Faithfulness and Content Selection in Long-Input Multi-Document Summarisation of U.S. Civil Rights Litigation

Isabel Sebire

Master of Science
School of Informatics
University of Edinburgh
2023

Abstract

Automatic summarisation of legal cases would reduce the burden on legal professionals and increase the accessibility of the law. However, the abstractive methods which dominate recent research are prone to hallucination. Despite the fact that faithful summaries are a requirement for practical use, preventing hallucination is currently an understudied area in the legal domain. In this paper, we produce a novel contribution by conducting the first study at the intersection of legal, multi-document, and faithful summarisation. We study the impact of content selection, legal pretraining, and planning through entity chaining on the quality and faithfulness of abstractive summaries of U.S. civil rights litigation, using the Multi-LexSum dataset and PEGASUS as our backbone model. While our full pipeline outperforms previous PEGASUS performance by 0.99 ROUGE-1 F1, when using oracle extracts, we achieve an improvement of 5.56 ROUGE-1 F1, 5.46 ROUGE-2 F1, 2.7 ROUGE-L F1, and 2.15 BERTScore over the state-of-the-art. Our results demonstrate the importance of content selection for summary faithfulness and quality for long-input legal abstractive summarisation, and that legal pretraining can further boost summary quality when an effective input representation is used. However, entity chaining shows little effectiveness in mitigating remaining hallucinations. Our work highlights that several issues, including the quality of the content selection method and addressing particular hallucination scenarios, remain to be addressed for long-input legal abstractive summarisation to see real-world use and adoption.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Isabel Sebire*)

Acknowledgements

I would like to express my gratitude to Claire Barale, Andrew Horne, and Mirella Lapata for their valuable guidance and advice, and to my friends and family for their constant support along the way.

Table of Contents

Chapter 1

Introduction

The intersection of AI and Law is an area of increasing academic and commercial interest [1]. In jurisdictions following the common law system, including the United Kingdom [2] and United States of America [1], judicial decisions are informed by relevant past cases (precedent cases) [1, 3] - thus, finding relevant precedent cases is of great importance [1, 2, 3, 4]. However, with ever-growing numbers of precedent cases, each typically hundreds of pages long [5], there is an increasing burden on legal professionals [6] as retrieving and understanding legal documents is a time-consuming task [1]. To assist with this, popular legal retrieval systems provide case summaries [7], however, these are currently produced by legal experts, which is costly and time consuming [8, 9], and not available for all cases. Automatic summarisation of legal cases using tools from natural language processing (NLP) would therefore bring significant benefits to various stakeholders:

- Legal professionals - case summaries would greatly reduce the time spent identifying relevant cases [1, 10], reducing the burden on legal professionals.
- Ordinary citizens – although case law is publicly available for transparency purposes, long and complex legal documents are inaccessible for the average citizen [11]. Summaries in more simple language would improve the accessibility of the law [9, 11]. Indeed, the existence of the U.S. 2010 Plain Writing Act shows the importance of making legal information more accessible [12].
- Summarisation researchers – as legal summarisation is a particularly challenging domain, research would also be helpful for automatic summarisation research in general, and for other complex and specialised domains (such as the biomedical

This chapter includes material adapted from this project's Informatics Project Proposal.

domain).

- Legal NLP researchers – for other legal NLP tasks, long input texts can be a problem for applying techniques developed in the general domain. Prior summarisation could thus be beneficial for such tasks, as has been demonstrated for case law retrieval [3, 4].

However, legal text has distinct characteristics, to the extent that it is often classified as a sublanguage [13, 14, 15]. Legal documents often contain terms uncommon or with different meaning to their use in standard English corpora [5, 16], and sentences are often long with usually complex syntax [5, 17]. Furthermore, legal texts are typically much longer than the maximum length that transformer-based models can handle, including models specifically designed for long text [1, 5, 9]. These considerations mean that legal text, in addition to being a domain with profound practical implications, is a particularly challenging and interesting domain for NLP tasks, as techniques primarily used and trained on more typical text are not always effective when applied out-of-the-box [5].

Our project contributes towards a trustworthy methodology for the abstractive summarisation of real-world legal text; in particular, we study Multi-LexSum [18], a challenging dataset focusing on multi-document summarisation of U.S. civil rights litigation. Automatic summarisation methods aim to condense input text into a fluent shorter text retaining the key information [19, 20, 21]. While legal summarisation research traditionally focused on extractive methods, which identify then assemble key elements from the source text [22], the development of large transformer-based models appropriate for legal text [17, 23] has resulted in research turning towards abstractive summarisation, where the summary is generated from scratch [22]. While this allows for more natural summaries, the ability to generate novel text means that the results of abstractive summarisation methods may contain information unrelated or unfaithful to the source [19]: this problem is called hallucination. However, despite the fact that hallucination has been highlighted as a major barrier for the practical applicability of abstractive summarisation tools [24], and especially in the high-stakes domain of law where trust in automated systems is already a known issue [25], this problem is understudied with respect to legal data. Our work will address this research gap.

The remainder of this report will be structured as follows: section 2 will review the relevant literature and background, section 3 will outline this project's research questions, and section 4 will explore the methods used. We present our and evaluate our results section 5. Finally, in section 6, we provide a critical assessment of the project.

Chapter 2

Background and Related Work

2.1 Summarisation

Summarisation is among the most challenging tasks in NLP. Traditionally, automatic summarisation research focused on extractive methods, which identify then assemble key information from the source text [22]. Typically, these methods proceed by first scoring textual units (such as sentences), then selecting the top scoring units to form the summary [16]. As extractive summaries only contain text directly from the source, this ensures a baseline level of grammaticality and accuracy, as hallucinations are not possible [26]. However, extractive approaches cannot perform paraphrasing or generalisation, which is critical for high quality summarisation, resulting in summaries of the same style as the original text, which may not flow logically [27, 28]. Traditional extractive summarisation methods include approaches based on graphs [11, 12, 16], statistical and semantic features [11, 16], latent semantic analysis [11], and KL divergence [22]. Modern extractive methods which frame the task as a binary classification problem, such as BERT-SUMM [29] and SummaRunner [30] are increasingly based on neural architectures [31, 18]. An alternative paradigm to summarisation is abstractive summarisation, where the summary is generated token-by-token, conditioned on the source text and previously generated tokens [22], allowing the summaries to contain novel words and phrases [32]. The development of large transformer based neural language models [33, 34, 35] has caused abstractive summarisation methods to receive increasing research interest [7].

In terms of the evaluation of system-generated summaries, high quality summaries cover the original text's key content cohesively, faithfully [19], and without redundancy

This chapter includes material adapted from this project's Informatics Project Proposal.

[36]. ROUGE, measuring lexical fluency and relevance [37], is the most widely used evaluation metric in the literature, and has a number of variants, each of which generates precision, recall, and F1 scores [22]. ROUGE-N is based on the n-gram overlap between generated and reference summaries [19, 22], whereas ROUGE-L is instead based on the longest common subsequence, which is the longest sequence of words shared between the predicted summary and reference summary [19, 22].

2.2 Legal Summarisation

The majority of legal summarisation research focuses on extractive summarisation. Approaches using the rhetorical structure of legal documents have seen a great deal of research attention [10, 38, 39, 40], though approaches based on feature engineering [11, 22, 41, 42], Maximum Marginal Relevance [20, 43], outcome prediction [43], linear programming [44], gravitational search [45], knowledge bases [9], and citations [46] have also been investigated. As with summarisation in the general domain, neural abstractive approaches are becoming more common [7, 47], and have been shown to significantly outperform extractive methods [16], especially as transformer-based models pretrained on legal corpora [2, 8, 17, 23, 48] have now been publicly released. Methodologies for legal abstractive summarisation have investigated incorporating chunking [2, 49], extractive summarisation [21], multitask learning [50], and argument roles [51, 52]. However, despite promising experimental results [16], the literature tackling legal abstractive summarisation is still relatively small [19], with a number of innovations from the general domain not yet investigated. Multi-document summarisation is particularly understudied in the legal domain.

2.3 Faithfulness and Hallucination

While abstractive summarisation approaches allow for more natural summaries, the ability to generate novel text introduces the possibility that (potentially plausible sounding) content which is not supported by the source text is included in the summary [24, 53, 54]. This problem is called hallucination [19], and presents a major barrier to the applicability of automatic abstractive summarisation methods [24, 55, 56, 57], especially in the legal domain [16, 19]. Faithfulness refers to how consistent the generated text is with respect to the provided input text [58]; thus, to decrease hallucination is to increase faithfulness [54]. It is however important to make the distinction between faithfulness and factuality. Hallucinated content (not consistent with the source) may

indeed be factual (consistent with world knowledge) [54, 59, 60], though the majority of current literature takes the assumption that 'any generated facts not appearing explicitly in the source are undesired hallucinations' [60].

Only one existing work [19] attempts to tackle the problem of hallucination for legal domain summarisation. [19] propose the LegalSumm method which generates summaries for multiple distinct chunks of the source text. A textual entailment model is then used to score the chunk-summary pairs, and the most faithful is used as the final summary. While this approach is promising, there are several limitations. The faithfulness model is trained with the 'faithful' examples as ground truth chunk-summary pairs from the dataset, but the information in the summary is not necessarily all contained within the input chunk. Similarly, the 'unfaithful' examples are constructed by selecting a summary from a different case for a given input chunk; this may not mirror real hallucination patterns. Overall, these factors limit the performance of the faithfulness model. Furthermore, as the final summary of the case is the summary derived from only one chunk, it is unlikely that all salient information from the document is included. We also note that this method is not applicable to all judicial documents through its use of specific case structure in the chunking process.

There are a variety of techniques to control hallucination in the general domain which have not yet been applied to legal data, presenting a large research gap. These methods include filtering training examples [61, 62]; maximising faithfulness metrics during training [53]; modifying beam search, for example through reranking [32, 58, 62] or constrained decoding [63]; and post-generation fact correction [24, 54]. Including additional information to guide generation has also been investigated: [60] provide additional information in the form of a knowledge graph, and [64] utilize explicit encodings of extracted facts in a sentence summarisation task - however the feasibility of this approach in a document (let alone multi-document) setting is questionable.

An alternative approach to providing additional guidance to summarisation models is through planning [57]. In planning-based methods, gold-standard summaries are prepended with a plan, where the plan and summary are separated by some special markers [57, 65]. At decoding time, the decoder therefore learns to first generate the plan conditioned on the input text, and then generate the target summary conditioned on both this plan and the input text, overall generating the concatenated content plan and summary [36]. As no alterations to the model architecture are required, planning is a model-agnostic and flexible strategy. [57] use a plan in the form of a question-answer pair blueprint to reduce hallucination. While promising results are achieved,

their blueprint creation process is involved and relies on the competence of question generation and question answering models; no high performing question generation or reading comprehension models yet exist for the legal domain [1]. A simple and flexible planning method is entity chaining [65]. In this approach, the plans take the form of entity chains, which are ordered sequences of the entities mentioned in the summary. [65], referring to this planning method as FROST, reported that finetuning PEGASUS in this way led to improved entity specificity and planning on all four datasets investigated. [37] independently implemented an analogous method (JAENS), reporting reduced hallucination with respect to entity-level metrics. Planning techniques have not yet been investigated as a mechanism for reducing hallucination in the legal domain - indeed, [66] suggests more guided generation of summaries as a future direction for the summarisation of long legal texts.

2.4 The Multi-LexSum Dataset

Our study will use the recently proposed Multi-LexSum [18] dataset; the first dataset for legal multi-document summarisation. This dataset contains 9,280 expert written summaries in accessible language pertaining to 4,539 U.S. civil rights lawsuits (cases) of diverse types (Figure 2.1) between 1950 and 2021 (Figure 2.2), obtained from the Civil Rights Litigation Clearinghouse (CRLC). In the U.S. legal system, civil lawsuits typically begin when the *plaintiff(s)* file a complaint against the *defendant(s)*. As the case proceeds, documents from these *parties* and the judge(s) are formally filed. We purposely choose a *case law* dataset; while some existing works attempt to summarise legal acts [47], case law provides a more practically useful application due to its larger and more rapidly increasing volume, and the fact that the general content of the document is useful (for example, for identifying precedent cases), unlike for legal acts, where the *exact wording* (which may be lost in abstractive summarisation) of the source document is often key. For each case in the Multi-LexSum dataset, the text to be summarised is the collection of relevant documents throughout the case. These documents include complaints, motions, court opinions, and settlement agreements, for example (see Appendix A). A document type of particular note is the docket, which contains a chronological record of every document that is filed in a given case. Each of a case's documents can be very long (over 100 pages) with a single case potentially involving hundreds of documents; a case's documents can extend to thousands of pages of text [18]. Indeed, for standard cases, writing summaries requires 1-4 hours, and

unusually long or complex cases can take over 10 hours, even for an experienced lawyer. Both the number of documents per case and the total number of words in a case's source documents exhibit a Zipfian distribution; large outlier cases include International Refugee Assistance Project ("IRAP") v. Trump, which contains over 250 long documents.

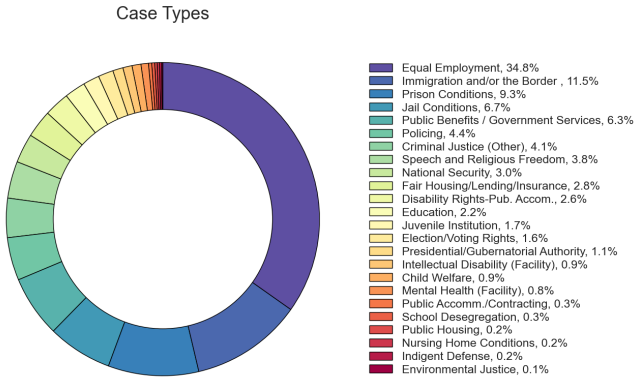


Figure 2.1: Prevalence of case types in the Multi-LexSum dataset.

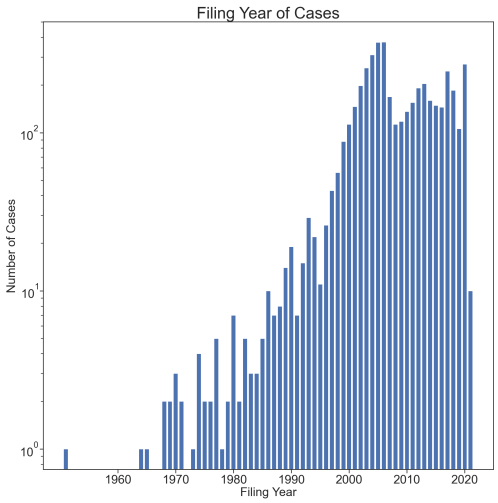


Figure 2.2: Distribution of number of cases per filing year.

While Multi-LexSum contains multiple levels of summary granularity (long, short, and tiny - examples of each are given in Appendix B), in this study we focus on short summaries (mean 130 words) - long summaries (mean 646.5 words) could feasibly exceed the maximum decoder token length (1024) for the backbone transformer model (PEGASUS) we will use, especially if including entity chains, and tiny summaries (mean 24.7 words) are sparse in the dataset and have limited practical applicability. Short summaries are typically a single paragraph, covering the background, parties involved, and outcome (so far) of the case [18]. Focusing on cases with short summaries, a mean of 99378.2 words (10.3 documents) must be summarised per case, giving a very high compression ratio of 840.7. The coverage (percentage of words in the summary

Dataset	Compression Ratio
Multi-LexSum / Short	840.7
SciTLDR	310.8
BookSum / Book	146.3
BigPatent	36.8
Multi-News	8.2

Table 2.1: Compression ratios of Multi-LexSum for short summaries and for other representative datasets, demonstrating the very high compression ratio for Multi-LexSum / Short [18].

which are found in the source documents) of 96% suggests few unexpected entities or hallucinations, and the density (average length of extractive fragments in summaries) of 3.33 suggests the summaries are highly abstractive [18].

The summaries in Multi-LexSum are authored by domain experts following carefully created guidelines, and are reviewed to ensure correctness and stylistic consistency. This ensures the quality of reference summaries and hence that supervised machine learning approaches are suitable; there is minimal risk of training on summaries containing facts unsupported by the source document, which can contribute to hallucination [18].

Overall, Multi-LexSum presents a challenging summarisation task, due to its high compression ratio and the complex real-world legal documents used, requiring understanding and synthesis of key events in a case. Multi-LexSum is also understudied; currently, the only experimental results on this dataset for short summaries are using off-the-shelf non-domain-specific models [18]. No legal-pretrained models were investigated, despite the fact that large language models pretrained in the legal domain have been shown to be more effective for other legal datasets [2], suggesting a promising future research avenue. The authors investigated BERT-EXT, PEGASUS, BART, LED-4096, LED-16384, PRIMERA models along with simple extractive heuristics, with the LED-16384 model achieving the best results for short summaries respect to ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1 scores. In general, the results of [18] suggest that models with longer inputs perform better. However, this is likely due to the naïve content selection method used for dealing with the models' maximum token length - the benefit of a longer input length is thus an increased likelihood of including salient information. The dataset was not cleaned or temporally ordered, and a simplistic method to deal with the token limitation of transformer-based models was used (for a maximum token length of L, and D documents, simply taking the first L/D tokens of each document)[1]. In a domain with noisy data, potentially hundreds of documents per

[1]This information is not contained within the publication, and was obtained by personal correspondence with the authors.

case (which may not be equally important) and so a very high compression ratio, and limited evidence of lead bias, this strategy means that salient information is not likely to be included as input to the summarisation model, reducing summary quality and potentially encouraging the model to hallucinate as the gold summaries and model inputs are not necessarily tightly coupled. The results of the authors' human evaluation study also suggest that an alternative content selection strategy is likely to improve model performance. In this investigation, a system for participants to select salient information when reading source documents was developed to aid the BART model in generating long summaries. While ROUGE-1 decreased, this method increased ROUGE-2 and ROUGE-L, and the study participants indicated that this approach led to higher summary quality than the fully automated system using the naïve content selection strategy. However, the generations still received a mean score of only 0.43 on a 0-3 Likert scale by human evaluators, indicating that the summaries are still far from acceptable [18]. Overall, investigating alternative input representations is a clear research opportunity. Additionally, in their evaluation of model outputs, the authors noted that their results were prone to hallucination [18] - a severe problem for trust and adoption. Indeed, research in general domain summarisation suggests that hallucinations are more likely when the sentence combines content from multiple source sentences [67], which is likely for a dataset with a low density which is as highly compressive as Multi-LexSum. Our research will attempt to address these research gaps.

Although less relevant to our study, we also note that the authors [18] investigated using all summary granularities in a multi-task learning framework, and generating short and tiny summaries with longer summary granularities as input. The latter investigation has limited real-world use as creating summaries from the *source documents* is the time-consuming task, and by inspection, short summaries are largely extracts of the long summaries. However, the authors do note that this provides 'evidence that inputs with more condensed information simplify the summarisation task' [18], again suggesting the promise of using a more sophisticated content selection strategy. The only other study using Multi-LexSum is [68] which investigated incorporating discourse structure into the generation of *long* summaries.

Chapter 3

Problem Statement

Informed by the opportunities suggested by the current literature, our work will address the following research questions:

1. RQ1: Can we improve abstractive summarisation results on the Multi-LexSum dataset simply by providing a better representation of the source data to the summarisation model - namely, by conducting dataset cleaning and using a content selector in a pipeline approach?
2. RQ2: Are transformer based models pretrained in the legal domain effective for legal multi document abstractive summarisation on the Multi-LexSum dataset?
3. RQ3: Is planning with entity chaining an effective method for reducing hallucinations in the abstractive summaries produced?
4. RQ4: Are the summaries produced easier to understand than the original cases?

In addressing these research questions, we contribute to the growing literature on faithfulness in abstractive summarisation, legal summarisation, and multi document summarisation by being the first work at this intersection. As faithfulness, understandability, trust, and applicability to realistic datasets are key factors to ensure that automatic legal summarisation can eventually benefit the relevant stakeholders, our methodology and research questions are designed with these factors in mind.

Chapter 4

Methods

In this section we detail each stage in the project pipeline, shown in Figure 4.1.

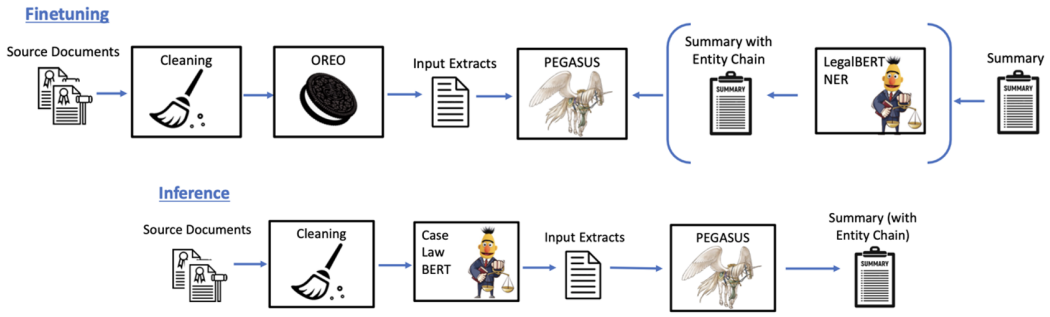


Figure 4.1: Overview of our methodology.

4.1 Preprocessing

Unlike [18], we first apply a number of preprocessing steps to the Multi-LexSum data, to allow for correct ordering and segmentation, and improve faithfulness.

4.1.1 Cleaning

The use of OCR (as required in real-world scenarios) to obtain plain text data from PDF court documents [18] of variable legibility containing formatting such as headers, footnotes, citations, and tables results in the source text in the Multi-LexSum dataset containing errors and noise. Therefore, despite the underlying quality of the judicial documents, we first conducted dataset cleaning to allow for subsequent steps such as segmentation to be meaningfully applied, as in many cases we find 'junk' in the middle of paragraphs or sentences, and erroneous line breaks.

The overall cleaning pipeline for each source document is illustrated in Figure 4.2. To define the rules for cleaning, we studied the text in the Multi-LexSum dataset and

the corresponding original documents available on the CRLC website for cases in the validation set. For each newly implemented rule, we tested their validity on subsequent documents in the validation set, and ensured that previously considered documents were not adversely affected. This process continued until a stable set of rules was reached, which was then applied to all source documents.

Figure 4.2: Summary of main stages of the cleaning pipeline.

- Removal of footers: We removed document footers containing irrelevant entities.
- Removal of headers: We only keep lines meeting at least one of the following conditions:
  - The line stripped of numerals only occurs once in the document - headers occur multiple times in the document, but may contain page numbers; thus, when stripped of numerals, this stripped line occurs multiple times.
  - The length of the stripped line is less than 20 characters - headers are long, we do not want to remove other information which may be repeated throughout the document, such as names, or terms such as ‘v.’ or ‘and’.
  - The line does not contain any numerals or hyperlinks - headers usually contain one or both, and we do not want to remove useful information.
- Removal of dirty lines: Dirty lines include page numbers, hyperlinks, lines not containing alphabetical characters, timestamps, and ‘junk’ resulting from OCR. Timestamp lines were identified using the dateutils parser - this parser handles most date formats, but throws an error if there is non-date information present. Thus, we can identify a line containing *only* date information if no error is thrown. To remove ‘junk’ lines resulting from the OCR process, we edited garbage_detector[1] (based on [69]), which identifies a line of text as ‘garbage’ if any one of several given conditions holds. We removed two of the conditions originally provided, as these gave many false positives in the legal domain: uppercase between lowercase; two distinct punctuation marks in the same line (legal sentences can be very long and can contain periods, commas, brackets, dashes in citations and legal acts, and colons and semicolons in lists, for example).

[1] https://github.com/foodoh/rmgarbage

We kept the remaining three original conditions, relating to a string's ratio of alphanumeric characters to total characters, ratio of consonants to vowels, and if a punctuation mark repeats consecutively (this condition was edited to reflect the fact that while periods and brackets can legitimately repeat consecutively, punctuation marks such as commas, colons, semicolons, and dollars cannot). Finally, we added a condition to capture the fact that certain punctuation marks appearing between lower-case letters is indicative of junk text.

- Line breaks: This includes removing blank lines, removing newlines in the middle of sentences or paragraphs, and correctly ensuring a newline before each new legal paragraph. We kept existing line breaks only after colons (used to precede legal lists), after periods where the previous character was not a capital letter or 'v' (to avoid line breaks after abbreviations such as v. or U.S.), or if the whole line consisted of upper case letters (indicative of a section title). To insert the correct line breaks between legal paragraphs, in judicial documents of 'standard' format a new legal paragraph can be identified by a numeral or letter (in the case of lists) followed by a period. At this phase, we had to consider a number of special cases. For example, we do not insert a newline after a colon if the colon is not followed by whitespace, so as not to insert a newline in the middle of a hyperlink.
- Clean remaining lines: We remove footnotes and floating punctuation.
- Additional docket processing: Docket documents have a distinct format to judicial documents of other types. In particular, dockets contain tables with two columns giving the date (left), and the action taking place (right), which are not well represented in plain text format. For dockets, we remove lines consisting solely of dates (the left column of the table), and numbers at the start of lines, as this is noise from attempting to linearise the table. In the vast majority of cases, no information is lost as the corresponding date is included in the main column entry.
- Address line breaks: Removing junk information often allows us to retrieve the correct line breaks. For docket documents, this phase is different, as due to the text originally being table cells, newline characters separate sentences.

An annotated representative excerpt from a case document before and after cleaning is given in Appendix C. This displays the effectiveness of our cleaning pipeline, however we note that cleaning cannot be perfect in all cases since documents have different formats and levels of OCR noise, and we do not want to erroneously remove valid text.

The cleaning process allows the source text to be correctly segmented into sentences and paragraphs, which is vital for subsequent stages in our methodology. The newline stages of the cleaning process allow for correct paragraph segmentation. For sentence segmentation, we use LexNLP [70] as this is specifically designed for legal text. Despite this, we still found that some postprocessing was required to achieve the best results as certain cases were not well handled. Following [38], we merge a sentence with the previous sentence if the previous sentence ends in an acronym (such as 'v.'), or if the current sentence begins with 'Section' (to address incorrect segmentation within legal articles). We also introduce a sentence boundary between ';' and '(' to segment long legal lists. For docket type documents, as there is no period at the end of entries in table cells, we must first divide the text into paragraphs, which correspond to each cell of the table, before applying sentence segmentation to each paragraph.

4.1.2 Ordering

Multi-LexSum presents the source documents for each case as a list of plaintext documents. However, these documents are not chronologically ordered, and chronological ordering is crucial for understanding. As not all document texts contain a date which could be extracted and used to sort (many court documents are physically stamped with the date, and this is not adequately picked up by the OCR software), we scraped the date of each document from the CRLC website, using the links provided in Multi-LexSum, and the BeautifulSoup[2] and requests[3] Python packages.

4.1.3 Data Filtering

As training examples which are unfaithful to the source text can encourage generative models to produce hallucinations, filtering out training examples with low entity extractivity is a standard method to discourage hallucination [37, 54, 60, 62, 65]. While the summaries in Multi-LexSum are expertly constructed and faithful to the source *documents* as on the CRLC website (verified by manual inspection for a sample of cases), the OCR process means that not all documents are adequately represented by the *plain text* format in Multi-LexSum - for example, Appendix D shows a handwritten source document for which the OCR software struggles to extract *any* text. Therefore, the Multi-LexSum dataset contains cases where the source text does not contain key information in the summary; these cases should be removed. We based our filtering on

[2]http://www.crummy.com/software/BeautifulSoup/bs4/
[3]https://requests.readthedocs.io

verifying if the named entities in the summary occur in the source text. However, this is nontrivial to determine, with several recurring scenarios causing difficulty:

- Dates - the same date can occur in different formats. We adopted a very optimistic approach to filter out obvious errors, however we note that this may give false positives by indicating entities are extractive when they are not. To deal with generalisations, such as 'September 2003' occurring in the summary while the source documents may only contain specific dates (i.e. - the day of the month is also specified), we parsed such expressions into multiple date formats and attempted to find a match in the source text for any of these formats, for any day of the month. Similarly, for expressions such as 'early 2003' we solely attempted to verify the year. For relative expressions such as 'the next day', we optimistically assumed these were valid.
- Paraphrases - for example, 'AT&T employee' and 'employed by AT&T Corp.'.
- Expansion and contraction of abbreviations - for example, 'Corporation' and 'Corp.'. Creating a dictionary to match all such abbreviations would be infeasible.
- Minor errors such as inconsistent spacing and punctuation.

We note that many of these issues occur due to basing our matching on an exact match of surface forms. While we considered strategies such as fuzzy string matching, we found this to lead to worse results, as for example, changing one letter is very important when referring to legal articles, but could still lead to a fuzzy string match with high confidence. Overall, our method is not reliable at the level of individual entities, and therefore methods such as the drop prompt mechanism in FROST++ [65], which restricts the entities in the entity chain to those present in the source text [36] and led to state-of-the-art faithfulness with respect to human and automatic evaluation approaches, would not be suitable for Multi-LexSum data unless further research is undertaken with respect to entity matching, which is outside the scope of this project. However, we found through manual inspection that our method suffices to filter out obviously low-quality sources. From inspection of the percentage of entities verified, summaries, court documents on the CRLC website, and source text in Multi-LexSum for a sample of cases, we removed cases where less than 75% of entities could be verified.

We found one legitimate case where summaries contained non-extractive entities: where the final sentence of the summary indicated whether the case was closed 'as of' the date of writing. In such cases, the date of writing was evidently not contained in the source documents. Therefore, if the last sentence of summaries in the training set

contained ‘as of’, we removed this sentence so as not to encourage hallucination.

4.1.4 Data Augmentation

Only 3,138 out of 4,539 cases in the Multi-LexSum dataset contain a short summary, however in many cases the long summary is of the length and style of a short summary. This is because the long, short, and tiny summary granularities are simply the summaries produced in descending order of length; if a case only has one summary, it is by default the ‘long’ summary. As training examples were removed in the previous filtering stage, to augment the dataset, if a case has no short summary but its long summary is within 671 words (the maximum length of short summaries observed), we include this in the short summary dataset. We only conducted filtering and augmentation for the training set. Table 4.1 presents the dataset splits after filtering and augmentation.

	Complete Dataset	Short Summaries (Original)	Short Summaries (Filtered and Augmented)
Train	4,539	3,138	3,436
Validation	3,177 (70%)	2,210 (70%)	2,508 (73%)
Test	454 (10%)	312 (10%)	312 (9%)
Total	908 (20%)	616 (20%)	616 (18%)

Table 4.1: Size of train, validation, and test splits after preprocessing.

4.2 Named Entity Recognition and Entity Chains

We augment the gold-standard summaries with entity chains to implement the entity chain planning approach [65] detailed in Section 2.3. In order to conduct the named entity recognition (NER) required to produce entity chains, we use a state-of-the-art NER system [71] developed specifically for the legal domain in collaboration with legal professionals. In this way, our entity chaining approach integrates symbolic domain knowledge with neural approaches, which has been suggested in the literature as a promising approach for legal NLP [1]. Another advantage of an approach directly targeting entities is that there is empirical evidence to suggest that hallucinations in abstractive summaries, including in the legal domain, often concern entities [16, 19, 72, 73], and entities often contain the most salient information [60]. Additionally, the inclusion of a planning mechanism may help to increase trust in the overall system, a critical factor to adoption, by mirroring the human process [25]; there is evidence that in humans, planning occurs at a higher level than individual words [57, 74, 75].

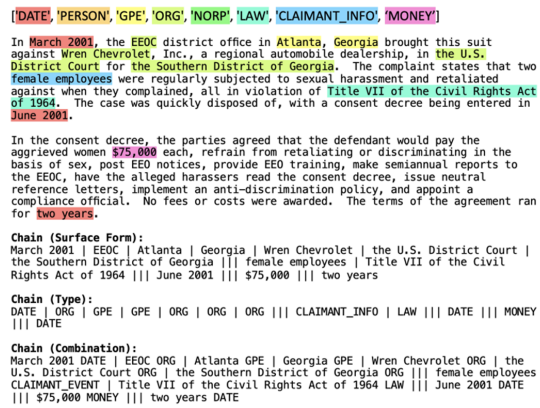
The NER system we use was trained on human annotated Canadian refugee law

cases. We use the LEGAL-BERT models[4] for all labels, as these are the best performing [71]. Details of all labels are given in Appendix E. As we do not have gold NER labels for Multi-LexSum (and the collection of such labels would be outside the scope of the project), we manually evaluated results of the LEGAL-BERT NER systems on a subset of the validation set, studying the performance and relevance of all categories. While overall the NER system performed well, we found one common error for the GPE and ORG categories - the system included additional words between two true entities, resulting in one false entity (eg *'AT&T employee* against *AT&T Corp.'*) being returned. To solve this, we used the nltk [76] part-of-speech tagger to postprocess these categories, removing words which were not nouns, adjectives, 'in', or 'of' from the entity and segmenting at the newly created boundaries. Overall, we chose to use all standard NER categories (DATE, PERSON, GPE, ORG, NORP, LAW) in addition to the CLAIMANT_INFO legal-specific category. We only selected this legal-specific category as the CLAIMANT_EVENT and PROCEDURE labels resulted in very long entity chains, and the other legal specific categories very rarely occurred, as these had limited relevance outside the refugee law domain. We also included the MONEY category from LexNLP [70] as monetary amounts are critical in law [77]. For comparison to other literature using entity chains, FROST [65] constructed entity chains using traditional named entities (such as PERSON, GPE, and ORG), dates, and numbers, while JAENS [37] excluded dates and numerals due to the 'difficult[y] [of] determin[ing] a match in the source document', which we discussed in Section 4.1.3.

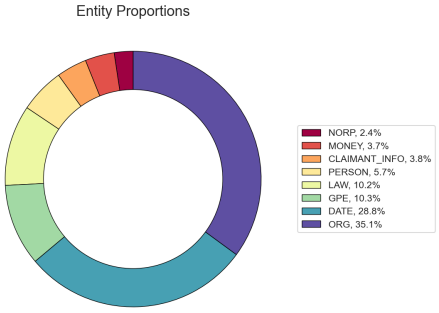
We present an example of the NER annotation in Figure 4.3(a) - Appendix F gives further examples. We also conducted exploratory analysis of entities in short summaries, finding that our chosen categories constituted a mean of 19.87% of the summary text.

After extracting the entities present in each summary, we construct three variants of entity chain to investigate the granularity of information needed for entity chaining to be beneficial as a planning mechanism. *Surface form* chains contain the span of text extracted for each entity, *type* chains contain the entity's type only, and *combination* chains contain both. An example of each type is given in Figure 4.3(a). In all cases, we construct the entity chains as in [65], and prepend the entity chain and summary with special tokens [ENTITYCHAIN] and [SUMMARY] respectively.

[4]Separate models are provided for traditional (DATE, GPE, ORG, PERSON, LAW, NORP) and legal-specific (DETERMINATION, CREDIBILITY, EXPLANATION, CLAIMANT_EVENT, DOC_EVIDENCE, PROCEDURE, CLAIMANT_INFO, LAW_REPORT, LAW_CASE) labels: https://github.com/clairebarale/refugee_cases_ner



(a) Final NER annotations for chosen categories, and associated entity chains.



(b) Proportion of entities of each category in short summaries.

Figure 4.3: NER annotation example and distribution of categories.

4.3 Transformers and PEGASUS

Following the state-of-the-art, we use a pretrained transformer architecture as our backbone abstractive summarisation model. Pretraining has been demonstrated to improve results with respect to both ROUGE and human judgements of faithfulness [78]. In particular, we use PEGASUS, a sequence-to-sequence encoder-decoder model. We chose this model for three reasons:

- PEGASUS has a **legal-pretrained** variant, Legal-PEGASUS[5], trained on U.S. case law. While other transformer models such as T5 and BART have shown promising results in general domain summarisation [57, 79], there is no available variant of these models pretrained on legal data, and pretraining on legal data has been consistently shown to increase performance on legal NLP tasks [1, 2, 23]. The standard variant of PEGASUS is trained on news data (CNN/Daily Mail).
- Results for PEGASUS are **reported in the original Multi-LexSum paper**.
- PEGASUS has a **pretraining objective**, gap sentence generation, designed specifically for abstractive summarisation [80]. While other models pretrained on legal corpora such as RoBERTa [81, 82] and Longformer [23, 81] are available, these do not have a training objective optimized for the summarisation task, as these models were developed primarily with classification tasks in mind.

PEGASUS has achieved state-of-the-art results on various summarisation datasets [28]. Of particular note is a study of multiple summarisation methods in the legal domain,

[5] https://huggingface.co/nsi319/legal-pegasus

where Legal-PEGASUS performed best, outperforming even LED-16384 which can consider 16 times longer input texts [2].

4.4 Content Selection

The transformer architecture has brought significant performance benefits throughout NLP [83], but the self-attention mechanism introduces a maximum input token length limitation [23] (1024 tokens for PEGASUS [80]), which is particularly significant for long-document and multi-document summarisation tasks. Therefore, a strategy is needed to reduce the length of the source text which is input to the transformer-based summarisation model. While a naïve approach of simply truncating the source text at the token limit is feasible, this 'may hide vital information' [19]. We want to ensure that the important information from a very lengthy collection of source documents is used in the abstractive summarisation step, calling for a more sophisticated strategy.

Sparse attention transformers such as Longformer [84], LongT5 [85], and BigBird [86] change the underlying attention mechanism to handle longer input sequence lengths, up to 16,348 tokens [84]. While such models have shown promising performance for summarisation [5, 23], and Longformer Encoder Decoder (LED) in particular is popular in the legal summarisation literature [3, 4, 8, 52, 87], the increased token length is still often insufficient for the input text in its entirety in the legal and multi document cases. For example, [52] attempt to 'avoid' the input token limitation problem by using LED; they study a dataset containing input documents of up to 26,000 words, yet 'truncat[e] the input length to 6144 words'. For Multi-LexSum, the mean length of source text for a given case is 83,340 tokens (maximum 4,423,683 tokens). Hence, sparse attention transformers would not adequately address the content selection problem in our case.

Another approach is dividing the source text into chunks which are each below the token limit, summarising each chunk separately, then concatenating these summaries - this strategy can incorporate information from the entire input in principle, and [2] finds that this strategy performs well for their legal datasets. However, the documents in these datasets were significantly shorter than for Multi-LexSum, and the chunking strategy introduces a number of issues: it is non-trivial to extract the corresponding sentences from the reference summary for each chunk, not all chunks may be (equally) informative, and independent chunk processing may lead to redundancy in the final summary. These problems persist even for more sophisticated segmentation methods, such as Se3 in [49]. Furthermore, for input text as long as in Multi-LexSum, summarising every

chunk of the input text would be computationally expensive. A variant on chunking based approaches are multi-stage frameworks such as SUMM^{N} [88], which iteratively uses the concatenated summary as the input to another phase of chunking and abstractive summarisation. Introducing multiple abstractive stages significantly increases computational complexity, and also introduces more opportunities for hallucination.

Due to the shortcomings of the above approaches, we choose to adopt a mixed-model approach, where we first coarsely identify salient information [89, 90], then use this information as the input to our backbone abstractive summarisation model. This strategy has shown promising results for scientific literature summarisation, for example [91], and has a number of advantages. A pipeline extract-then-abstract approach can mitigate the fact that abstractive summarisation models can perform poorly at content selection [92], as the output of the pipeline benefits from extractive methods' superior content selection capabilities [7, 93]. Perhaps most importantly in the legal domain, the pipeline approach better mirrors the human summarisation process, and hence may contribute to user trust in the summarisation system. There is evidence that a human summarising long input text would first understand the text, then highlight the important information, then paraphrase this information to form a summary [21, 25, 90, 94], and a study on legal text summarisation demonstrates participants' increased trust in systems for which they understand the summary's creation process (also reported in [95, 96]) and feel that this process is similar to their own [25].

There are various approaches to the selection of salient information for pipeline approaches, including simply using extractive text summarisation [89, 90]. In the legal domain, [21] use GPT-2 perplexity scores to select salient sentences, which they use to train a binary salience classifier. The sentences classified as salient are then fed as input to a finetuned BART model to produce the final summary. Their approach resulted in a 5.05 ROUGE-L improvement compared to truncation and TextRank baselines, but the dataset used in the study contained much shorter input text than for Multi-LexSum - source documents were only compressed by 61% on average. Furthermore, despite the authors' claims, the salience classifier recovering only 64.7% of sentences marked by human annotators does not suggest a particularly strong 'correlation with human judges' [21]. This suggests that further research into salient information retrieval approaches in the legal domain would be valuable. In the general domain, [90] and [97] score input text paragraphs, and select the top scoring as input to abstractive summarisation models. As they investigate Wikipedia data, [90] and [97] both use tf-idf with the document title as a potential scoring method, however a 'document title' is not a

provided or meaningful concept for legal summarisation such as for Multi-LexSum. [90] achieve their best result with a 'cheating' method, using the recall of bigrams in the gold-standard summary to score paragraphs. Indeed, following this result, the authors suggested training a supervised model to predict relevance. Our approach mirrors this; we adopt a ranking-based approach to select salient information, by training a BERT-based salience classifier to extract useful information from the source text at a sentence level, using the state-of-the-art OREO method to obtain gold standard training labels. This allows us to create a list of all source text sentences ranked by the classifier's confidence that the sentence contains salient information. At inference time, the top-ranked sentences are used to construct the input (detailed in Section 4.4.3) to the PEGASUS model (when finetuning PEGASUS, we instead use the gold standard sentences from OREO, rather than from the BERT classifier, as the model input).

4.4.1 OREO and Oracle Construction

To train a classifier to predict if a given source sentence is salient, we must first obtain reference labels to use as training data. As annotation of sentences containing salient information by legal professionals would be prohibitively costly and time consuming (even for a single case in Multi-LexSum), we use an automatic labelling approach, by converting the gold-standard abstractive summaries to their extractive equivalent (*oracle extracts*). Various approaches have been proposed to create oracle extracts, among which greedily maximising the ROUGE overlap with the gold-standard summary is most common [26, 44, 47]. However, oracles constructed in this way do not always lead to high-performing summaries [26] - indeed, a recent study on legal extractive summarisation [47] suggests that 'alternative methods to create oracle extractive summaries' should be considered. Furthermore, this greedy approach considers only a *single* oracle summary, Y^*, but there can be *multiple* valid oracle summaries for the same source text; systems trained on greedy oracles are optimised by maximising the probability at Y^* and assigning zero probability to all other hypotheses, regardless of quality.

For this reason, we use the OREO algorithm to create oracles, which incorporates the idea of learning from *multiple* oracle summary hypotheses. Formally, the summary-worthiness of a sentence x_i is defined as the expectation of its associated oracle evaluation:

$$\ell'_i := \sum_{Y^*}^{\mathbb{Y}} \mathcal{R}(Y^*,\ S)\ p(x_i|Y^*,D)\ p(Y^*|D,S) = \mathop{\mathbb{E}}_{Y^* \sim\ p(Y^*|D,S)} [\underbrace{\mathcal{R}(Y^*,S)}_{oracle\ quality} \underbrace{p(x_i|Y^*,D)}_{oracle\ membership}]$$

where $\mathcal{R}$ denotes the mean of ROUGE-1 and ROUGE-2, $D = \{x_i\}_1^m$ denotes the

source text, S is the reference (abstractive) summary, and $\mathbb{Y}$ is the oracle summary space. The 'oracle membership' term refers to if the oracle hypothesis Y^* is in the oracle distribution, which is a uniform distribution over the t top results of the k oracle summary hypotheses returned by beam search. The final sentence labels are given by the scaled expectation $\ell(x_i) = (\ell'_i - \bar{\ell}_{min})/(\bar{\ell}_{max} - \bar{\ell}_{min})$ [26].

Compared to the greedy approach, OREO achieved a superior performance on a variety of summarisation benchmarks for extractive summarisation. The authors also showed that the extracts created by OREO can better guide the learning and inference of an abstractive summarisation system [26]. While this experiment used the GSUM model, which uses extractive summaries to *guide* the generation of abstracts, rather than as the *input* as in a pipeline approach, this still suggests that OREO could be a promising strategy for our methodology.

To obtain the OREO labels for Multi-LexSum, we set the beam size hyperparameter k to 16, and the oracle distribution hyperparameter t to 16, as in the hyperparameter search performed in [26], these were the best parameters for the most highly compressive dataset evaluated, Multi-News. We set the summary size hyperparameter to 30 (approx. 1024 / 34) sentences, based on the mean number of tokens (34, very long tail distribution) per source sentence. However, after running OREO, in many cases fewer than 30 sentences were extracted (received a non-zero score) for a given case.

4.4.2 BERT Classification

Given the 'oracle' sentences containing salient information output by OREO, we can now train a classifier to predict if unseen sentences are summary-worthy. We use BERT as the classifier architecture as BERT models are well suited for classification tasks [17, 71], and we use a model pretrained on legal text, as this has been shown to improve classification results in the legal domain [17, 52]. We use CaseLawBERT [48] as opposed to LegalBERT [17] as CaseLawBERT is trained on 37GB [5] of *U.S. case law* [82], providing a better domain match to our U.S. civil rights data than LegalBERT, which is trained on 12GB of legal data, only 27.8% of which is U.S. case law [17].

The OREO labels of sentence salience create a huge dataset with severe class imbalance, as shown in Table 4.2. While there exist multiple strategies to deal with class imbalance such as editing the cross entropy loss to include class weights [52], the sheer number of training examples in the dataset makes this approach computationally infeasible. Thus, we instead carried out random downsampling of the training data to bring the number of examples in the negative class in line with the number of examples

	Number of Instances	Positive Instances	Negative Instances
Train	6,230,772	34,296 (0.55%)	6,196,476
Validation	1,122,744	4,355 (0.39%)	1,118,389
Test	1,672,233	8,021 (0.48%)	1,664,212

Table 4.2: Number of instances of each class for sentence salience classification.

in the positive class, for a total of 68,592 training examples. We note that a more sophisticated method to deal with this class imbalance may lead to improved results - indeed, the true proportions of classes are important for classifying a new point, and this information has been lost when conducting downsampling.

We trained the CaseLawBERT model using its Pytorch implementation in the Huggingface [98] library on a single NVIDIA RTX A6000 GPU. The model was trained for 3 epochs [48], with a batch size of 16 and using BertAdam with a learning rate of 2e-5 and warmup of 0.01. Inputs were truncated at 128 tokens for feasibility reasons due to the huge number of sentences in the test set; we acknowledge that not truncating may lead to improved results. As output, we obtained the probability of the sentence containing salient information. As we are not working with a threshold (to construct the inputs to PEGASUS, we use a ranked list by probability) and as metrics such as accuracy, precision, and recall are not very informative for highly skewed data, we report the classifier's ROC-AUC score of 0.884 (curve in Appendix G) - this indicates excellent [99] performance, despite the computational considerations made.

4.4.3 Input Construction

To construct the PEGASUS inputs from the ranked list of sentences (from OREO at training time, or BERT at inference time), we consider several strategies:

- Sentences - we add the top scoring sentences with non-zero scores, as in [26], until the token limit is reached.
- Windows - for each selected sentence, we also add the preceding and following sentence. This provides context, but may lead to irrelevant information being contained and subsequent salient information no longer being able to fit in the maximum input length. We add windows until the token limit is reached.
- Paragraphs - for every selected sentence, we add the whole paragraph the sentence is contained within. We add paragraphs until the token limit is reached.

For the window and paragraph methods, we took care to avoid including duplicated sentences. In all cases, we concatenate the extracted information in order of appearance

in the (temporally ordered) source documents. While there are alternative methods of representing extracted information for multi-document summarisation, such as [97] which can capture complex cross document relations such as contrasting opinions in the news domain, for Multi-LexSum, the documents can be flatly concatenated in a meaningful way (temporal), so more complex methods would introduce additional complexity for limited benefit. We also consider three baseline methods:

- First-1024 - we take the first 1024 tokens of the temporal concatenation of all the case's source documents. This is because 1024 is the maximum token limit for PEGASUS [88]. This approach is likely to discard valuable information.
- First-K - like [18], for a case with D documents, we take the first $1024/D$ tokens of each. Unlike [18], the dataset has been cleaned and temporally ordered.
- TextRank - this is a simple general-domain unsupervised extractive summarisation method, which is frequently used as a content selection baseline [90], including in the legal domain [21, 47]. The central idea of TextRank is to represent the source text as a graph, where textual units (sentences) are nodes, and similarity is represented as edges [22]. We used the underlying implementation of TextRank from the pytextrank[6] module, extracting 30 sentences, as for OREO. As our preliminary investigation found that the model primarily returned non-salient docket entries, we excluded docket documents from the input to TextRank. As the input text sometimes exceeded the maximum for this implementation of TextRank, we chunked the input text by document, and by paragraph if a single document exceeded the input length. Then, we found the proportion p of source characters contained in each chunk, and extracted $\lfloor 30 * p \rfloor$ sentences for each chunk. However, we note that the assumption made that there is an even distribution of summary sentences across chunks is unlikely to hold in practice.

We also considered legal unsupervised extractive summarisation methods as baselines, however we found that many were not applicable out-of-the-box to Multi-LexSum. For example, GraphicalModel [100] and LetSum [40] both require preidentified cue words, and CaseSummariser [42] requires a sentence's proximity to headings (not every document in Multi-LexSum has a format with headings), dates, and entities (NER tagging the source would be computationally impractical). CaseSummariser, which was developed originally for Australian case judgements, has furthermore been demonstrated to perform poorly on documents from other jurisdictions [7, 11].

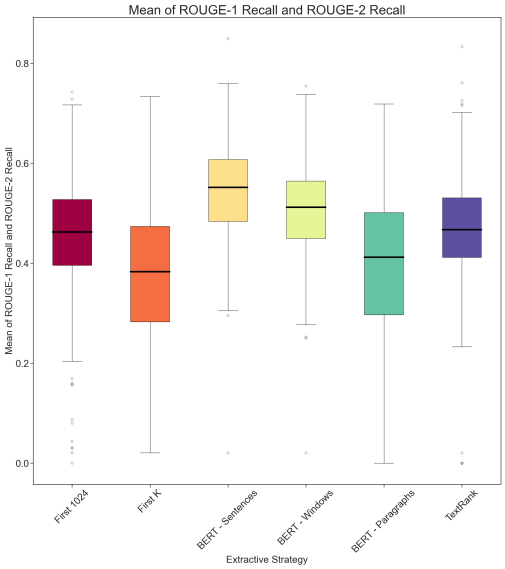
[6] https://derwen.ai/docs/ptr/

	ROUGE-1	ROUGE-2	ROUGE-L
First-1024	67.51	24.35	41.50
First-K	57.36	19.25	35.94
BERT-Sentences	76.61	32.61	46.15
BERT-Windows	73.88	28.00	41.95
BERT-Paragraphs	58.30	19.99	32.66
TextRank	70.28	23.47	43.93

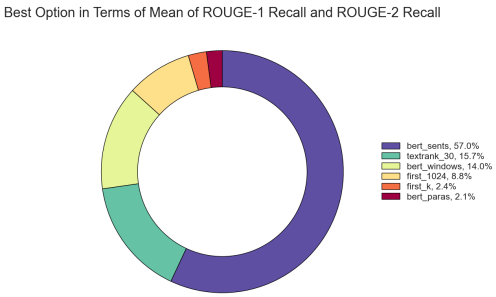
Table 4.3: Mean ROUGE recall scores against corresponding reference summary.

As a preliminary experiment, we investigate the ROUGE recall between the extracts produced and the corresponding gold standard summary for the test set (we use the test set as we have already performed the expensive inference process for the BERT classifier on this data); we use recall as we wish to consider if the salient information has been included, not the specificity of salient information. Table 4.3 presents these results. The BERT-Sentences and BERT-Windows methods clearly outperform the naïve First-1024 and First-K baselines, with the TextRank baseline performing surprisingly well. The BERT-Paragraphs method performs poorly, which is likely due to the fact that including a large amount of additional context for each selected sentences means that some sentences containing relevant information may not fit within the token limit, and also due to the fact that as only longer chunks of text are selected and we only add complete paragraphs until the token limit is reached, the total number of tokens extracted is typically fewer than for other methods (see Table 4.5). At the level of individual cases (Figure 4.4(b)), the BERT-Sentences method is most frequently the best performing option, with TextRank again performing surprisingly well.

We can also compare the three BERT based strategies to their OREO counterparts (Table 4.4). OREO-Windows gives the best performance overall with respect to ROUGE-1 and ROUGE-2 recall. All OREO strategies outperform their BERT-based counterparts with respect to ROUGE-2, however BERT-Sentences notably outperforms OREO-Sentences on ROUGE-1. At the level of individual cases and evaluating with respect to the mean of ROUGE-1 and ROUGE-2 recall, OREO outperforms BERT in 84% (519) and 85% (525) of cases for windows and paragraphs respectively, but for sentences, BERT outperforms OREO in 66% (405) of cases. We also note that OREO extracts significantly fewer tokens than BERT in the sentence case - this is because OREO frequently assigns non-zero scores to less than 30 sentences, but as BERT outputs confidences, we do not see zero scores. This is an interesting result as it suggests that an input token length of 1024 tokens is sufficient and thus that using a sparse transformer architecture allowing for longer input lengths may not improve results if a sophisticated content selection mechanism is used.



(a) Distributions of ROUGE recall scores against corresponding reference summary for all strategies investigated.



(b) Method giving highest ROUGE recall against corresponding reference summary for individual cases.

Figure 4.4: ROUGE recall of content selection strategies against reference summaries, demonstrating that BERT-Sentences is the best performing approach by this metric.

	ROUGE-1	ROUGE-2	ROUGE-L
OREO Sentences	68.07	32.73	35.43
OREO Windows	79.43	37.13	45.25
OREO Paragraphs	73.83	33.24	41.59

Table 4.4: Mean ROUGE recall scores of OREO strategies against corresponding ref-

	Mean Number of Tokens Extracted	
	OREO	BERT
Sentences	264.15	1000.78
Windows	821.31	966.10
Paragraphs	679.73	596.47

Table 4.5: Mean number of tokens extracted for BERT-based and OREO-based input strategies.

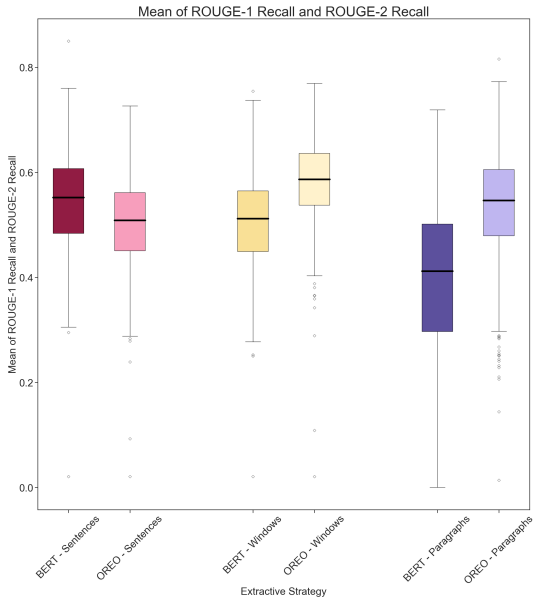


Figure 4.5: Distributions of ROUGE recall scores against corresponding reference summary for BERT-based and OREO-based strategies, demonstrating the difference in salient information retrieval between BERT-based and OREO-based counterparts.

4.5 Experimental Setup

We have now constructed a variety of input (Section 4.4) and output (Section 4.2) representations for finetuning our backbone summarisation model, PEGASUS. This gives three dimensions to vary in our experiments, corresponding to research questions:

- RQ1: input representation and content selection - First-1024; First-K; TextRank; BERT-Sentences; BERT-Windows; BERT-Paragraphs.
- RQ2: PEGASUS model pretraining - standard variant trained on CNN/Daily Mail; Legal-PEGASUS trained on U.S. case law.
- RQ3: output representation - no entity chain; surface form entity chain; type entity chain; combination entity chain.

All experiments were conducted on a single NVIDIA RTX A6000 GPU, using the PyTorch implementations of PEGASUS and Legal-PEGASUS available from the Huggingface [98] library. For comparison to the PEGASUS results reported for Multi-LexSum in [18], we use the same hyperparameters values where provided: we train the models for 6 epochs with a learning rate of 5e-5, and for inference we use beam search with 5 beams and n-gram repetition blocks for n $>$3. For additional hyperparameters, we trained the models with a batch size of 4, 64 gradient accumulation steps, gradient checkpointing enabled, and a weight decay of 0.01. For our models at inference, we used a minimum of 24 tokens and maximum of 960 tokens for experimental settings with no entity chain, and a minimum of 34 tokens and maximum of 1154 for experimental settings including some form of entity chain, as these were the boundaries observed for our gold-standard data. We also added a length penalty of 2.0 to encourage the generation of long sequences, as [18] observed that PEGASUS undergenerated the number of words when producing short summaries for Multi-LexSum.

To ensure the validity of our code, we also attempted to replicate the original PEGASUS results reported in [18]; we note that our results differ slightly (see Table 5.1), falling behind the results reported in [18], which is likely due to incomplete knowledge of full hyperparameter settings used in the original work. However, as optimising hyperparameters falls outside the scope of our current research as we focus on the *impacts* our proposed improvements can bring rather than strictly on outperforming the state-of-the-art, comparison to our reproduction is sufficient to demonstrate the potential of our methods.

Chapter 5

Results and Analysis

We now present and analyse our experimental results.

5.1 RQ1 - Input Representation and Content Selection

To address RQ1, we must evaluate the quality of the summaries produced. We report ROUGE-1, ROUGE-2, and ROUGE-L scores (see Section 2.1), as is standard [18], using the implementation in the rouge-score[1] Python library. However, despite its popularity, the ROUGE metric has received various criticisms, both with respect to legal summarisation and more generally [2, 7, 9, 22, 43, 101, 102]. ROUGE assumes that a high quality summary uses the same words as the reference summary [19], and thus is not effective in cases of terminology variation or paraphrasing [101], failing to capture deeper semantic similarity [43]. Additionally, ROUGE considers all of the text equally, but in reality, only a small fraction of ngrams carry most of the semantic content [55]. Thus, following recent literature, we also report BERTScore [103] to capture the semantic similarity between generated and gold summaries without relying solely on lexical overlap [101]. We used the implementation provided in the bert-score[2] Python library, using the DeBERTA model for embedding for comparison with [18].

Table 5.1 shows ROUGE and BERTScore F1 scores for each of our six assessed input strategies, the PEGASUS and state-of-the-art results reported in [18], and our reproduction of the PEGASUS results in [18]. On all metrics apart from ROUGE-2, BERT-Windows was the most effective of the six tested input strategies - this is likely due to the balance of the number of relevant sentences included, and providing context for each sentence. First-1024, First-K, BERT-Sentences, and BERT-Windows

[1] https://pypi.org/project/rouge-score/
[2] https://pypi.org/project/bert-score/

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
First-1024	43.39	18.96	28.42	34.47
First-K	43.24	18.96	28.40	34.94
BERT-Sentences	43.61	19.33	27.58	34.52
BERT-Windows	44.17	19.28	28.53	35.62
BERT-Paragraphs	40.14	16.28	25.95	31.39
TextRank	42.36	17.23	27.31	33.45
PEGASUS [18] Reproduction	43.23	19.26	29.35	36.15
PEGASUS [18]	43.35	19.91	29.99	37.88
LED-16384 [18] (SOTA)	**46.54**	**22.08**	**31.91**	**40.00**

Table 5.1: Mean ROUGE and BERTScore F1 scores with respect to corresponding reference summary, with best PEGASUS-based results highlighted.

all outperform our reproduction of [18]'s PEGASUS strategy with respect to ROUGE-1, as expected, and BERT-Sentences and BERT-Windows outperform the reproduction baseline with respect to ROUGE-2. TextRank fails to outperform this baseline, which is consistent with its poor performance as a content selector for abstractive summarisation in [21]. BERT-Paragraph also fails to outperform this baseline, which is likely due to the fact that including longer context for each sentence means that information for fewer relevant sentences can be included within the token limit. Interestingly, none of our proposed strategies outperform the reproduction baseline on ROUGE-L or BERTScore metrics. We expected a greater improvement from the First-K baseline over the reproduction baseline, which intuitively should improve results as its only difference to the content selection strategy in [18] is the introduction of dataset cleaning and temporal ordering. We hypothesise therefore that the dataset filtering process may have resulted in decreased ROUGE scores[3], consistent with [37] - although this is likely to contribute to increased faithfulness. With respect to the PEGASUS results reported in [18], BERT-Windows, our most effective method, led to an improvement of 0.82 ROUGE-1. However, we did not observe improvements in comparison to the results of [18] with respect to other metrics.

Although not a realistic scenario, we also analysed the model's performance using oracle extracts from OREO as inputs, to ascertain the maximum possible improvements

[3]This was *not* due to the augmentation process - we performed an ablation study without dataset augmentation for the Lead-K baseline and achieved poorer results: ROUGE-1 F1 42.56, ROUGE-2 F1 18.41, ROUGE-L F1 27.93.

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
OREO-Sentences	**50.99**	**26.47**	**33.31**	**40.27**
OREO-Windows	47.97	23.28	31.55	38.92
OREO-Paragraphs	47.15	22.42	30.83	37.84

Table 5.2: Mean ROUGE and BERTScore F1 scores with respect to corresponding reference summary.

(i.e. - if the BERT salience classifier had 100% accuracy) which could be obtained when using a content selection strategy based on OREO. We found that for all three scenarios (sentences, windows, and paragraphs), the OREO inputs vastly outperformed their BERT counterparts, and all outperform the state of the art and thus also the PEGASUS results in [18] on ROUGE-1 and ROUGE-2. For ROUGE-L and BERTScore, the OREO-Windows and OREO-Paragraphs methods outperform the PEGASUS results in [18], and the OREO-Sentences strategy additionally outperforms the state of the art. As OREO-Sentences extracts typically consist of much fewer tokens (mean 264.15) than BERT-sentences (mean 1000.78), yet BERT-Sentences extracts have a greater ROUGE recall with the reference summary (see Section 4.4.3), this suggests that the specificity of the inputs provided to PEGASUS is key. Indeed, when measuring the mean of ROUGE-1 and ROUGE-2 precision between the OREO and BERT extracts used as input to PEGASUS with the gold summary, the OREO extracts display a greater precision (28.57 vs 7.67 for sentences, 10.68 vs 7.73 for windows, and 12.82 vs 12.15 for paragraphs). The increasing similarity in precision scores between OREO and BERT variants as the number of sentences for which information is included in the extracts decreases also suggests that the BERT classifier performs best for its high confidence outputs. To investigate this further, we investigated an alternative variant of BERT-Sentences where the same number of sentences as returned by OREO is returned (again, this is not feasible in real-world inference, as the oracle extracts and hence the number of sentences in the oracle extracts are not known). However, this led to worse results than the standard BERT-Sentences approach, again indicating that the classifier itself is an area for improvement - results are given in Appendix I. Overall, the OREO results suggest the promise of improved content selection and that research to further improve the salience classifier would be of great benefit.

Overall, we can establish that content selection does have the potential to improve summarisation outputs with respect to ROUGE and BERTScore, but that the classifier

	PEGASUS	Legal-PEGASUS	Legal-PEGASUS + Surface Form Chain	Legal-PEGASUS + Type Chain	Legal-PEGASUS + Combination Chain
BERT-Sentences	43.61 / 19.33 / 27.58 / 34.52	42.77 / 19.08 / 27.25 / 34.81	40.92 / 16.61 / 25.32 / 32.06	42.57 / 18.76 / 26.82 / 33.79	41.56 / 17.17 / 25.57 / 32.67
BERT-Windows	44.17 / 19.28 / 28.53 / 35.62	44.34 / 19.55 / 28.91 / 36.35	41.54 / 15.98 / 26.12 / 32.69	43.81 / 18.75 / 27.86 / 34.59	42.42 / 16.44 / 26.45 / 33.75
OREO-Sentences	50.99 / 26.47 / 33.31 / 40.27	**52.10 / 27.54 / 34.61 / 42.15**	48.71 / 22.89 / 31.14 / 38.27	50.92 / 26.40 / 33.05 / 40.67	49.91 / 24.27 / 31.45 / 38.97
OREO-Windows	47.97 / 23.28 / 31.55 / 38.92	48.41 / 23.72 / 31.91 / 39.44	45.57 / 19.92 / 29.12 / 36.13	47.27 / 22.69 / 30.49 / 37.60	45.86 / 20.21 / 28.94 / 36.19

Table 5.3: ROUGE-1 / ROUGE-2 / ROUGE-L / BERTScore F1 scores for different experimental setups.

performance limits these improvements in practice. As the sentence and window-based strategies offer the most promising results, we henceforth only report results for these strategies.

5.2 RQ2 - Domain-Specific Pretraining

ROUGE and BERTScore results for Legal-PEGASUS are given in Table 5.3. When introducing legal pretraining, we observe improvements in BERTScore for all input settings. For all settings apart from BERT-Sentences, we also observed improvements in ROUGE F1. Overall, our best results for the complete pipeline are given by BERT-Windows. However, these results still only outperform the PEGASUS results reported in [18] with respect to ROUGE-1 (by 0.99 F1). In contrast, OREO-Sentences further outperforms the state of the art, achieving an improvement of 5.56 ROUGE-1 F1, 5.46 ROUGE-2 F1, 2.7 ROUGE-L F1, and 2.15 BERTScore. Overall, we observed greater improvements for better (in terms of vanilla PEGASUS ROUGE-1 and BERTScore) content selection strategies (with performance for BERT-Sentences actually decreasing). Overall, our results again indicate the importance of content selection.

5.3 RQ3 - Faithfulness and Entity Chaining

Lexical overlap based metrics such as ROUGE, while appropriate to some extent for capturing general summary content [62], do not correlate well with human judgements of faithfulness [32, 62, 72]. An unfaithful summary hypothesis can still achieve a high

ROUGE score [54] - for example, the sentences 'I am studying in Edinburgh' and 'I am not studying in Edinburgh' share nearly all unigrams and bigrams despite having opposite semantic polarity. For this reason, several metrics have been proposed to evaluate the faithfulness of summaries to their source text. However, not all have been shown to give meaningful results, especially in the legal domain [2].

Following [36, 57], we evaluate faithfulness using textual entailment. Various works have demonstrated a correlation between entailment scores and human judgements of faithfulness [36, 56, 58, 73, 78, 104], but it is notable that entailment has been found to underpredict faithfulness scores compared to human judgements [28]. We report the probability of a generated summary (PEGASUS output) being entailed by its source text (PEGASUS input) returned by a BART-large classifier finetuned on Multi-NLI [56]. As BART has the same maximum token length as PEGASUS, no chunking is required as the source text can be considered at once. We use the implementation provided by [56], using the document-to-sentence mode, which returns the average probability across all summary sentences of each sentence being entailed by the source text. However, despite entailment being the most promising option, the suitability of (particular) automatic metrics to assess faithfulness is still a subject of research debate. While human evaluation is the ideal [7], this has infeasible time and cost requirements [55], especially for long documents in the legal domain [19] - we would require highly trained annotators who are willing to invest significant amounts of time [2].

ROUGE and BERTScore results when including entity chains are given in Table 5.3. Overall, all entity chain variants led to decreases in ROUGE and BERTScore metrics, with the surface form type chains leading to the most drastic decreases, type chains decreasing summary quality metrics the least, and the combination chains (as would be expected) falling between. The introduction of entity chaining leading to decreased ROUGE scores is consistent with [37]. With respect to entailment (Table 5.4), we find the inclusion of the type chains to improve faithfulness over the Legal-PEGASUS only baseline. The combination chains improved faithfulness in for all input variants apart from BERT-sentences. The surface-form chains, which were the form of entity chain investigated in [37] and [65], decreased entailment in all cases. However, this is consistent with [65], who observed that while entailment decreased, human faithfulness evaluations increased. Surprisingly, Legal-PEGASUS reduced entailment scores in all experimental settings compared to vanilla PEGASUS. Overall, BERT-based methods always had higher faithfulness than their OREO-based counterparts, and window-based methods had higher faithfulness than sentence-based methods. This

	PEGASUS	Legal-PEGASUS	Legal-PEGASUS + Surface Form Chain	Legal-PEGASUS + Type Chain	Legal-PEGASUS + Combination Chain	Reference
BERT-Sentences	0.5134	0.4954	0.4852	0.5106	0.4952	0.3582
BERT-Windows	0.5551	0.5551	0.5423	0.5650	0.5578	0.3630
OREO-Sentences	0.4915	0.4680	0.4406	0.4920	0.4759	0.2122
OREO-Windows	0.5457	0.5469	0.5304	0.5587	0.5492	0.3727

Table 5.4: Mean probability of generated summary being entailed by its corresponding source text (using BART-large finetuned on Multi-NLI) for different experimental setups.

may be due to BERT-based methods and window-based methods having longer source texts. Our results agree with the literature [32, 62, 72] that ROUGE does not correlate to faithfulness; although OREO-Sentences receives the worst entailment scores, this method performs best on ROUGE and BERTScore.

While we do not have entailment scores for the exact PEGASUS setup in [18] as we do not have access to the original model outputs, and we acknowledge that our reproduction leads to slightly different results, in all cases it is evident that all our experimental setups vastly improve the probability of the source text entailing the summary text in comparison for this reproduction baseline (mean entailment probability 29.37). Once again, this suggests the importance of content selection.

We also calculated the probability that the each of the input strategies entailed the reference summary; these scores were low, indicating that the fundamentally ROUGE based [26] content selection strategies used are perhaps not retrieving the most salient information in relation to the reference summaries.

We note that while we did calculate the entity precision metric (the percentage of entities in the summary that occur in the source text) proposed in the entity faithfulness evaluations of [65] (FROST) and [37] (JAENS), due to the matching issues discussed in Section 4.1.3 (and by [37] - who noted that both the matching heuristics used and the NER system itself can lead to inaccuracy in the metrics), we did not find these results to be reliable based on manual inspection. In addition, entity-specific metrics also do not capture the relations between or context surrounding entities, which influence overall faithfulness. For this reason, together with a greater robustness to lexical variability than matching-based approaches, we focus on entailment as our automatic measure of summary faithfulness and manually analyse issues of entity hallucination for a sample of results in Section 5.5. Regardless, we report entity precision in Appendix J for completeness' sake, and while the exact figures themselves are unreliable, the results nevertheless appear to suggest similar trends to those observed using entailment.

It is also worth discussing why we did not assess faithfulness using other metrics:

- BERTScore between the source text and generated summary as a faithfulness metric has been shown to be poor at handling negations and numerical values [56], which are especially important in legal text and for three of the entity types we use in the construction of entity chains (DATE, MONEY, LAW).
- FactCC [73], a trained metric returning a binary prediction of whether a sentence is faithful to a source [53], is finetuned on synthetically hallucinated summaries [58] using rule based transformations [73] which are not necessarily reflective of real hallucination patterns [58]. FactCC is also not well suited to highly abstractive summaries [53] such as those in Multi-LexSum, and in some experiments has been found to have no correlation with human faithfulness judgements at all [28].
- Information extraction (IE) based metrics have been proposed to tackle the relation hallucination problem, where entities from source document appear in the wrong relation in the generated text [56]. However, IE models are not yet sufficiently reliable, even in the general domain [54, 56], and this approach is not feasible for long source texts due to the huge number of facts that would be extracted.
- Question answering (QA) metrics such as QAGS [55] and QUALS [53] are based on the intuition that if we ask questions about a summary and its source, we will receive similar answers if the summary is faithful [24]. Such strategies rely on the quality of question generation and question answering models [53, 55] which, as mentioned in Section 2.3, are not yet sufficiently robust in the legal domain. The matching of answers is also an issue - synonyms, generalisations, and abbreviations have all been shown to be problematic [28, 56]. Additionally, experimental results showing a similar correlation to human faithfulness judgements suggest that QA methods could simply be an overly complex method of entity comparison, as answers are primarily named entities [28].

5.4 RQ4 - Readability

Metrics such as ROUGE do not account for linguistic qualities such as human readability [102]. We evaluate readability using common readability metrics, following [12]:

- Flesch-Kincaid formula - dependant on the number of words in a sentence and the number of syllables per word [12]; used by Various U.S. states to score financial forms and legal documents such as insurance policies [105].

- Coleman-Liau index - dependant on the number of letters per 100 words and the average number of sentences per 100 words [12].
- SMOG - dependant on the number of polysyllable words per sentence [12]; used in healthcare, another specialised domain [105].
- Automated readability index (ARI) - dependant on the number of characters per word and number of words per sentence [12].

Each of these metrics returns the 'minimum reading age' of the text, so a higher score indicates lower readability. We use the implementations available here[4].

	Source Documents (Test Set)	Reference Summaries (Test Set)	Reproduction of [18]
Flesch Kincaid	19.57	17.72	16.46
Coleman Liau	15.66	18.46	17.15
ARI	22.31	20.92	19.61
SMOG	19.06	19.08	17.57

Table 5.5: Mean readability scores for Flesch Kincaid, Coleman Liau, ARI, and SMOG.

	PEGASUS	Legal-PEGASUS	Legal-PEGASUS + Surface Form Chain	Legal-PEGASUS + Type Chain	Legal-PEGASUS + Combination Chain
BERT-Sentences	17.78 / 17.77 / 20.99 / 18.74	17.84 / 17.91 / 21.08 / 18.88	17.68 / 17.92 / 20.98 / 18.68	18.49 / 17.92 / 21.86 / 19.23	17.93 / 17.90 / 21.26 / 18.92
BERT-Windows	17.72 / 17.71 / 20.97 / 18.58	17.65 / 17.81 / 20.88 / 18.65	17.45 / 17.97 / 20.75 / 18.50	18.47 / 17.76 / 21.83 / 19.08	18.07 / 18.08 / 21.43 / 18.95
OREO-Sentences	16.87 / 17.40 / 19.79 / 18.06	16.97 / 17.64 / 19.93 / 18.22	16.88 / 17.59 / 19.87 / 18.15	17.62 / 17.62 / 20.70 / 18.59	17.25 / 17.64 / 20.18 / 18.47
OREO-Windows	17.80 / 17.75 / 21.04 / 18.60	17.83 / 17.84 / 20.99 / 18.69	17.69 / 18.02 / 20.97 / 18.66	18.44 / 17.81 / 21.73 / 19.07	17.86 / 17.92 / 21.09 / 18.89

Table 5.6: Mean readability scores for Flesch Kincaid, Coleman Liau, ARI, and SMOG.

Our summaries improve readability over the original documents in all considered metrics apart from Coleman-Liau, suggesting an increase in accessibility. However, the improvements are not to the extent found in [12]; in addition to the fundamental characteristics of the text, a potential reason for this is the high level of extractivity observed in Section 5.5. Decreased readability in the Coleman-Liau index indicates the type of linguistic complexity present in our summaries: long words and sentences [106]. The readability scores of the generated summaries are broadly in line with the reference summaries; as the reference summaries are known to be written in accessible language

[4]https://github.com/wimmuskee/readability-score

[18], this suggests that our generated summaries are of an acceptable readability. We note that the readability of the summaries generated by the PEGASUS reproduction [18] baseline are lower than for the reference summaries in all cases.

Analysing the difference in readability between our experimental settings, we observe that the OREO-Sentences input representation produces the most readable summaries, and in all cases, the summaries produced when including type chains had highest Flesch Kincaid and SMOG and ARI, indicating a lower degree of readability. We note that this is also the output representation giving the highest textual entailment scores, indicating that a higher degree of extractivity may be occurring in these settings.

5.5 Qualitative Analysis

We noted previously that evaluation of the summaries with human participants is infeasible, however to better understand our models' behaviour and failure modes, we manually analysed generated summaries for a sample of 10 cases across model settings. We present example summaries in Figure 5.1; further examples are given in Appendix K. In general, the outputs of the reproduction of the PEGASUS method in [18] were comparatively good at reproducing the correct date when the case began, as this is frequently mentioned in headers at the start of the document. Background information for the case (which often occurs at the start of the initial complaint document) are also reflected fairly reliably. However, the outputs often hallucinate the law which is alleged to be violated, which is extremely vital information, and struggle to accurately represent the case's process and procedure. This is likely because this information is not included in the source text captured using the authors' content selection strategy.

In contrast, our models generally produce longer summaries which better match the reference summaries in overall content - this is not reflected by the ROUGE results. In general, our models perform well for the background and laws involved in the case, but performance declines for a case's procedural actions; key information is often missed, and the models often fail to provide the reasoning for decisions made. This may be because these aspects involve a higher degree of understanding and assimilation of information, and also follow a less standard format so are less easily identified by the BERT classifier, as noted in [43]. While our models generally contain less hallucinatory content than our reproduction of [18]'s approach, two common hallucination scenarios remain. Firstly, dates and monetary amounts which truly occur in the source text are often contained in the summary in the incorrect context; such intrinsic hallucination is

less suited to entity-focused methods like entity chaining. The second frequent scenario for hallucination stems from issues in relation to case understanding. A notable subtype of this error case is the model outputs including information from *cited cases* as if it pertains to the main case under discussion. This mistake is made when using both BERT and OREO inputs; this may be because discussions of cited cases often include a high density of common legal keywords. As the BERT and OREO methods both make this mistake, this suggests that selecting relevant sentences may be more suited to human annotation than automatic overlap-based methods; while full-scale human annotation would be infeasible, as discussed in Section 4.4.1, semi-supervised approaches, which have been applied with success in other areas of legal AI [107], may be promising. At the BERT classifier stage, including context for the sentence under consideration may help to distinguish between information relating to main or cited cases.

The planning methods appeared to have little effect with respect to summary content, with outputs often being very similar across experimental scenarios for the same input strategy, suggesting that the content selection component largely dictates summary content, and that our model's improved faithfulness compared to the baseline is likely due to improved content selection. We also note that for settings including entity chains, the chains and summaries are often not tightly coupled - not all entities in the chain occur in the summary (and if they do, not necessarily in the same order), and not all entities in the summary appear in the chain. [65] proposed checklist models [108] and entity-chain constrained decoding [109] as possible strategies to deal with this issue. In rare cases, entity chains fail to generate at all.

Another issue with our models is the tendency to include large extractive fragments from the source text, as evidenced by artefacts (such as numerals) remaining after the cleaning process being reproduced in the generated summaries. This extractivity limits the readability of the summaries in some cases by replicating complex legal terminology and syntax from the source text. We hypothesise that this high degree of extractivity may be due to limited text occurring in the input text for each point, and that these fragments are not well-flowing text, which models such as PEGASUS are trained on.

The points observed are consistent across input strategies, including OREO-based strategies. Overall, our method demonstrates an improvement in faithfulness compared to our reproduction of the PEGASUS method in [18], although our results suggest that these improvements are largely due to content selection. Faithfulness issues and issues in terms of readability and case understanding persist, indicating that legal abstractive summarisation requires further research to be applicable in the real world.

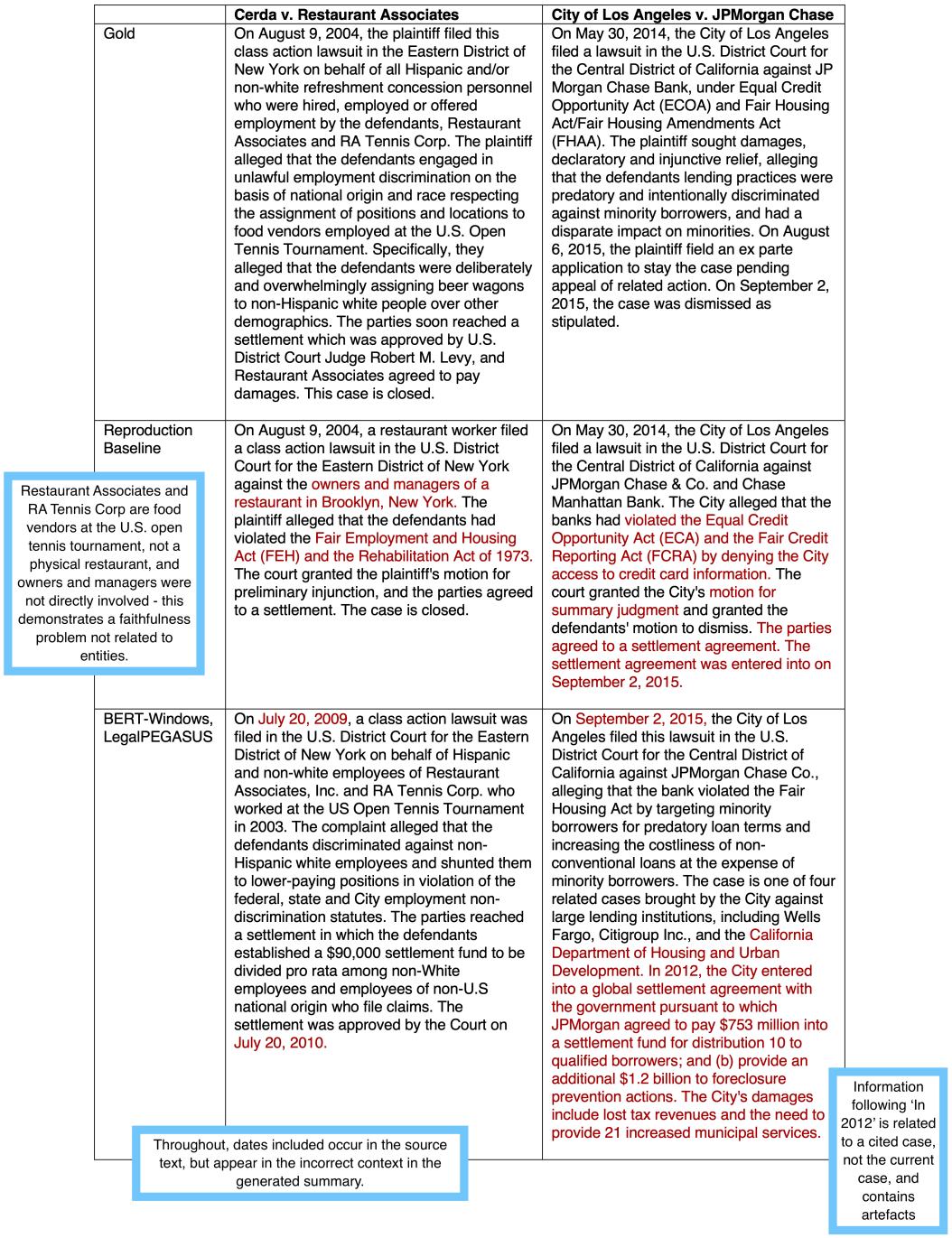


Figure 5.1: Annotated examples of representative model outputs for two cases, with facts inconsistent with the case documents and other errors (such as assimilating information from cited cases) highlighted in red.

Chapter 6

Discussion and Conclusion

We have conducted the first study at the intersection of legal, multi-document, and faithful summarisation. We investigated the impact of content selection, legal pretraining, and the entity chain planning mechanism on the abstractive summarisation of U.S. civil rights litigation, using PEGASUS as our backbone model. Our full test-time pipeline outperforms the PEGASUS results in [18] by 0.99 ROUGE-1 F1. Furthermore, although not realistic, we show that using oracle extracts vastly outperforms the state-of-the-art, with legal pretraining further boosting results: we achieve an improvement of 5.56 ROUGE-1 F1, 5.46 ROUGE-2 F1, 2.7 ROUGE-L F1, and 2.15 BERTScore. Overall we provide evidence that more sophisticated content selection improves summary faithfulness and quality, that legal pretraining can further boost results when an effective input representation is used, and that the summaries generated by our models improve readability compared to the original judicial documents. While our model outputs demonstrate a clear improvement over our baseline, the generated summaries can fail to cover relevant information and still contain hallucinations which are not adequately addressed by introducing entity chaining as a planning mechanism without additional modifications. As noted in Section 5.5, several issues, including the quality of the content selection method and addressing particular hallucination scenarios, remain to be addressed for such automatic summarisation to see real-world adoption.

Our project is limited by the resources available. Firstly, while presenting a realistic summarisation scenario, the OCR process used to construct the Multi-LexSum dataset from the original court documents introduces noise, which despite dataset cleaning, can adversely affect both summarisation outputs and the metrics used to evaluate these outputs. The availability of computational resources also limits the range of experiments that can be conducted and the hyperparameter settings used. Perhaps most importantly,

the lack of a thorough human evaluation of our models' outputs by domain experts limits our interpretation of our findings, as metrics such as ROUGE and entailment are only *proxies* for summary quality and faithfulness. This factor is especially important in the legal domain, where a lack of correlation between automatic metrics and human expert judgements has previously been demonstrated [2, 7], and as the utility of automatic metrics to judge faithfulness remains a topic of research debate.

Our work also has limitations pertaining to the intended use case of legal summarisation - namely, by legal professionals or ordinary civilians without resources or expertise in machine learning. The powerful GPUs required for finetuning and performing inference for the transformer models used throughout our pipeline are unlikely to be available in non-academic environments. Furthermore, our methodology's performance on other datasets (for example - for other legal areas, jurisdictions, or languages) has not yet been tested. Overall, our work certainly contributes towards a legal summarisation system which will benefit the relevant stakeholders, by providing a trustworthy methodology for the summarisation of realistic datasets, which improves summary quality and faithfulness. However, issues of generalisation and GPU requirements, in addition to the limitations of our model outputs, mean that further research is required to develop a summarisation solution which could bring widespread benefit to legal practitioners.

Our project's limitations and error cases suggest several future research directions. Our results strongly suggest that further research into content selection is promising - investigating methods of content selection not fundamentally based on ROUGE, such as using human salience annotations in a semi-supervised framework, could be fruitful. With respect to entity chaining, implementing strategies such as checklist models [108] or entity-chain constrained decoding [109] could improve the correspondence between entity chains and their associated summaries. This would then enable entity chains to more effectively control the generated output, either by introducing mechanisms such as FROST++'s [65] drop-prompt (which would also require additional work on entity matching heuristics), or by modifying the plan to provide summaries related to user queries in a larger retrieval system (as topical diversity and emphasis can be achieved by modifying which entities are included and their order [37, 65]), for example. More generally, the application of our generated summaries as the input to other legal NLP tasks could be interesting. In addition, future work could conduct similar investigations to ours on different legal domains and jurisdictions, or using different backbone models - for example, it would be interesting to observe the effect of content selection on models able to handle longer input texts, such as LED, or GPT-based models.

Bibliography

[1] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online, July 2020. Association for Computational Linguistics.

[2] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 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, November 2022. Association for Computational Linguistics.

[3] A. A. Askari, S. V. Verberne, O. Alonso, S. Marchesin, M. Najork, and G. Silvello. Combining lexical and neural retrieval with longformer-based summarization for effective case law retrieva. In *Proceedings of the second international conference on design of experimental search & information REtrieval systems*, pages 162–170. CEUR, 2021.

[4] Sophia Althammer, Arian Askari, Suzan Verberne, and Allan Hanbury. DoSSIER@COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval, August 2021. arXiv:2108.03937 [cs].

[5] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Emre Mumcuoğlu, Ceyhun E. Öztürk, Haldun M. Ozaktas, and Aykut Koç. Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Information Processing and Management: an International Journal*, 58(5), September 2021.

[7] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 413–428, Cham, 2019. Springer International Publishing.

[8] Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, Raghav R, and Roshni Kar. An Evaluation Framework for Legal Document Summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753, Marseille, France, June 2022. European Language Resources Association.

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

[10] Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law*, 27(2):141–170, June 2019.

[11] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388, May 2021.

[12] Laura Manor and Junyi Jessy Li. Plain English Summarization of Contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Rupert Haigh. *Legal English*. Routledge, London, 5 edition, June 2018.

[14] Peter M. Tiersma. *Legal Language*. University of Chicago Press, Chicago, IL, May 2000.

[15] Christopher Williams. *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*. Cambridge University Press, May 2007. Google-Books-ID: HwF3XIuLflYC.

[16] Diego Feijo and Viviane Moreira. Summarizing Legal Rulings: Comparative Experiments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 313–322, Varna, Bulgaria, September 2019. INCOMA Ltd.

[17] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics.

[18] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, June 2022.

[19] Diego de Vargas Feijo and Viviane P. Moreira. Improving abstractive summarization of legal rulings through textual entailment. *Artificial Intelligence and Law*, 31(1):91–113, March 2023.

[20] Anastassia Kornilova and Vladimir Eidelman. 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, November 2019. Association for Computational Linguistics.

[21] Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, and Andrew McCallum. Long Document Summarization in a Low Resource Setting using Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research*

Workshop, pages 71–80, Online, August 2021. Association for Computational Linguistics.

[22] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. Bayesian Optimization based Score Fusion of Linguistic Approaches for Improving Legal Document Summarization. *Knowledge-Based Systems*, 264:110336, March 2023.

[23] Joel Niklaus and Daniele Giofre. Can we Pretrain a SotA Legal Language Model on a Budget From Scratch? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 158–182, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics.

[24] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey, April 2023. arXiv:2104.14839 [cs].

[25] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–7, New York, NY, USA, May 2021. Association for Computing Machinery.

[26] Yumo Xu and Mirella Lapata. Text Summarization with Oracle Expectation, September 2022. arXiv:2209.12714 [cs].

[27] Abigail See, Peter J. Liu, and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks, April 2017. arXiv:1704.04368 [cs].

[28] Tim Fischer. Finding Factual Inconsistencies in Abstractive Summaries. Master's thesis, Universität Hamburg, June 2021.

[29] Yang Liu and Mirella Lapata. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics.

[30] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), February 2017. Number: 1.

[31] Yue Dong. A Survey on Neural Network-Based Summarization Methods, March 2018. arXiv:1804.04589 [cs].

[32] Zheng Zhao, Shay B. Cohen, and Bonnie Webber. Reducing Quantity Hallucinations in Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online, November 2020. Association for Computational Linguistics.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[34] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551, January 2020.

[36] Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1319–1339, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[37] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online, April 2021. Association for Computational Linguistics.

[38] Vedant Parikh, Vidit Mathur, Parth Mehta, Namita Mittal, and Prasenjit Majumder. LawSum: A weakly supervised approach for Indian Legal Document Summarization, October 2021. arXiv:2110.01188 [cs].

[39] M. Saravanan and B. Ravindran. Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artificial Intelligence and Law*, 18(1):45–76, March 2010.

[40] Atefeh Farzindar and Guy Lapalme. Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles. In *Text Summarization Branches Out*, pages 27–34, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[41] Chao-Lin Liu and Kuan-Chun Chen. Extracting the Gist of Chinese Judgments of the Supreme Court. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, pages 73–82, New York, NY, USA, June 2019. Association for Computing Machinery.

[42] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. CaseSummarizer: A System for Automated Summarization of Legal Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[43] Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D. Ashley, and Matthias Grabmair. Automatic Summarization of Legal Decisions using Iterative Masking of Predictive Sentences. In *Proceedings of the Seventeenth International*

Conference on Artificial Intelligence and Law, ICAIL '19, pages 163–172, New York, NY, USA, June 2019. Association for Computing Machinery.

[44] Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pages 22–31, New York, NY, USA, July 2021. Association for Computing Machinery.

[45] Ambedkar Kanapala, Srikanth Jannu, and Rajendra Pamula. Summarization of legal judgments using gravitational search algorithm. *Neural Computing and Applications*, 31(12):8631–8639, December 2019.

[46] Filippo Galgani, Paul Compton, and Achim Hoffmann. Summarization based on bi-directional citation analysis. *Information Processing & Management*, 51(1):1–24, January 2015.

[47] Svea Klaus, Ria Van Hecke, Kaweh Djafari Naini, Ismail Sengor Altingovde, Juan Bernabé-Moreno, and Enrique Herrera-Viedma. Summarizing Legal Regulatory Documents using Transformers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2426–2430, New York, NY, USA, July 2022. Association for Computing Machinery.

[48] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel 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*, ICAIL '21, pages 159–168, New York, NY, USA, July 2021. Association for Computing Machinery.

[49] Gianluca Moro and Luca Ragazzi. Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11085–11093, June 2022. Number: 10.

[50] Ahmed Elnaggar, Christoph Gebendorfer, Ingo Glaser, and Florian Matthes. Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification. In *Proceedings of the 2018 Artificial Intelligence and*

Cloud Computing Conference, AICCC '18, pages 9–15, New York, NY, USA, December 2018. Association for Computing Machinery.

[51] Huihui Xu, Jaromir Savelka, and Kevin 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*, ICAIL '21, pages 250–254, New York, NY, USA, July 2021. Association for Computing Machinery.

[52] Mohamed Elaraby and Diane Litman. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[53] Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. Improving Factual Consistency of Abstractive Summarization via Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online, August 2021. Association for Computational Linguistics.

[54] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38, March 2023.

[55] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics.

[56] Tim Fischer, Steffen Remus, and Chris Biemann. Measuring Faithfulness of Abstractive Summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany, 2022. KONVENS 2022 Organizers.

[57] Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Dipanjan Das, and Mirella Lapata. Conditional Generation with a Question-Answering Blueprint, July 2022. arXiv:2207.00397 [cs].

[58] Arvind Krishna Sridhar and Erik Visser. Improved Beam Search for Hallucination Mitigation in Abstractive Summarization, December 2022. arXiv:2212.02712 [cs].

[59] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[60] Yue Dong, John Wieting, and Pat Verga. Faithful to the Document or to the World? Mitigating Hallucinations via Entity-Linked Knowledge in Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[61] Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. Improving Truthfulness of Headline Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online, July 2020. Association for Computational Linguistics.

[62] Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, and Alexander Gray. X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[63] Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics*

(GEM), pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[64] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pages 4784–4791, New Orleans, Louisiana, USA, February 2018. AAAI Press.

[65] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. Planning with Learned Entity Prompts for Abstractive Summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492, December 2021.

[66] Dennis Aumiller, Ashish Chouhan, and Michael Gertz. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[67] Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. Analyzing Sentence Fusion in Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China, November 2019. Association for Computational Linguistics.

[68] Dongqi Pu, Yifan Wang, and Vera Demberg. Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada, July 2023. Association for Computational Linguistics.

[69] Kazem Taghva, Tom Nartker, Allen Condit, and Julie Borsack. Automatic Removal of "Garbage Strings" in OCR Text: An Implementation.

[70] Michael James Bommarito, Daniel Martin Katz, and Eric Detterman. LexNLP: Natural Language Processing and Information Extraction For Legal and Regulatory Texts, June 2018.

[71] Claire Barale, Michael Rovatsos, and Nehal Bhuta. Automated Refugee Case Analysis: A NLP Pipeline for Supporting Legal Practitioners. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2992–3005, Toronto, Canada, July 2023. Association for Computational Linguistics.

[72] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online, June 2021. Association for Computational Linguistics.

[73] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics.

[74] Willem J. M. Levelt. *Speaking: From Intention to Articulation*. The MIT Press, August 1993.

[75] Markus Guhe. *Incremental Conceptualization for Language Production - 1st Edition* -. Lawrence Erlbaum Associates Publishers, 2007.

[76] Steven Bird and Edward Loper. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[77] Wenmian Yang, Weijia Jia, XIaojie Zhou, and Yutao Luo. Legal Judgment Prediction via Multi-Perspective Bi-Feedback Network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4085–4091, August 2019. arXiv:1905.03969 [cs].

[78] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[79] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke 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*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[80] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 11328–11339. JMLR.org, July 2020.

[81] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. MultiLegalPile: A 689GB Multilingual Legal Corpus, June 2023. arXiv:2306.02069 [cs].

[82] Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada, July 2023. Association for Computational Linguistics.

[83] Jiaru Jia, Xin Chen, Aiqing Yang, Qiulin He, Pengyu Dai, and Mingzhe Liu. Link of Transformers in CV and NLP: A Brief Survey. In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 735–743, August 2022.

[84] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. arXiv:2004.05150 [cs].

[85] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. LongT5: Efficient Text-To-Text Transformer for Long Sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States, July 2022. Association for Computational Linguistics.

[86] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang,

and Amr Ahmed. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020.

[87] Ayesha Sarwar, Seemab Latif, Rabia Irfan, Adnan Ul-Hasan, and Faisal Shafait. Text Summarization from Judicial Records using Deep Neural Machines. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6, November 2022.

[88] Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. Summˆ N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[89] Yen-Chun Chen and Mohit Bansal. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[90] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by Summarizing Long Sequences, January 2018. arXiv:1801.10198 [cs].

[91] Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online, November 2020. Association for Computational Linguistics.

[92] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[93] Duy Duc An Bui, Guilherme Del Fiol, John F. Hurdle, and Siddhartha Jonnalagadda. Extractive text summarization system to aid data extraction from

full text in systematic review development. *Journal of Biomedical Informatics*, 64:265–272, December 2016.

[94] Hongyan Jing and Kathleen R. McKeown. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 129–136, New York, NY, USA, August 1999. Association for Computing Machinery.

[95] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics.

[96] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. Conference Name: IEEE Access.

[97] Yang Liu and Mirella Lapata. Hierarchical Transformers for Multi-Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy, July 2019. Association for Computational Linguistics.

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

[99] Jayawant N. Mandrekar. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, September 2010.

[100] M. Saravanan, B. Ravindran, and S. Raman. Improving Legal Document Summarization Using Graphical Models. In *Proceedings of the 2006 conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, pages 51–60, NLD, June 2006. IOS Press.

[101] Arman Cohan and Nazli Goharian. Revisiting Summarization Evaluation for Scientific Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813. European Language Resources Association (ELRA), May 2016.

[102] Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. Single Document Summarization based on Nested Tree Structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[103] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, February 2020. arXiv:1904.09675 [cs].

[104] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics.

[105] Carmine DiMAscio. py-readability-metrics: Score text "Readability" with popular formulas and metrics including Flesch-Kincaid, Gunning Fog, ARI, Dale Chall, SMOG, Spache and more.

[106] Meri Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975. Place: US Publisher: American Psychological Association.

[107] K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, and A. Yeh. Semi-Supervised Methods for Explainable Legal Prediction. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence*

and Law, ICAIL '19, pages 22–31, New York, NY, USA, June 2019. Association for Computing Machinery.

[108] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, Texas, November 2016. Association for Computational Linguistics.

[109] Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. Constrained Abstractive Summarization: Preserving Factual Consistency with Constrained Generation, December 2021. arXiv:2010.12723 [cs].

Appendix A

Legal Document Types

Here we present an overview of the types of documents occurring in the Multi-LexSum dataset, adapted from [18].

Document Type	Description
Complaint	The document that starts a case, usually the first thing filed. Plaintiffs can also file amended complains to add or subtract parties or claims.
Opinion/Order	Created by judges, opinions or orders memorialise rulings in the case.
Pleading/Motion/Brief	Broadly covers documents filed by the parties in order to make requests or explain their arguments.
Monitor/Expert/Receiver Report	Reports created by non-parties to help with the litigation in various ways. A "monitor" is a court-appointed expert, usually superintending compliance with a court order; an "expert" works for one or the other side, during the litigation; a "receiver" is an entity appointed by the court to run defendant operations because the defendant has somehow demonstrated incapacity.
Settlement	An agreement among parties that resolves some or all of the issues in the lawsuit.
Press Release	A press release.
Docket	The court's index of everything that has happened in a case, in that court.
Correspondance	Letters NOT directed to the court. In some jurisdictions (particularly in New York City), parties will conduct lots of litigation through letters to the court or "letter motions"—these are classified as motions or briefs, not as correspondence).
Declaration/Affidavit	Documents in which someone provides information under penalty of perjury.
Discovery/FOIA Material	Discovery material is evidence turned over by one party to another. FOIA materials are documents produced in response to a Freedom of Information Act (FOIA) request.
FOIA Request	A request for information under the Freedom of Information Act or a state equivalent.
Internal Memorandum	An organization's internal memo (different from litigation documents with "memorandum" in the title).
Magistrate Report/Recommendation	Decisions from magistrate judges.
Statute/Ordinance/Regulation	A law or rule of government entity—federal, state, city or county, or agency. This document type includes policies created by prisons, school districts, police departments, immigration authorities, etc.
Transcripts	Verbatim transcripts of court proceedings or depositions.

Table A.1: Document types represented in Multi-LexSum.

Appendix B

Summary Granularities

Here we present examples of the three summary granularities in Multi-LexSum: long, short, and tiny.

Source Input Excerpt ... And, even if the agency had made an internal decision to maintain the status quo, the documents at issue would not lose their predecisional status because plaintiff has not shown that they have been "adopted, formally or informally, as the agency position on an issue or is used by the agency in its dealings with the public."1 Coastal States Gas Corp., 617 F.2d at 866; Sears, 421 U.S. at 161 ("[I]f an agency chooses expressly to adopt or incorporate by reference an intraagency memorandum previously covered by Exemption 5 in what would otherwise be a final opinion" that memorandum may not be withheld under Exemption 5). Plaintiff does not point to any public statements that OMB has made referencing, adopting, or incorporating the records or the subject matter at issue, nor has plaintiff provided the Court with any evidence that the records were informally adopted as the agency's position. Plaintiff references a statement made by Karen Battle, chief of the Census Bureau's Population Division, on January 26, 2018, where she explained that additional research and testing were necessary before the Census Bureau could proceed to implement a separate Middle Eastern or North African category. Pl.'s Cross-Mem. at 13. Plaintiff argues that "[t]o the extent that Ms. Battle's explanation about the need for more research, and indeed the entire underlying decision to maintain the status quo, is evidenced in the withheld documents, it has been adopted as the agency's policy." Id. But, this statement was made by a Census Bureau official, not an OMB official. And, in any event, the statement 1 Courts in this district have held that the plaintiff carries the burden to show that the agency has formally or informally adopted a record as policy. See, e.g., Heffernan v. Azar, 317 F. Supp. 3d 94, 122 (D.D.C. 2018), citing Sec. Fin. Life Ins. Co. v. U.S.

Dep't of Treasury, No. 03-102, 2005 WL 839543, at *7 (D.D.C. Apr. 12, 2005). ...

Long Summary: On April 13, 2018, the Arab American Institute ("AAI") sued the Office of Management and Budget ("OMB") under the Freedom of Information Act ("FOIA"), 5 U.S.C. § 552, in the U.S. District Court for the District of Columbia. AAI alleged that OMB violated FOIA by failing to disclose requested records pertaining to OMB's decision not to include a combined race and ethnicity question or a Middle Eastern or North African (MENA) category on the 2020 Census. AAI asked the court to declare that OMB violated FOIA and to issue an injunction ordering the agency to release the requested records. This case was assigned to Judge Amy Berman Jackson One month later, on May 18, 2018, the court ordered OMB to file a dispositive motion or a status report setting a schedule for OMB's production of documents to AAI. OMB chose the latter, filing its first status report on June 15, 2018. Over the next two years, the parties filed several joint status reports detailing which documents OMB had disclosed to AAI and which documents were still outstanding or disputed. By May 13, 2020, OMB had reviewed approximately 2,000 potentially responsive documents, producing "a number" of them to AAI and withholding 161 of them, claiming they were FOIA exempt. AAI objected to the withholding of five of the allegedly exempt documents. OMB filed a motion for summary judgment on February 10, 2020, arguing that the five disputed documents were exempt under FOIA Exemption 5, which allows agencies to withhold "inter-agency or intra-agency memorandums or letters that would not be available by law to a party other than an agency in litigation with the agency," including "predecisional and deliberative" documents that reflect internal Executive Branch deliberations. AAI filed a cross-motion for summary judgment on March 12, 2020, arguing that OMB had not provided a sufficient basis for exempting the documents and that the exemption didn't apply because the documents were not "predecisional." On August 13, 2020, after conducting in camera review, the court granted OMB's motion for summary judgment and denied AAI's cross-motion, finding that the disputed documents were predecisional and exempt from FOIA. 2020 WL 4698098. As of December 25, 2020, AAI has not appealed the court's decision.

Short Summary: On April 13, 2018, the Arab American Institute sued the Office of Management and Budget under the Freedom of Information Act in the U.S. District Court for the District of Columbia. AAI alleged that OMB violated FOIA by failing to disclose requested records pertaining to OMB's decision not to include a combined race and ethnicity question or a Middle Eastern or North African (MENA) category on the 2020 Census. In May, the court ordered OMB to file a dispositive motion or a

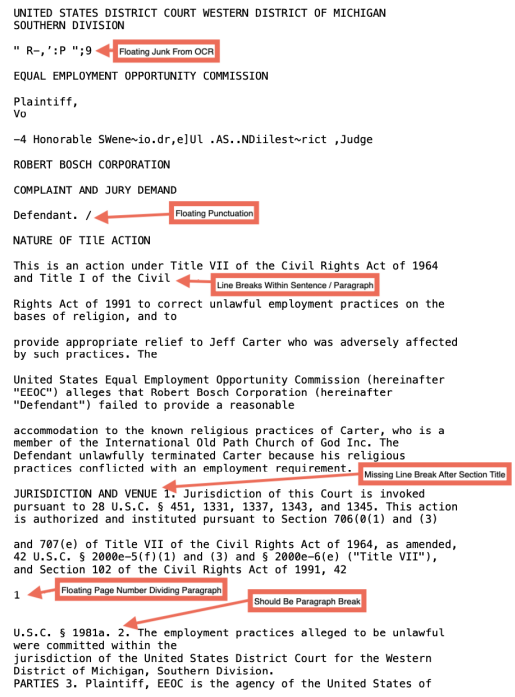
status report setting a schedule for OMB's production of documents to AAI. Over the next two years, the parties filed several joint status reports detailing which documents OMB had disclosed to AAI and which documents were still outstanding or disputed. OMB produced a number of documents to AAI but withheld some, claiming they were FOIA exempt. AAI objected to five claimed exemptions. The parties both filed motions for summary judgment. After conducting in camera review, on August 13, 2020, the court granted OMB's motion for summary judgment and denied AAI's cross-motion, finding that the disputed documents were predecisional and exempt from FOIA. As of December 25, 2020, AAI has not appealed the court's decision.

Tiny Summary: The Office of Management and Budget is forced to disclose documents requested by the Arab American Institute under the Freedom of Information Act. (D.D.C.)

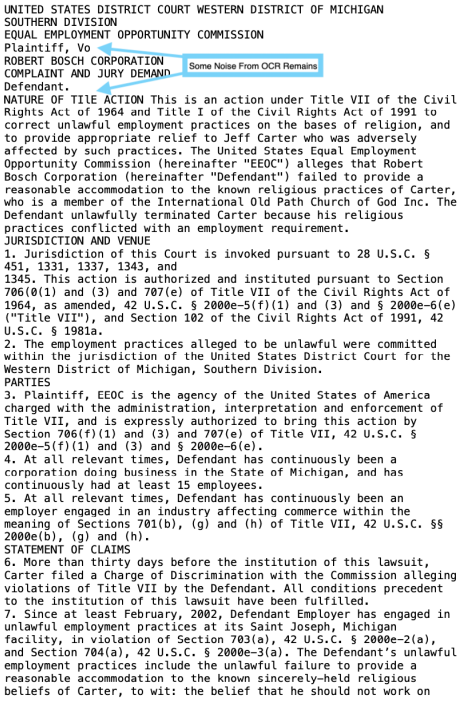
Appendix C

Cleaned Document Example

Here we present an annotated representative excerpt from a case document, before and after applying the cleaning process outlined in Section 4.1.1.



(a) Document before cleaning.



(b) Document after cleaning.

Figure C.1: Annotated representative excerpt, before and after cleaning process.

Appendix D

Source Documents Leading To Noise

Here we present examples of PDF documents on CRLC which lead to noisy plaintext in Multi-LexSum.

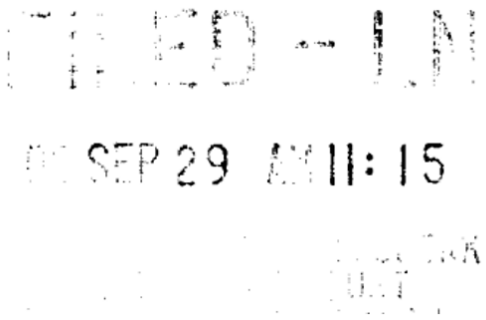
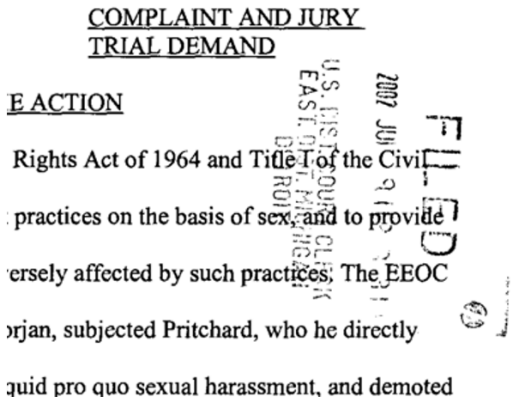


Figure D.1: Physically stamped dates are not highly legible in PDF format and thus not adequately recognised by the OCR process.


COMPLAINT AND JURY
TRIAL DEMAND

E ACTION

Rights Act of 1964 and Title I of the Civil

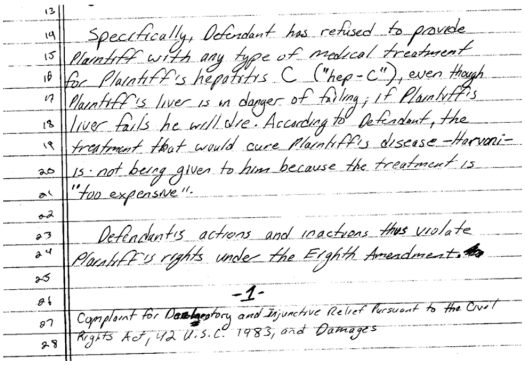
practices on the basis of sex, and to provide

ersely affected by such practices. The EEOC

orjan, subjected Pritchard, who he directly

quid pro quo sexual harassment, and demoted

Figure D.2: Stamps over text lead to ‘junk’ being produced when OCR is applied. For example, ‘and Title I of the Civil’ is reflected as ‘and Tifl.ĩf;Gtthe CiviF’ in the plaintext present in Multi-LexSum. This also leads to incorrect segmentation.


Specifically, Defendant has refused to provide Plaintiff with any type of medical treatment for Plaintiff's hepatitis C ("hep-C"), even though Plaintiff's liver is in danger of failing; if Plaintiff's liver fails he will die. According to Defendant, the treatment that would cure Plaintiff's disease -Hervoni- is not being given to him because the treatment is "too expensive".

Defendant's actions and inactions thus violate Plaintiff's rights under the Eighth Amendment.

-1-

Complaint for Declaratory and Injunctive Relief Pursuant to the Civil Rights Act, 42 U.S.C. 1983, and Damages

Figure D.3: Multi-LexSum contains examples where the original judicial documents are handwritten. In such cases, very little text is recovered when OCR is applied.

Appendix E

NER Categories and Performance on Canadian Refugee Law Dataset

Here we explain all labels of the NER system used and report their performance on the Canadian refugee law dataset this system was developed for. Label descriptions are taken from [71] and performance metrics for the specific model versions used in our study were obtained by personal correspondence with the author.

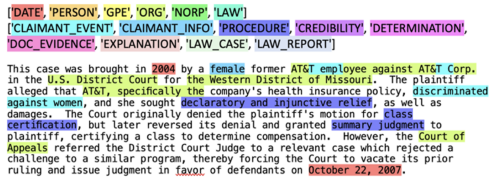
Label	Description	Precision	Recall	F1
DATE	absolute or relative dates or periods	85.19	93.88	89.32
GPE	cities, countries, regions	96.24	98.90	97.55
ORG	tribunals, NGOs, and companies	89.00	90.82	89.90
PERSON	names	73.58	78.00	75.73
LAW	legislation and international conventions	51.61	60.38	55.65
NORP	nationalities, religious, political, or ethnic groups or communities	90.00	85.71	87.80
DETERMINATION	outcome of the decision (accept or reject)	72.41	67.74	70.00
CREDIBILITY	mentions of credibility in the determination	80.43	77.08	78.72
EXPLAINATION	reasons given by the panel for the determination	81.43	61.29	69.94
CLAIMANT_EVENT	verbs or nouns describing an event of the story of the claimant	66.29	64.80	65.54
DOC_EVIDENCE	evidence, proof, and supporting documents	82.95	86.90	84.88
PROCEDURE	legal procedure events	69.51	65.52	67.46
CLAIMANT_INFO	age, gender, citizenship, occupation	88.64	88.64	88.64
LAW_REPORT	country reports written by NGOs or the United Nations	100.00	100.00	100.00
LAW_CASE	case law and past decided cases	62.50	71.43	66.67

Table E.1: NER system label details.

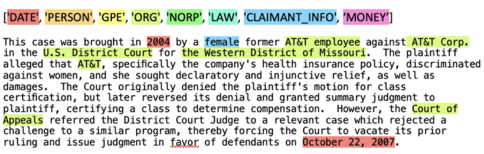
Appendix F

Further NER Tagging Examples

In this section we present further examples of NER tagging Multi-LexSum data.

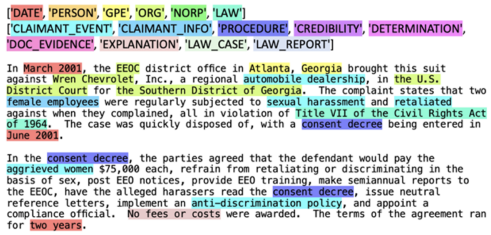


(a) All categories of NER system, no postprocessing.

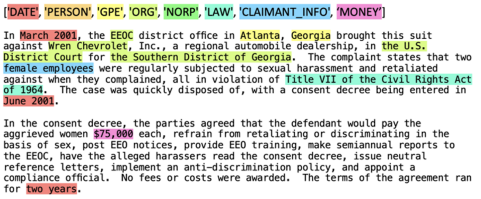


(b) Chosen categories, with MONEY category added and postprocessing performed.

Figure F.1: Example of NER annotation.



(a) All categories of NER system, no postprocessing.



(b) Chosen categories, with MONEY category added and postprocessing performed.

Figure F.2: Example of NER annotation.

Appendix G

CaseLawBERT Classifer ROC Curve

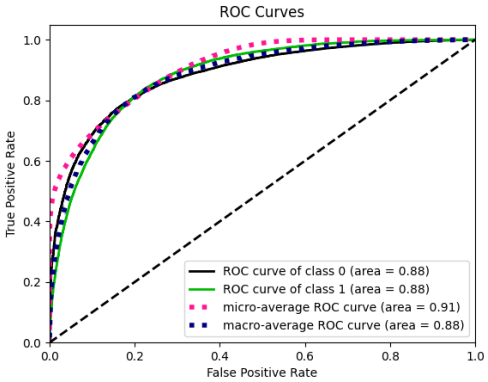


Figure G.1: ROC curve for CaseLawBERT classifier.

Appendix H

Short Summary Results Reported in Multi-LexSum Paper

We include the results for short summaries reported in [18].

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
First K	21.97	7.17	13.61	-1.60
Random K	24.37	3.79	12.92	5.28
BERT-EXT	24.79	5.48	13.89	4.49
PEGASUS	43.35	19.91	29.99	37.88
BART	43.55	19.98	29.84	37.41
LED-4096	45.44	21.00	30.99	39.33
LED-16384	46.54	22.08	31.91	40.00
PRIMERA	45.51	21.04	30.81	39.32
BART MULTITASK	43.80	20.14	29.89	38.00

Table H.1: Mean ROUGE and BERTScore F1 scores with respect to corresponding reference summary.

Appendix I

BERT-Sentences Results With Restricted Number Of Sentences

As explained in Section 5.1, we investigated an alternative variant of BERT-Sentences where the only same number of sentences as returned by OREO are used. This approach led to worse results than the standard BERT-Sentences approach.

	ROUGE-1	ROUGE-2	ROUGE-L
BERT-Sentences	76.61	32.61	46.15
BERT-Sentences (Short)	56.18	19.18	30.29

Table I.1: Mean ROUGE recall between extracts produced and corresponding gold summary.

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
BERT-Sentences	43.61	19.33	27.58	34.52
BERT-Sentences (Short)	42.65	17.98	27.40	33.08

Table I.2: ROUGE and BERTScore F1 scores for BERT-Sentences and BERT-Sentences (Short) input representations with PEGASUS as the abstractive summarisation model.

Appendix J

Entity Precision Results

Here we report entity precision results, as discussed in Section 5.3.

	PEGASUS	Legal-PEGASUS	Legal-PEGASUS + Surface Form Chain	Legal-PEGASUS + Type Chain	Legal-PEGASUS + Combination Chain
BERT-Sentences	0.8328	0.8169	0.7526	0.8184	0.7720
BERT-Windows	0.7927	0.7921	0.7066	0.7952	0.7434
OREO-Sentences	0.7164	0.6850	0.6223	0.7167	0.6700
OREO-Windows	0.7121	0.6948	0.6564	0.7205	0.6928

Table J.1: Mean proportion of entities in generated summary confirmed to occur in the corresponding source text.

Appendix K

Further Model Outputs

Here we present additional examples of model outputs - facts inconsistent with the case documents and other errors (such as assimilating information from cited cases) are highlighted in red. Entity chains are omitted, where applicable.

K.1 Cerda v. Restaurant Associates

Gold: On August 9, 2004, the plaintiff filed this class action lawsuit in the Eastern District of New York on behalf of all Hispanic and/or non-white refreshment concession personnel who were hired, employed or offered employment by the defendants, Restaurant Associates and RA Tennis Corp. The plaintiff alleged that the defendants engaged in unlawful employment discrimination on the basis of national origin and race respecting the assignment of positions and locations to food vendors employed at the U.S. Open Tennis Tournament. Specifically, they alleged that the defendants were deliberately and overwhelmingly assigning beer wagons to non-Hispanic white people over other demographics. The parties soon reached a settlement which was approved by U.S. District Court Judge Robert M. Levy, and Restaurant Associates agreed to pay damages. This case is closed.

Reproduction: On August 9, 2004, a restaurant worker filed a class action lawsuit in the U.S. District Court for the Eastern District of New York against the owners and managers of a restaurant in Brooklyn, New York. The plaintiff alleged that the defendants had violated the Fair Employment and Housing Act (FEH) and the Rehabilitation Act of 1973. The court granted the plaintiff's motion for preliminary injunction, and the parties agreed to a settlement. The case is closed.

BERT-Windows, PEGASUS: On July 15, 2005, a class action lawsuit was filed in the

U.S. District Court for the Eastern District of New York against Restaurant Associates, Inc. and RA Tennis Corp. alleging that the defendants discriminated against non-white and Hispanic employees in violation of the federal, state and City employment non-discrimination statutes. The complaint alleged that defendant had given the higher-paying concession stand jobs to white employees and that the Hispanic and non- white employees were shunted off to lower-paying positions. The parties reached a settlement in which the defendant established a $90,000 settlement fund to be divided pro rata among non-White employees and employees of non-US national origin who file claims. The case was closed on July 20, 2009.

BERT-Windows, Legal-PEGASUS: On July 20, 2009, a class action lawsuit was filed in the U.S. District Court for the Eastern District of New York on behalf of Hispanic and non-white employees of Restaurant Associates, Inc. and RA Tennis Corp. who worked at the US Open Tennis Tournament in 2003. The complaint alleged that the defendants discriminated against non-Hispanic white employees and shunted them to lower-paying positions in violation of the federal, state and City employment non-discrimination statutes. The parties reached a settlement in which the defendants established a $90,000 settlement fund to be divided pro rata among non-White employees and employees of non-U.S national origin who file claims. The settlement was approved by the Court on July 20, 2010.

BERT-Windows, Legal-PEGASUS, Surface Form Chain: On July 16, 2006, a class action lawsuit was filed in the United States District Court for the Eastern District Of New York by a Hispanic employee of Restaurant Associates. The complaint alleged that defendant Restaurant Associates had discriminated against the plaintiff and his fellow employees by giving higher-paying jobs to white employees and shunting them to lower-paying positions in violation of the federal, state, and City employment non-discrimination statutes. The parties reached a settlement in which the defendant established a $90,000 settlement fund to be divided pro rata among the plaintiffs. The settlement was approved by the Court on July 16, 2006.

BERT-Windows, Legal-PEGASUS, Type Chain: On July 20, 2009, a class action was filed in the U.S. District Court for the Eastern District of New York on behalf of Hispanic and non-white employees of Restaurant Associates, Inc. and RA Tennis Corp. who worked at the US Open Tennis Tournament in 2003. The complaint alleged that defendant Restaurant Associates had given the higher-paying concession jobs to white employees and shunted the Hispanic and Non-White employees off to lower-paying positions, in violation of the federal, state and City employment non-discrimination

statutes. The parties reached a settlement in which the defendant established a $90,000 settlement fund to be divided pro rata among non- white employees who filed claims. The settlement is subject to approval by the Court.

BERT-Windows, Legal-PEGASUS, Combination Chain: Lawsuit was filed on July 16, 2009 by a Hispanic employee of Restaurant Associates and RA Tennis Corp. who claimed that he and other Hispanic and non-white employees were shunted off to lower-paying positions in violation of the federal, state and City employment non-discrimination statutes. The complaint alleged that defendant Restaurant Associates hired several hundred persons to staff these facilities for approximately a three-week period during the summer of 2003 and that the defendants had given the higher-paying jobs to white employees. The settlement calls for defendant to establish a $90,000 settlement fund to be divided pro rata among non- white employees and employees of non-US national origin who file claims.

OREO-Windows, Legal-PEGASUS, Type Chain: On July 20, 2006, a Hispanic man filed a class action lawsuit in the U.S. District Court for the Eastern District of New York against Restaurant Associates, Inc. and RA Tennis Corp. alleging that the defendants discriminated against Hispanic and/or non-white refreshment vendors at the 2003 US Open Tennis Tournament. The complaint alleged that, because of the short time-span of the US Open, the defendants deliberately and overwhelmingly assigned the ”Deer wagon” to non-Hispanic white people, 18. The class consisted of all Hispanic and non-White refreshment vendors who worked at the tournament. The defendants agreed to pay $90,000 to each of the 279 members of the class. The court approved the settlement on July 20, 2006.

OREO-Sentences, Legal-PEGASUS: On June 16, 2003, a Hispanic employee of Restaurant Associates, Inc. and RA Tennis Corp. filed a class action lawsuit in the U.S. District Court for the Eastern District of New York against the defendants. The complaint alleged that the defendants discriminated against Hispanic and non-white employees of the US Open Tennis Tournament in violation of Title VII of the Civil Rights Act of 1964. The complaint sought injunctive relief, monetary damages, and other relief. The parties reached a settlement in October 2003. The settlement was approved by the Court in March 2004.

K.2 City of Los Angeles v. JPMorgan Chase

Gold: On May 30, 2014, the City of Los Angeles filed a lawsuit in the U.S. District Court for the Central District of California against JP Morgan Chase Bank, under Equal Credit Opportunity Act (ECOA) and Fair Housing Act/Fair Housing Amendments Act (FHAA). The plaintiff sought damages, declaratory and injunctive relief, alleging that the defendants lending practices were predatory and intentionally discriminated against minority borrowers, and had a disparate impact on minorities. On August 6, 2015, the plaintiff field an ex parte application to stay the case pending appeal of related action. On September 2, 2015, the case was dismissed as stipulated.

Reproduction: On May 30, 2014, the City of Los Angeles filed a lawsuit in the U.S. District Court for the Central District of California against JPMorgan Chase & Co. and Chase Manhattan Bank. The City alleged that the banks had violated the Equal Credit Opportunity Act (ECA) and the Fair Credit Reporting Act (FCRA) by denying the City access to credit card information. The court granted the City's motion for summary judgment and granted the defendants' motion to dismiss. The parties agreed to a settlement agreement. The settlement agreement was entered into on September 2, 2015.

BERT-Windows, PEGASUS: In 2012, the City of Los Angeles filed this lawsuit in the U.S. District Court for the Central District of California against JPMorgan Chase Co. and four other large mortgage servicers. The City alleged that JPMorgan targeted minority borrowers for predatory loan terms and increased the costliness of non-conventional loans at the expense of minority borrowers. The settlement relates to unethical mortgage origination and 26 servicing activities similar to the activities alleged herein. The case is one of four related cases brought by the City against large lending institutions.

BERT-Windows, Legal-PEGASUS: On September 2, 2015, the City of Los Angeles filed this lawsuit in the U.S. District Court for the Central District of California against JPMorgan Chase Co., alleging that the bank violated the Fair Housing Act by targeting minority borrowers for predatory loan terms and increasing the costliness of non-conventional loans at the expense of minority borrowers. The case is one of four related cases brought by the City against large lending institutions, including Wells Fargo, Citigroup Inc., and the California Department of Housing and Urban Development. In 2012, the City entered into a global settlement agreement with the government pursuant to which JPMorgan agreed to pay $753 million into a settlement fund for

distribution 10 to qualified borrowers; and (b) provide an additional $1.2 billion to foreclosure prevention actions. The City's damages include lost tax revenues and the need to provide 21 increased municipal services.

BERT-Windows, Legal-PEGASUS, Surface Form Chain: The City of L.A. filed this suit in September 2015 in the United States District Court for the Central District of California. The City alleged that in 2012, JPMorgan Chase and four other large mortgage servicers agreed to a global settlement. The settlement related to unethical mortgage origination and servicing activities similar to the activities alleged in this suit. The case is one of four related cases brought by the City against large lending institutions, 8 alleging identical claims. All of the claims were dismissed with prejudice.

BERT-Windows, Legal-PEGASUS, Type Chain: The City of Los Angeles filed this suit in the U.S. District Court for the Central District of California against JPMorgan Chase Co. and four other large mortgage servicers, alleging that the defendants violated the Fair Housing Act by targeting minority borrowers for predatory loan terms and increasing the costliness of non-conventional loans at the expense of minority borrowers. The case is one of four related cases brought by the City against large lending institutions. In 2012, the City entered into a global settlement agreement with the government pursuant to which JPMorgan agreed to make a cash payment of $753 million into a settlement fund for distribution 10 to qualified borrowers; and to provide an additional $1.2 billion to foreclosure prevention actions. The settlement related to unethical mortgage origination and 16 servicing activities similar to the activities alleged herein. The City's damages include lost tax revenues and the need to provide 21 increased municipal services.

BERT-Windows, Legal-PEGASUS, Combination Chain: On September 2, 2015, the City of Los Angeles filed a lawsuit in the U.S. District Court for the Central District of California against JPMorgan Chase Co. alleging that the bank violated the Fair Housing Act by targeting minority borrowers for predatory loan terms and increasing the costliness of non-conventional loans at the expense of minority borrowers. The case is one of four related cases brought by the City against large lending institutions. In 2012, JPMorgan Chase and four other large mortgage servicers agreed to a global settlement with the government pursuant to which JPMorgan agreed to pay $753 million into a settlement fund for distribution 20 to qualified borrowers; and provide an additional $1.2 billion to foreclosure prevention actions. The settlement related to unethical mortgage origination and 16 servicing activities similar to the activities alleged in this case. The City's damages include lost tax revenues and the need to provide 21 increased municipal

services.

OREO-Sentences, Legal-PEGASUS: On May 30, 2014, the City of Los Angeles filed suit in the U.S. District Court for the Central 5 District of California against JPMorgan Chase Bank, N.A. and Wells Fargo 4 Co. alleging that the defendants violated the Fair Housing Act and the Federal Housing Act by targeting minority neighborhoods with predatory and otherwise discriminatory lending practices. The case was dismissed on July 29, 2015. On September 2, 2015, the City filed an appeal to the Ninth Circuit Court of Appeals. The Court granted the City's request for a stay of the case pending resolution of the appeal.

K.3 Hopson v. Baltimore

Gold: This is a race discrimination class action brought by black police officers in Baltimore, alleging discriminatory discipline and hostile work environment. It was filed in December 2004, and settled in June 2009 for $2.5 million in damages to the 15 named plaintiffs and injunctive relief for the class.

Reproduction: On June 13, 2005, a group of African-American employees of the Baltimore City Police Department filed a class action lawsuit in the U.S. District Court for the District of Maryland against the City of Baltimore and the State of Maryland. The plaintiffs alleged that the City had violated the Civil Rights Act of 1964 and the Rehabilitation Act of 1973. The case was settled on June 23, 2009, and the parties entered into a settlement agreement.

BERT-Windows, PEGASUS: In 2003, a group of African-American police officers (the "Plaintiffs") filed a class action lawsuit in the U.S. District Court for the District of Maryland against the Baltimore City Police Department (" BPD"). The plaintiffs alleged that the BPD had engaged in a pattern and practice of disparate discipline with respect to African-Americans. The plaintiffs sought declaratory and injunctive relief and compensatory damages to secure future protection and to redress the past deprivation of rights guaranteed by Title VII of the Civil Rights Act of 1964 ("Title VII") and the Equal Protection Clause of the United States Constitution ("Equal Protection Clause"), as well as monetary and other relief. The parties reached a settlement agreement in 2005 and the case was dismissed on March 26, 2007. The settlement fund will be used to pay the claims of the plaintiffs and all of their attorneys.

BERT-Windows, Legal-PEGASUS: On February 4, 2005, a group of African-American police officers filed a class action lawsuit in the U.S. District Court for

the District of Maryland against the Baltimore City Police Department. The plaintiffs alleged that the BPD had engaged in a pattern and practice of disparate discipline with respect to African-Americans. The parties entered into a settlement agreement in which the BPD agreed to fund and hold no less than five (5) leadership training classes during the five years following the Effective Date of the Agreement. The settlement also required the BPD to pay the plaintiffs' attorneys' fees and costs.

BERT-Windows, Legal-PEGASUS, Surface Form Chain: On February 2004, African-American police officers in Baltimore City, Maryland filed a class action lawsuit against the City of Baltimore and the Baltimore Police Department. The plaintiffs alleged that the City had engaged in a pattern and practice of discrimination against African-Americans in violation of Title VII. The suit sought declaratory and injunctive relief and compensatory damages. The parties entered into a settlement agreement in 2005. The parties agreed to fund and hold no less than five (5) leadership training classes during the five years following the Effective Date of the Agreement. The amount of the settlement will be used to pay the claims of the fifteen individual plaintiffs as well as all of the class members.

BERT-Windows, Legal-PEGASUS, Type Chain: This is a class-action lawsuit brought by African-American police officers against the Baltimore City Police Department in the U.S. District Court for the District of Maryland. The plaintiffs alleged that the BPD had engaged in a pattern and practice of discrimination against African-Americans in violation of Title VII of the Civil Rights Act of 1964 ("Title VII") and the Equal Protection Clause of the Fourteenth Amendment to the United States Constitution ("Ecclestone"), and that the disciplinary process of the BPD, although purportedly neutral on its face, was administered through a centralized, excessively subjective policy and procedures that have a disparate impact on African- Americans. The parties entered into a settlement agreement in 2005 that provided for the BPD to fund and hold no less than five (5) leadership training classes during the five years following the Effective Date of the Agreement. The settlement was subject to court approval.

BERT-Windows, Legal-PEGASUS, Combination Chain: On February 2004, African-American police officers in Baltimore City, Maryland filed a class action lawsuit against the City of Baltimore in the United States District Court for the District of Maryland. The plaintiffs alleged that the City had engaged in a pattern and practice of discrimination on the basis of color and race in violation of Title VII. The parties entered into a settlement agreement on February 2005. The parties agreed to fund and hold no less than five (5) leadership training classes during the five years following the

Effective Date of the Agreement. The settlement is subject to approval by the court.

OREO-Windows, Legal-PEGASUS, Type Chain: This class action was filed in the U.S. District Court for the Southern District of New York on November 26, 2006 by two African-American police officers against the Boston Police Department (BPD) alleging racial discrimination in the disciplinary process of the BPD. The complaint alleged that the BPD's policies and practices for disciplinary decisions had a disparate impact on African-Americans, and that the defendants had engaged in a pattern and practice of disparate discipline with respect to African- Americans. The case was dismissed on October 31, 2006.

OREO-Sentences, Legal-PEGASUS: This class action lawsuit was filed in the U.S. District Court for the District of Columbia against the Baltimore Police Department (BPD) in August 2004 on behalf of African-American police officers. The plaintiffs alleged that the BPD had engaged in a pattern and practice of disparate discipline with respect to African-Americans. The class sought injunctive relief, monetary damages, and other relief. The court granted the plaintiffs' motion for summary judgment, and the case was closed.

K.4 Perez-Farias v. Global Horizons, Inc.

Gold: On July 12, 2005, three U.S. farm works from Yakima Valley who were illegally denied or terminated from agricultural employment brought this class action lawsuit in the U.S. District Court for the Eastern District of Washington against Global Horizons, Green Acre Farms, and Valley Fruit Orchards under the Migrants and Seasonal Agricultural Worker Protection Act, 29 U.S.C. ¬ß¬ß 1801, the Farm Labor Contractors Act, 42 U.S.C ¬ß 1981 and the Washington Law Against Discrimination. The parties agreed to go to trial, and after years of litigation the plaintiffs received a judgment in their favor under FLCA, as well as attorneys fees, for both trial and appellate legal work. In January 2014, the Court granted the plaintiffs' motion for incentive payments. There has been no further movement in the case.

Reproduction: This case was brought in the U.S. District Court for the Eastern District of Washington by a former employee of the Department of Health and Human Services (HHS) who alleged that the agency violated the Equal Protection and Due Process Clauses of the Fourteenth Amendment by failing to comply with the requirements of the Freedom of Information Act (FFRA) and the Administrative Procedure Act (APA). The parties reached a settlement agreement in which HHS agreed to revise

its policies and procedures to ensure compliance with the FFRA and APA. The parties agreed to dismiss the case with prejudice. The case is closed.

BERT-Windows, PEGASUS: On June 3, 2005, the Washington State Department of Labor and Industries (D.L.I.) filed this class action lawsuit in the U.S. District Court for the Western District of Washington against Global Horizons, Inc., Green Acre Farms, and Valley Fruit Orchards. The complaint alleged that the defendants failed to pay wages due to 12 members of the Valley Fruit subclass by failing to pay the promised piece rate, failed to provide adequate written pay statements, and failed to disclose 2 production standards for pruning, thinning, harvesting and other activities in the written statement of the terms and conditions of employment. The parties reached a settlement agreement in which the defendants admitted to violating the 17 Washington Farm Labor Contractor Act.

BERT-Windows, Legal-PEGASUS: On June 3, 2005, the Washington State Department of Labor and Industries (W.D.L.I.) filed this class action lawsuit in the U.S. District Court for the Western District of Washington against Global Horizons, Inc., a farm labor contractor, and Valley Fruit Orchards, LLC. The complaint alleged violations of the Washington Farm Labor Contractor Act. According to the complaint, the defendants failed to disclose two production standards for pruning, thinning, harvesting, and other activities in the written statement of the terms and conditions of employment. A significant number of class members were not fluent or literate in English; the language common to the workers was Spanish. The defendants also failed to pay wages due to 12 members of the Valley Fruit subclass by failing to pay the promised piece rate.

BERT-Windows, Legal-PEGASUS, Surface Form Chain: On June 2005, the Washington State Labor Department filed this class action lawsuit in the U.S. District Court for the Western District of Washington on behalf of resident farm workers employed by Global Horizons. The complaint alleged that Global Horizons violated the Washington Farm Labor Contractor Act by failing to disclose production standards for pruning, thinning, harvesting, and other activities in the written statement of the terms and conditions of employment, failing to pay wages due to 12 workers, and failing to provide adequate written pay statements to workers. The case was eventually settled by a settlement agreement between the defendants and the state of Washington.

BERT-Windows, Legal-PEGASUS, Type Chain: On June 3, 2005, the Washington State Department of Labor and Industries (D.L.I.) filed this class action lawsuit in the U.S. District Court for the Western District of Washington against Global Horizons, Inc., Green Acre Farms, and Valley Fruit Orchards, LLC, alleging violations of the

Farm Labor Contractor Act. The complaint alleged that the defendants violated the terms and conditions of employment by failing to disclose two production standards for pruning, thinning, harvesting, and other activities in the written statement of the terms of employment. A significant number of the class members were not fluent or literate in English; the language common to the workers was Spanish. The defendants also failed to pay wages due to 12 members of the Valley Fruit subclass. The case was preceded by an investigation by the Washington Department 20 of Labor that ultimately resulted in a settlement agreement between the state of Washington, the Global Defendants and the Grower Defendants. As part of the 22 settlement process, Global admitted to violating Washington state and federal laws.

BERT-Windows, Legal-PEGASUS, Combination Chain: On June 3, 2005, the Washington State Department of Labor and Industries (D.L.I.) filed this class action lawsuit in the U.S. District Court for the Western District of Washington against Global Horizons, Inc., Green Acre Farms, and Valley Fruit Orchards LLC. The complaint alleged that the defendants violated the Farm Labor Contractor Act by failing to disclose production standards for pruning, thinning, harvesting, and other activities in the written statement of the terms and conditions of employment, failing to pay wages due to 12 members of the Valley Fruit subclass, and failing to provide adequate written pay statements to workers. As part of the settlement process, the defendants entered into a settlement agreement in which they agreed to pay the workers a total of $2,200. The settlement agreement also required the defendants to provide the workers with a copy of the employer's address and telephone number, the beginning and ending dates of the pay period, and the rate of pay for "other" hours.

OREO-Windows, Legal-PEGASUS, Type Chain: In July 2005, a group of Yakima Valley resident workers filed a class action lawsuit in the U.S. District Court for the Eastern District of Washington against Global Horizons, Inc. and Mordechai Orian, alleging violations of the Migrant and Seasonal Agricultural Worker 26 Protection Act (AWPA) and the Farm Labor Contractor's Act 27 (FLCA). The complaint alleged that the defendants violated AWPA and FLCA by failing to adequately disclose the terms and conditions of employment, failing to provide adequate disclosure of the job offers, and providing false and misleading information about job offers. The court granted class certification and ordered the defendants to pay $1,458, which was the cost of the interpretation fees that were provided at 12 trial on a cost-reimbursable basis. The case is now closed.

Oreo-Sentences, Legal-PEGASUS: In 2005, a group of immigrant farm workers filed

a class action lawsuit in the U.S. District Court for the Eastern District of Washington against Global Horizons, Inc. and Mordechai Orian, alleging violations of the Migrant and Seasonal Agricultural Worker Protection Act and the Farm Labor Contractor's Act. The plaintiffs sought injunctive relief, monetary damages, and attorney's fees. In July 2005, the court entered a judgment in favor of the plaintiffs in the amount of $1,452,831.

K.5 Adar v. Smith

Gold: Plaintiffs are a same-sex couple who asked a federal court in Louisiana to declare as unconstitutional Louisiana's refusal to issue an amended birth certificate for an adopted child listing both individuals in an unmarried same-sex couple as the legal parents, and, for injunctive relief requiring Louisiana to issue such birth certificate. After the district court held for plaintiffs, the Fifth Circuit reversed and found that Louisiana's recognition of a New York adoption order was not entitled to the Full Faith and Credit Clause of the U.S. Constitution, and that there was no Equal Protection violation.

Reproduction: On October 9, 2007, the parents of a minor filed a lawsuit in the U.S. District Court for the Eastern District of Louisiana against the State Registrar of Vital Records and Statistics of the Louisiana Department of Health and Hospitals. The plaintiffs alleged that the State was violating the Equal Protection Clause of the Fourteenth Amendment by denying them the right to equal protection under the law. The case was dismissed on March 18, 2009, and the Fifth Circuit affirmed the district court's decision.

BERT-Windows, PEGASUS: A same-sex couple who jointly parent their child as legal parents sued the Louisiana State Registrar of Vital Records under 42 U.S.C. 1983 for declaratory and in relief, asserting that her action denies full faith and credit to the New York adoption decree and equal protection to them and Infant J. The district court ruled in favor of the plaintiffs, and the Seventh Circuit affirmed.

BERT-Windows, Legal-PEGASUS: In October 2007, a same-sex couple who jointly parent their child as legal parents filed a lawsuit in the U.S. District Court for the Eastern District of Louisiana against the Louisiana State Registrar of Vital Records. The plaintiffs sought declaratory and injunctive relief, asserting that the Registrar's refusal to issue a birth certificate violated the Full Faith and Credit Clause (the "Clause") and the Equal Protection Clause of the United States Constitution. The district court granted the plaintiffs' summary judgment motion based solely on their FF&C Clause claim.

The Seventh Circuit affirmed the district court's decision.

BERT-Windows, Legal-PEGASUS, Surface Form Chain: In April 2006, a same-sex couple who jointly parented their child in Louisiana filed a lawsuit against the State of Louisiana for failing to issue a birth certificate for their child, who was adopted in New York. The couple claimed that the failure to issue the certificate violated the Full Faith and Credit Clause and the Equal Protection Clause of the United States Constitution. The district court granted the couple's motion for summary judgment, and the Seventh Circuit affirmed the district court's decision. The case is now closed.

BERT-Windows, Legal-PEGASUS, Type Chain: Plaintiffs are a same-sex couple who jointly parent their child as legal parents. Their child was adopted in a final adoption order entered outside the State of Louisiana. The plaintiffs filed suit in the U.S. District Court for the Eastern District of Louisiana against the State Registrar of Vital Records, alleging that the Registrar's refusal to issue a birth certificate violated the Full Faith and Credit Clause (the "Clause") and the Equal Protection Clause of the United States Constitution. The district court granted the plaintiffs' summary judgment motion based solely on their FF&C Clause claim, and the Seventh Circuit affirmed the district court's decision.

BERT-Windows, Legal-PEGASUS, Combination Chain: Plaintiffs are a same-sex couple who jointly parent their child as legal parents. Their child was adopted in a final adoption order entered outside the State of Louisiana. In October 2007, the plaintiffs sued the Louisiana State Registrar of Vital Records for declaratory and injunctive relief under the Full Faith and Credit Clause and the Equal Protection Clause of the United States Constitution. The district court ruled in favor of the plaintiffs on their full faith and credit claim. The court held that Louisiana law, properly understood, required the Registrar to reissue the birth certificate.

OREO-Windows, Legal-PEGASUS, Type Chain: On October 9, 2007, a same-sex couple who jointly parent their child as legal parents filed a lawsuit in the U.S. District Court for the Western District of Louisiana against the State of Louisiana, the Office of Vital Records and Statistics, the Louisiana Department of Health and Hospitals Smith, and the Illinois state police department under No. 09-30036 1983, alleging that Louisiana's refusal to issue an accurate birth certificate for their child violated the Equal Protection Clause of the Fourteenth Amendment of the United States Constitution. The district court granted summary judgment to the plaintiffs, holding that Louisiana owes full faith and credit to the New York state adoption decree and that there is no public policy exception to the Clause. The Seventh Circuit remanded the case to the district

court for further proceedings.

OREO-Sentences, Legal-PEGASUS: The plaintiffs, a same-sex couple who jointly parent their child as legal parents, filed this lawsuit in the U.S. District Court for the Western District of Louisiana. The plaintiffs alleged that Louisiana's refusal to respect their out-of-state adoption decree and refusal to issue an amended birth certificate violated the Equal Protection Clause of the Fourteenth Amendment of the United States Constitution. The district court granted summary judgment to the plaintiffs, holding that Louisiana owes full faith and credit to the New York adoption decree. The Seventh Circuit affirmed the district court's decision.