BB25HLegalSum: Leveraging BM25 and BERT-based clustering for the summarization of legal documents

Leonardo Bonalume
Informatics Institute
Federal Univ. of Rio Grande do Sul
Porto Alegre - Brazil
lbandrade@inf.ufrgs.br

Karin Becker
Informatics Institute
Federal Univ. of Rio Grande do Sul
Porto Alegre - Brazil
karin.becker@inf.ufrgs.br

Abstract

Legal document summarization aims to provide a clear understanding of the main points and arguments in a legal document, contributing to the efficiency of the judicial system. In this paper, we propose BB25HLegalSum, a method that combines BERT clusters with the BM25 algorithm to summarize legal documents and present them to users with highlighted important information. The process involves selecting unique, relevant sentences from the original document, clustering them to find sentences about a similar subject, combining them to generate a summary according to three strategies, and highlighting them to the user in the original document. We outperformed baseline techniques using the BillSum dataset, a widely used benchmark in legal document summarization. Legal workers positively assessed the highlighted presentation.

1 Introduction

Pending judicial processes are a prevalent and significant issue affecting legal systems worldwide. The number of pending cases can vary significantly depending on population size, legal system, and backlog of cases. While in some countries, there may be only a few thousand pending processes, it can amount to millions in others. This scenario motivates the research of computational techniques that can help accelerate judicial analysis, select similar cases for judging in batches, or identify patterns that could lead to better decision-making. The automatic summarization of legal documents to synthesize their essence is critical in this context.

The goal of automatic text summarization is to create summaries that are similar to human-created summaries (Allahyari et al., 2017). This is a challenging task since natural language is complex and nuanced. Text summarization algorithms must consider the intended audience, the purpose of the summary, as well as the type and format of the original text. Text summarization is valuable for various applications, such as news aggregation, document management, and legal document summarization.

Most works use extractive summarization to generate summaries, defined in (Anand and Wagh, 2019), as "the generation of a summary containing a sentence subset of the original text after identifying the important sentences". Several techniques were explored for extractive legal text summarization, including word relevance (Polsley et al., 2016), graph-based ranking models (Dalal et al., 2023; Jain et al., 2023), statistical models (Jain et al., 2022; Merchant and Pande, 2018), and deep learning (Anand and Wagh, 2019). More recently, BERT (Devlin et al., 2018) has been leveraged in the legal area (Furniturewala et al., 2021), inspired by state-of-the-art results achieved in general extractive text summarization (Liu, 2019).

Another approach used in the legal documents area is BM25 (Robertson et al., 2009), a ranking function commonly used in information retrieval to determine the relevance of a document concerning a search query. The combined use of BERT and BM25 is recurrent for information retrieval in legal documents (Askari et al., 2022; Althammer et al., 2021), but it is still in the initial stages in the legal documents summarization area. BERT is a powerful language model that captures complex relationships between words and sentences, while BM25 is an effective information retrieval algorithm to rank documents. The strengths of these techniques can be joined to produce high-quality summaries and help to overcome some of the traditional methods' hurdles (e.g., feature engineering, long documents).

According to (Jain et al., 2021), there needs to be more analysis of the readability of the generated summaries, and how to present them. In the legal area, summary presentation is addressed using highlighting (Licari et al., 2023) and heatmaps

Proceedings of Recent Advances in Natural Language Processing, pages 255–263
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_029

(Polsley et al., 2016) representing the relevance of sentences within the original document. However, the relevance of a sentence may be a secondary aspect for legal workers, who seek the main arguments within their context.

In this article, we propose BB25HLegalSum (BERT + BM25 + Highlighting Legal Documents Summarization), a novel method for the extractive summarization of legal documents. It leverages BERT and BM25 to identify relevant sentences in a legal document and combine clusters of sentences to generate candidate summaries, which are selected using metrics against a reference summary. We generate summaries using three strategies to identify the best parts of a document, focused on the precision of the selected sentences, their coverage of the text (recall), and a trade-off between these two criteria. Another distinctive feature is the presentation of the generated summary. We propose a subsidiary highlighting approach that represents, using different colors, the sentences contained in the summaries generated according to each strategy. In this way, the user can identify and distinguish in their original context the relevant sentences of the document according to distinct points of view that emphasize precision, coverage, or both.

Our experiments address the following research questions: (1) How does the performance of BB25HLegalSum compare to baseline methods for legal document summarization? (2) How does the length of the reference summary impact the recall and precision of the generated summary using BB25HLegalSum? (3) Which type of document summary is more readable in the legal context: focused on precision, recall, or f-measure?

Our method outperformed baseline works in a benchmark dataset (Jain et al., 2021). We observed that the length of the reference summary impacts the recall and precision of the generated summaries and that BB25HLegalSum performs better for larger-than-average summaries. A qualitative assessment by legal workers has shown that highlighting with distinct colors enables identifying different types of information captured by each summarization strategy. They pointed out that higher recall is the most critical criterion for summarization in the legal context, since it avoids missing relevant information.

The main contributions of our article are:
(1) a method that leverages BERT and BM25 to generate legal document summaries. It outperforms baselines (Anand and Wagh, 2019; Mihalcea and Tarau, 2004; Erkan and Radev, 2004) in a benchmark dataset;
(2) a presentation method for the generated summaries using different colors that highlights in their original context the importance of sentences according to distinct points of view (precision vs. coverage). Legal workers positively assessed this presentation.

The remaining of this work is structured as follows. Section 2 presents related work. Section 3 describes BB25HLegalSum in detail. Section 4 presents our experiments. Section 5 outlines the conclusions and future work.

2 Related Work

Extractive summarization forms summaries by selecting and concatenating the most important spans (typically sentences) in a document (Liu, 2019). Legal document summarization has explored various techniques. CaseSummarizer (Polsley et al., 2016) combines standard summary methods based on word relevance (i.e., TF-IDF) with domain-specific knowledge to summarize legal documents. Graph-based ranking models, notably LexRank (Dalal et al., 2023) and TextRank (Jain et al., 2023), explore the relationships and similarities between nodes representing the text to select the relevant portions of legal documents. Statistical models have been utilized for scoring the relevance of sentences in legal documents, including Bayesian optimization (Jain et al., 2023), Kullback-Leibler (Jain et al., 2022), and Latent Semantic Analysis (Merchant and Pande, 2018). The contextual nuances and semantic dependencies in legal documents are explored for generating summaries using deep learning (Anand and Wagh, 2019). More recently, a trend is to deploy pre-trained models such as BERT (Furniturewala et al., 2021), which capture complex relationships between words and sentences.

The focus in some works is the presentation of the generated legal summary. (Licari et al., 2023) uses different colors to highlight the top-5 sentences, and (Polsley et al., 2016) proposes a heatmap to distinguish the importance of sentences. However, the relevance of a sentence may be a secondary aspect for legal workers, given that they generally seek the key arguments within a legal document.

The quality of generated summaries is typically assessed by comparing the generated summary against some reference summary using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (ROUGE, 2004). In the context of ROUGE, recall refers to how much of the reference summary is captured in the system summary, precision measures how much of the system summary is relevant, and F1 combines recall and precision. Necessary assessments on legal text summarization remain unaddressed, such as properties of the readability of the summaries (e.g., the trade-off between conciseness and completeness) and the relationship between performance efficiency and reference summaries, typically used as the gold standard to evaluate the proposed summary systems (Jain et al., 2021).

BM25 (Robertson et al., 2009) is a well-established information retrieval algorithm that ranks documents based on their relevance concerning a query. The combined use of BERT and BM25 is recurrent for document retrieval in the Competition on Legal Information Extraction/Entailment (COLIEE) (Askari et al., 2022; Rosa et al., 2021; Althammer et al., 2021), but its potential has not been fully examined for legal document summarization. The resulting summarization model can benefit from the strengths of both approaches to produce high-quality summaries and help to overcome some of the traditional methods' hurdles, such as the reliance on feature engineering and the difficulty in handling long documents.

Our work contributes with a solution that leverages BERT and BM25 to produce legal document summaries, and with a method for presenting the generated summaries using highlighting that enables the examination of the trade-off between conciseness and completeness for readability of legal documents summaries.

3 BB25HLegalSum overview

BB25HLegalSum is a novel method for the extractive summarization of legal documents. It assumes as input a legal document D, composed of a legal description (*desc*), and a reference summary (*refSum*). Given a document D, the goal is to select from *desc* a set of relevant sentences and to combine them to produce a *generated summary*, hereafter *GSum*. Our premise is that, for a lawyer, the most important aspect of legal document summarization is the extraction of the most relevant arguments and the ability to identify their importance within a context. Hence, the *refSum* may synthesize the document, but it does not necessarily provide all the useful information a legal worker needs.

Our method comprises four main steps: (1) select from *D.desc* a set of unique, relevant sentences by leveraging BERT to explore similarity thresholds and BM25 to rank sentences; (2) aggregate relevant sentences to select a set of *candidate summaries* ($candSum_m$) by combining clusters of related sentences; and (3) select among the candidate summaries the most representative one, as measured by ROUGE against the reference summary (*D.refSum*); (4) present the generated summary *GSum* in the original document by highlighting the selected sentences using different colors, combining multiple perspectives of importance.

A significant concern in our work is understanding the trade-off of conciseness and completeness as a measure of the quality of the generated summaries. Hence, our method proposes and assesses three strategies to select the best-generated summary, given a set of possible candidates, according to the metrics used for the selection (precision, recall, and f-measure, respectively). The remaining of this section provides details on our method.

3.1 Extracting BERT and BM25 candidate sentences

Given a legal document D(desc, refSum), the goal is to decompose D.*desc* into a set of sentences s_i (where $0 < i <$ *D.desc.length*), and explore BERT and BM25 to select the most relevant ones. We refer to these as *sentence filters*. The goal is to output three sets with sentence indices (minSizeFilterIDX, BERTFilterIDX, BM25FilterIDX), where each index is a set $\{a \mid 0 \leq a \leq$ *D.desc.length*$\}$, such that there exists a sentence $s_a \in$ D.*desc*.

(a) minimum size filter: the first issue is the minimum sentence size required for each sentence to be a candidate, using a *size_threshold*. The rationale is to remove sentences that are too short because in legal datasets usually the reference summary is comprised of long sentences. Given a set of documents, we defined the value of *size_threshold* experimentally. First, we measured the shortest sentence in the reference summary of all documents and then calculated the average (*shortestSentsrefSumAvg*). In our experiments, $size\_threshold = 2 * shortestSentsrefSumAvg$. The list *minSize-*

FilterIDX contains the index of the D.*desc* sentences with minimum size.

(b) BERT Filter: the goal of the BERT filter is to eliminate duplicated sentences. Initially, each sentence s_i is transformed into an equivalent BERT representation br_i according to a pre-trained BERT model. To determine that a sentence br_i is duplicated, we calculate its similarity with regard to all other br_j previously selected. We defined uniqueness according to a maximum *similarity_threshold*; otherwise, it is considered a duplicate and it is discarded. We defined $similarity\_threshold = 0.9$ experimentally, as a good trade-off to distinguish between repetitive sentences and sentences about a similar topic. The list *BERTFilterIDX* contains the index of the non-duplicate sentences in D.*desc*, considering the *similarity_threshold*.

(c) BM25 Ranking filter: BM25 is a bag-of-words retrieval function that ranks documents based on the query terms appearing in each document. The rationale of this filter is to select the sentences that are more representative according to the overall document, affecting the precision of the generated summary. We used as query terms all the tokens extracted from *D.desc*, and then ranked the sentences s_i according to their relevance. We select the top-n best-ranked sentences as the relevant ones. Experimentally, we defined $top - n = 50\%$. The list *BM25FilterIDX* contains the index of the top-n most relevant sentences according to BM25 ranking.

Finally, we compute *filteredSentencesIDX* as the intersection between *minSizeFilterIDX*, *BERTFilterIDX* and *BM25FilterIDX*. *FilteredSentences* is a set of sentences fs_i, where $i \in$ *FilteredSentences*.

3.2 Generating and selecting candidate summaries

In this stage, we generate a set of candidate summaries $candSum_m$, selecting the best one in terms of ROUGE metrics concerning D.*refSum*. To that end, we cluster all *FilteredSentences* from the previous step, and interactively aggregate clusters of sentences to generate a set of $candSum_j$. The generated candidates are compared against the *D.refSum* at each iteration, and the best one is selected. *GSum* is the set of sentences from the best combination of clusters (i.e., the best $candSum_j$).

(a) Clustering of relevant sentences: the goal of this step is to find groups of related relevant sentences. Recall that due to the BERT filter, sentences in a cluster are more related than strongly similar. The rationale is to group related sentences according to a subject or topic and combine them to compose the candidate summaries. This approach also has the advantage of reducing the search space of sentences to include in the generated summary, since instead of testing combinations of sentences, we assess combinations of sentence clusters. In this way, we reduce the possible combinations and, consequently, the execution time.

As the input, we used the BERT representations of the sentences from *FilteredSentences*, created in the previous step. We performed the clustering using the K-means algorithm, comparing the BERT representation of the sentences using a similarity function. This step results in a set *C* of k clusters. One of the challenges of using K-Means is to find the appropriate value for k. To do that, we varied the value of k from 2 to 50, selecting the best clustering. We tested two approaches for this selection: the clustering with the best Silhouette score (Rousseeuw, 1987) and the Elbow method using SSE (Sum of the Squared Error) (Umargono et al., 2020). For the silhouette scores, we used the *silhouette_score* function of the sklearn.metrics library, choosing the clustering with the highest silhouette. The Elbow method consists of plotting the explained variation (measured using SSE) as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The results reported in this paper were produced using the best Silhouette score as the criterion for selecting k.

(b) Generating candidate summaries: given a set of k clusters, the goal of this step is to generate candidate summaries by combining clusters of sentences encompassing different topics. We iteratively create candidate summaries $candSum_j$ from the combination of l clusters from C, compare them with *D.refSum* using ROUGE-1 scores and then use the winning candidate to create combinations of $l + 1$ clusters. At each step, we save the combination of l clusters with the best score ($candSum_l$). *GSum* is the final winning $candSum_j$ for a particular criterion. Due to computational restrictions, in our experiments we varied $l = 2..6$ (i.e. combinations of sentences of 2 up to 6 clusters).

Rouge-1, Rouge-2, and Rouge-L can be used to evaluate the quality of generated summaries. They measure the overlap between a generated summary and the *refSum* regarding unigrams, bigrams, and

longest common subsequences, respectively. We adopted the Rouge-1 given that precise wording and specific terminology are critical in legal documents.

We select the best candidate summaries, and ultimately the *GSum* for a document, according to three strategies, as represented by Rouge metrics: a) *precision-oriented summary* (***PoSum***), focused on conciseness; b) *recall-oriented summary* (***RoSum***), focused on completeness; and c) (*f-measure-oriented summary* (***FoSum***), as a trade-off. Conciseness refers to conveying the message clearly and succinctly without including unnecessary details. Completeness relates to the inclusion of key information from the original text. A summary with good conciseness and completeness will be easy to read and understand, ensuring that produced summaries convey key information from the legal document to the target audience. Conducting a qualitative assessment of summary readability is crucial to ensuring that the research findings can have a real-world impact on legal workers, and we qualitatively assessed the summaries generated according to each strategy in terms of conciseness and completeness.

3.3 Highlighting summaries in the legal document

To be useful, it is important that the generated summaries are readable. We propose to present them as highlights in the original text. Highlighting text improves the reader's knowledge and understanding of the topic being explored (Roy et al., 2021) and it allows the reader to fully grasp not only the relevant words but their context, which can be inspected whenever necessary.

We chose to present the three types of summaries within a single document, using three different colors, one for each criterion-focused summary (green for PoSum, blue for FoSum, and red for RoSum). This allows the reader to understand the different nuances for each highlighted color while condensing the three generated summaries into a single text. We chose to highlight with three colors in a subsidiary way (*subsidiary highlighting*) instead of highlighting the colors of the intersections (*intersectional highlighting*), since the latter could make the reading more difficult. Compared to related work (Polsley et al., 2016; Licari et al., 2023), we provide the context for the relevant sentences and highlight them according to different points of view (precision vs. coverage).

Our method relies on the premise that PoSums are shorter than the FoSums, which in turn are shorter than RoSums. Given the PoSum, FoSum and RoSum generated for a given document D, we start by highlighting with green every tri-grams that appear in the PoSum. Then we highlight in blue every tri-grams that appear in the FoSum that were not included in the PoSum. Finally, we highlight in red all tri-grams that appear in the RoSum and which have not been highlighted yet.

4 Experiments and Results

4.1 Datasets and model

Our experiments are based on the BillSum dataset[1], which is extensively used to measure the performance of summarization methods over legal documents (Kornilova and Eidelman, 2019). It is a dataset that contains the summarization of US Congressional and California state bills. Each bill contains a title, a textual legal description, and a summary. This dataset is divided into training data and test data. Since our method is unsupervised, we used only the test datasets. US test data contains 3269 bills, and CA test data has 1238 bills.

We run our method in all bills in the test datasets. Since we use three criteria to select the winning summaries (f-measure, precision, and recall), for each bill, we generated three types of summaries (FoSum, PoSum and RoSum), measuring the respective precision, recall, and f1 measures for ROUGE-1, ROUGE-2, and ROUGE-L. We implemented our solution using Python 3.6 and libraries such as itertools, sklearn, SentenceTransformer, gensim and numpy. We used the embedder 'distiluse-base-multilingual-cased-v1'.

4.2 Experiment 1

This experiment addresses the following research question: "*How does the performance of BB25HLegalSum compare to baseline methods for legal document summarization?*". As a baseline, we have used the best results compiled in (Jain et al., 2021), namely LSTM with word2vec, LexRank and TextRank. We report the results considering all three strategies for selecting the winning summary (FoSum, PoSum, RoSum). The results presented in Tables 1 and 2 are the average of the scores for all bills in the US test data and CA test data, respectively.

[1] https://github.com/FiscalNote/BillSum

US Dataset	Rouge-1			Rouge-2			Rouge-L		
	F	P	R	F	P	R	F	P	R
LSTM-with-w2v	0.3615	N/A	0.6539	0.2086	N/A	0.3720	0.3664	N/A	**0.5358**
Lexrank	0.3704	N/A	0.5415	0.1811	N/A	0.2604	0.3365	N/A	0.4230
Textrank	0.3269	N/A	0.6295	0.1793	N/A	0.3423	0.3383	N/A	0.5037
BB25HLS FoSum	**0.4425**	**0.3941**	0.5946	**0.2550**	**0.2264**	0.3506	**0.3722**	**0.3482**	0.4539
BB25HLS PoSum	**0.4000**	**0.4839**	0.4676	**0.2295**	**0.2796**	0.2762	0.3446	**0.4215**	0.3749
BB25HLS RoSum	**0.4022**	**0.3090**	**0.6936**	**0.2464**	**0.1894**	**0.4293**	0.3661	**0.3011**	0.5330

Table 1: Performance on US test data

CA Dataset	Rouge-1			Rouge-2			Rouge-L		
	F	P	R	F	P	R	F	P	R
LSTM-with-w2v	0.4073	N/A	0.4638	0.1883	N/A	0.2093	0.3312	N/A	0.3588
Lexrank	0.4144	N/A	0.4529	0.1936	N/A	0.2083	0.3406	N/A	0.3531
Textrank	0.4069	N/A	0.5055	0.2015	N/A	0.2461	0.3457	N/A	0.3848
BB25HLS FoSum	**0.4481**	**0.4338**	**0.5425**	**0.2441**	**0.2356**	**0.3000**	**0.3593**	**0.3485**	**0.4116**
BB25HLS PoSum	0.4031	**0.5707**	0.3307	0.1656	**0.2383**	0.1697	0.2596	**0.3449**	0.2748
BB25HLS RoSum	**0.4131**	**0.3358**	**0.6188**	0.1979	**0.1599**	**0.3433**	0.2986	**0.2521**	**0.4577**

Table 2: Performance on CA test data

We outperformed the baselines in most cases. Overall, the RoSums yielded the best scores in the US test data, while the FoSums display the best performance in the CA test data. If we consider 9 criteria for each dataset, by combining the types of ROUGE score and metric, we outperformed 15 out of 18 criteria for the FoSum strategy and 14 out of 18 for the RoSum strategy. The PoSum strategy outperformed all baselines in terms of precision.

Although we did not achieve the best results in all cases, there are many comparable results. In the US dataset, for the ROUGE-L f-measure and recall, BB25HLegalSum RoSum scores 0.3661 and 0.5330 in comparison to LSTM-with-w2v 0.3664 and 0.5358, respectively. The same can be observed in CA dataset for the ROUGE-2 f-measure criterion, where BB25HLegalSum RoSum scores 0.1979 in comparison to TextRank 0.2015. Therefore, the performance was encouraging even when our system did not outperform the baselines.

4.3 Experiment 2

In this experiment, we address the following research question: "*How does the length of the reference summary impact the recall and precision of the generated summary using BB25HLegalSum in the legal document summarization?*". We divided the reference summaries into different length intervals (number of characters), and aggregated the different scores for each interval. We analyzed the summaries generated using the three strategies (PoSum, FoSum, RoSum). The results of our evaluation provide insights into the effectiveness of different summarization techniques for different lengths of reference summaries.

Results for the US test data are presented in Figures 1, 2 and 3 for PoSum, RoSum and FoSum strategies, respectively. All tables provide ROUGE-1, ROUGE-2 and ROUGE-L precision, f-measure, and recall averaged values according to the reference summary length intervals in characters.

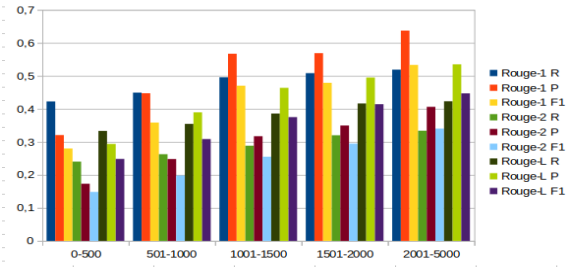

Figure 1: PoSum scores (US test Data)

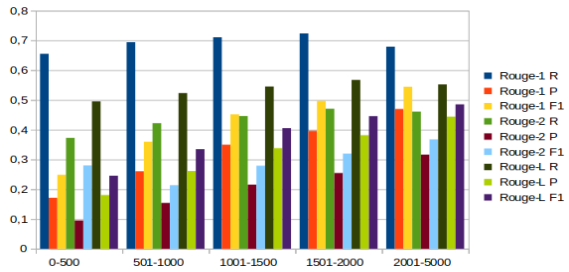

Figure 2: RoSum scores (US test Data)

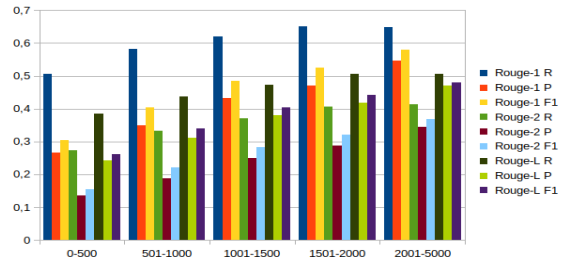

Figure 3: FoSum scores (US test Data)

As shown in Figure 1, the ROUGE-1 precision scores of PoSums more than double when comparing 0-500 to 2001-5000 reference summary range. As we can see in Figure 2, using the RoSum strategy, BB25HLegalSum behaved well on longer reference summaries, with a slight recall decrease on reference summaries longer than 2000 characters. The scores for the FoSums, displayed in Figure 3, present a more balanced score, having a positive impact on score values as the length of the reference summaries increase. Regardless of the summarization strategy, in general all scores increased with longer reference summaries.

We conclude that the length of the reference summary impacts the recall (RoSum) and precision (PoSum) scores of the generated summaries. On the other hand, the proposed solution performs better when the reference summary has a size larger

than average.

4.4 Experiment 3

4.4.1 Method

This final experiment targets the following research question "*Which type of document summary is more readable in the legal context: focused on precision, recall, or f-measure?*". To assess the most suitable strategy for generating legal documents summaries, we have selected specific bills from a set of US test data, and used them to assess the quality of the summaries produced by BB25HLegalSum for creating accurate and useful legal document summaries.

To be able to assess a significant amount of summaries about the trade-offs between completeness and conciseness, we adopted two criteria for selecting bills from the US dataset: a) the generated PoSums have at most 1000 characters, a criterion met by 30% of this type of summary; and b) the FoSums are larger by at least 250 characters the corresponding PoSums. The bills meeting these two criteria were then sorted in ascending order of difference in length between the respective FoSum and PoSum. We selected the first 50 bills of this ranking.

The assessment was performed by three lawyers, who received 50 highlighted bills to read, and the corresponding reference summary. The highlights were produced using the sentences of the respective PoSum, RoSum and FoSum, as described in Section 3.3. Table 3 displays a representative example of how the text was highlighted. It compares the reference summary and the highlighted text of bill 723 from US test data. The first column shows the reference summary, while the second column displays the bill's text with different colors.

Upon reading, they were asked to answer the following questions:
(1) Regarding the reference summary, do the three colored highlights outline the main arguments?
(2) Regarding the highlights in GREEN, do the highlights in BLUE or RED seem to bring new relevant information?
(3) Based on the highlights alone, can you understand the context, only the main arguments, or both?
(4) Among the three forms of highlighting, which method do you believe is the most suitable for lawyers and jurists and why? Consider the following options: (a) emphasis only in GREEN; (b) highlight in GREEN + BLUE; (c) griffin in GREEN + BLUE + RED. Write your observations in a few lines.

4.4.2 Results and Discussion

All participants answered *yes* to the first and second questions. One of the lawyers emphasized that the highlights helped better understand the context. For example, the green color (i.e., extracted from the PoSum) exposes the topic, while the red highlights (RoSum) complement it with more details, such as the bill's purpose. The usefulness of the blue griffin (FoSum) was perceived as limited.

Regarding the third question, two participants agreed on the possibility of inferring context and the main arguments from the highlights alone. The other subject responded that it is not possible to inquire about the main arguments by the highlights alone, but since they are being presented with the full document, the inference of context from reading the highlighted and its surrounding non-highlighted text is uncontested.

In the fourth question, all participants selected the three-colored method (GREEN + BLUE + RED) as the most appropriate one for all bills assessed, considering the perspective of lawyers and jurists. This encompasses the entire content of the RoSum with the inclusion of words related to PoSum and FoSum. They all have agreed that distinct colors help to understand the nuances and that despite conciseness being important, completeness is more useful in real-life court decisions. They justified the usefulness by noting that the highlights using all colors included in general the meaning of some of the terms, as well as relevant details such as objectives/purpose, criteria, and requirements. At times, it also included the name of the act. Hence, the level of detail provided was regarded as appropriate.

For instance, Bill 723 in Table 3 deals with the requirements for a particular relocation subsidy. It shows that the words in the PoSum and FoSum do not encompass key arguments. Examples are the requirement highlighted in blue in line 7 (not included in the PoSum) and the one highlighted in red in line 10 (not encompassed by the FoSum). On the other hand, the words from the RoSum sometimes bring unnecessary words, such as "For purposes of this section" given that it benefits completeness, rather than conciseness. However, this is deemed irrelevant in comparison to missing key arguments because it is a lot better for the lawyer to have all key arguments highlighted, even if some unneces-

Reference summary	Precision oriented (green), F-measure oriented (green + blue) and Recall oriented (green + blue + red) summaries
American Worker Mobility Act of 2014 - Authorizes the Secretary of Labor to grant a relocation subsidy of up to $10,000 to an individual who: (1) has been totally unemployed for at least 26 consecutive weeks. (2) has exhausted all rights to state or federal unemployment compensation. (3) has not received a relocation subsidy for the two-year period preceding the subsidy application. And (4) is able to work, available to work, and actively seeking work. Prescribes subsidy program requirements. Directs the Secretary to issue regulations to prevent program fraud or abuse.	SECTION 1. SHORT TITLE. This Act may be cited as the "American Worker Mobility Act of 2014". SEC. 2. RELOCATION SUBSIDIES FOR THE LONG-TERM UNEMPLOYED. (a) In General.–The Secretary of Labor may grant a relocation subsidy to an eligible individual who meets the requirements of this section. (b) Meaning of Eligible Individual.–For purposes of this section, an eligible individual is an individual who, as of the date of the application for a relocation subsidy under this section– (1) is totally unemployed and has been totally unemployed for at least 26 consecutive weeks; (2) has exhausted all rights to regular compensation under the law of a State or under Federal law with respect to a benefit year (excluding any benefit year ending before July 1, 2008); (3) has not received a relocation subsidy under this section in the 2-year period preceding such date of application; and (4) is able to work, available to work, and actively seeking work. (c) Requirements for Grant.–The Secretary of Labor may not grant a relocation subsidy to an eligible individual under this section unless the Secretary determines that– (1) the relocation subsidy will assist such individual in relocating within the United States, at least 60 miles from the individual's current residence, for the purpose of attaining employment; (2) such individual filed an application with the Secretary not later than January 1, 2019; and (3) such individual– (A) has obtained a bona fide offer of suitable employment affording a reasonable expectation of long- term duration in the area in which the individual wishes to relocate; or (B) wishes to relocate to an area that has an unemployment rate that is at least 2 percentage points less than the unemployment rate of the area of the individual's initial residence. (d) Amount of Subsidy.–A relocation subsidy granted to an eligible individual under this section shall be equal to the lesser of $10,000 or the amount that any contribution by a potential employer of the individual to the individual's relocation expenses is exceeded by the sum of– (1) 90 percent of the reasonable and necessary expenses incurred in transporting the worker, the worker's family, and household effects, plus (2) a lump sum equivalent to 3 times the individual's weekly benefit amount for the most recent benefit year (as such terms are defined in the State law), up to a maximum payment of $1,250. (e) Regulations.–Prior to granting any relocation subsidies under subsection (a), the Secretary of Labor shall issue regulations designed to prevent fraud or abuse relating to the program established under this Act. (f) No Additional Funds Authorized.–No additional appropriations are authorized for any fiscal year to carry out this Act. (g) Definitions.–For purposes of this section– (1) the term "regular compensation" has the meaning given the term in section 205(2) of the Federal-State Extended Unemployment Compensation Act of 1970 (26 U.S.C. 3304 note), as in effect prior to January 1, 2014; and (2) the term "suitable work"– (A) means suitable work as defined in the applicable State law for claimants for regular compensation; and (B) does not include self-employment or employment as an independent contractor. (h) Reports.–Not later than March 15 of each of calendar years 2015 and 2017, the Secretary of Labor shall submit a report to Congress that identifies, by geographic region– (1) the total number of relocation subsidies granted to individuals under this section during the calendar year preceding each such calendar year; (2) the total number of relocation subsidies granted to individuals pursuant to subsection (c)(3)(A) during such calendar year; (3) the total number of relocation subsidies granted to individuals pursuant to subsection (c)(3)(B) during such calendar year, and the number of such individuals who obtained employment within 1 month, 3 months, and 6 months, respectively, after the individual's relocation; (4) the average amount of a relocation subsidy granted during such calendar year; (5) the average distance traveled for relocation by each individual receiving a relocation subsidy during such calendar year; and (6) the number of individuals who received a relocation subsidy under this section during such calendar year and subsequently applied for unemployment benefits.

Table 3: Bill 723: Reference summary and highlighted bill according to the three strategies.

sary words are highlighted as well, than to have a lack of highlights, as it happens in the PoSum and FoSum summaries shown in Table 3.

Given this assessment, we observe that the PoSums and FoSums are shorter because they usually lack key arguments. In a legal document context, having a higher recall as a suitable criterion is important because failing to identify a relevant piece of information can have serious consequences, such as missing an essential element of context or failing to make a critical argument. Another important remark is that highlighting with multiple colors allows the reader to select pieces of information more easily, faster, and more intuitively.

5 Conclusions

In this paper we described BB25HLegalSum, a method that leverages BM25 and the combination of BERT clusters to summarize legal documents. We generate a summary using three strategies to understand the role of preciseness and completeness in legal documents: PoSum, RoSum, and FoSum. The summaries are presented to users within the original document with three-colored highlighted sentences that indicate the relevant sentences according to a summarization perspective.

Our experiments revealed that this unsupervised method outperforms the baselines for the BillSum dataset (US and CA test data), and that the length of the reference summary impacts the recall and precision of the generated summaries. The larger the reference summary, the better is the performance of our system. We also conducted a qualitative assessment with three lawyers, who evaluated that summaries that target higher recall (RoSum) are more appropriate in the legal context, since they avoid missing relevant information. They also positively evaluated the three-coloring approach proposed, arguing that it provides the context of the sentences and the relevance perspective.

Future work includes improving the combination of clusters to generate summaries, and a more comprehensive readability assessment.

Acknowledgements: This study was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq (131178/2020-2) and the PETWIN Project (FINEP financing and LIBRA Consortium).

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Sophia Althammer, Arian Askari, Suzan Verberne, and Allan Hanbury. 2021. Dossier@ coliee 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. *arXiv preprint arXiv:2108.03937*.

Deepa Anand and Rupali Wagh. 2019. Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University-Computer and Information Sciences*.

Arian Askari, Georgios Peikos, Gabriella Pasi, and Suzan Verberne. 2022. Leibi@ coliee 2022: Aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. *arXiv preprint arXiv:2205.13351*.

Sarthak Dalal, Amit Singhal, and Brejesh Lall. 2023. Lexrank and pegasus transformer for summarization of legal documents. In *Machine Intelligence Techniques for Data Analysis and Signal Processing: Proceedings of the 4th International Conference MISP 2022, Volume 1*, pages 569–577. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Shaz Furniturewala, Racchit Jain, Vijay Kumari, and Yashvardhan Sharma. 2021. Legal text classification and summarization using transformers and joint text features.

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

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2022. Improving kullback-leibler based legal document summarization using enhanced text representation. In *2022 IEEE Silchar Subsection Conference (SILCON)*, pages 1–5. IEEE.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2023. Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization. *Knowledge-Based Systems*, 264:110336.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Daniele Licari, Praveen Bushipaka, Gabriele Marino, Giovanni Comandé, and Tommaso Cucinotta. 2023. Legal holding extraction from italian case documents using italian-legal-bert text summarization.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Kaiz Merchant and Yash Pande. 2018. Nlp based latent semantic analysis for legal text summarization. In *2018 international conference on advances in computing, communications and informatics (ICACCI)*, pages 1803–1807. IEEE.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. 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.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686*.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Peter Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the highlight: incorporating active reading tools in a search as learning environment. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 229–238.

Edy Umargono, Jatmiko Endro Suseno, and SK Vincensius Gunawan. 2020. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In *The 2nd International Seminar on Science and Technology (ISSTEC 2019)*, pages 121–129. Atlantis Press.