentation mining, Artificial Intelligence and Law 19 (2011) 1–22.

[24] R. M. Palau, M.-F. Moens, Argumentation mining: the detection, classification and structure of arguments in text, in: Proceedings of the 12th international conference on artificial intelligence and law, 2009, pp. 98–107.

[25] H. Xu, J. Savelka, K. D. Ashley, Toward summarizing case decisions via extracting argument issues, reasons, and conclusions, in: Proceedings of the eighteenth international conference on artificial intelligence and law, 2021, pp. 250–254.

[26] H. Xu, J. Savelka, K. D. Ashley, Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences, in: Legal Knowledge and Information Systems, IOS Press, 2021, pp. 33–42.

[27] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 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), 2019, pp. 3982–3992.

[28] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, D. E. Ho, When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings, in: Proceedings of the eighteenth international conference on artificial intelligence and law, 2021, pp. 159–168.

[29] W. Yin, Y. Pei, Optimizing sentence modeling and selection for document summarization, in: Twenty-fourth international joint conference on artificial intelligence, 2015.

[30] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, M. Nagata, Single-document summarization as a tree knapsack problem, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1515–1520.

[31] D. Miller, Leveraging bert for extractive text summarization on lectures, arXiv e-prints (2019) arXiv–1906.

[32] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1073–1083.

[33] T. Goyal, J. J. Li, G. Durrett, News summarization and evaluation in the era of gpt-3, arXiv preprint arXiv:2209.12356 (2022).

[34] H. Xu, J. Šavelka, K. D. Ashley, Using argument mining for legal text summarization, in: Legal Knowledge and Information Systems, IOS Press, 2020, pp. 184–193.

[35] F. Y. Choi, Advances in domain independent linear text segmentation, in: 1st Meeting of the North American Chapter of the Association for Computational Linguistics, 2000.

[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[37] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[38] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[39] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[40] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.