CLSJUR.BR - A Model for Abstractive Summarization of Legal Documents in Portuguese Language based on Contrastive Learning

Alex Aguiar Lins[1] and **Cecilia Silvestre Carvalho**[2] and **Francisco das Chagas Jucá Bomfim**[3]
Daniel de Carvalho Bentes[4] and **Vládia Pinheiro**[5]
University of Fortaleza, Fortaleza, Brazil
[1]alexaguiarlins@yahoo.com.br, [2]ceciliacarvalhoo@gmail.com, [3]franciscojuca@gmail.com
[4]daniel.bentes@unifor.br, [5]vladiacelia@unifor.br

Abstract

In the legal domain, there has been a growing interest among Natural Language Processing (NLP) researchers in the Automatic Legal Document Summarization. However, legal documents differ from the general texts, as the former involves technical texts of a legal nature, which are generally longer and contain more sophisticated vocabulary than the general domain texts. In this article, we propose the CLSJUR.BR, a Contrastive Learning model for automatic and abstractive summarization of legal documents in Portuguese language, that applies the reference-free evaluation technique. CLSJUR.BR was trained and evaluated using the Ruling.BR corpus, composed of judicial decisions from the Supreme Federal Court of Brazil. The results indicating their good applicability to the task of summarizing legal documents.

1 Introduction

Automatic Text Summarization (ATS) is one of the most challenging tasks in Natural Language Processing (NLP), as its objective is to transform long texts into smaller texts that are understandable and that cover the most important points of the original text (Alomari, 2022). It can also be defined that ATS is the process that uses computer programs to retrieve relevant information from texts, to automatically generate summaries similar to those written by humans (Jindal and Kaur, 2020; Feijó, 2021). There are two main approaches to ATS. The first is extractive summarization, which performs summarization by selecting entire sentences directly from the source text, and the second approach is abstractive summarization, in which new sentences are generated in the summary, maintaining the ideas and facts of the text original (Alomari, 2022).

In the legal domain, given the large quantity of legal documents available, both on the internet and in court systems, there has been a growing interest among NLP researchers in the automatic processing of legal texts. According to Turtle (1995 apud Feijó (2021)), legal documents have some distinctive characteristics compared to other types of texts (for example, newspaper articles or scientific articles), namely: (i) they tend to be longer ; (ii) they have their own internal structure; (iii) they have many technical and specific terms from the legal domain (e.g. *ratio decidendi, sub judice, In dubio pro reo, ex post facto, amicus curiae*); (iv) they generally mention many ambiguous terms that lead to different legal interpretations; and (v) they reference citations to other legal processes and norms, which play a prominent role in the legal domain (by supporting decisions, arguments, challenges and petitions).

Regarding the task of Automatic Legal Document Summarization (ALDS), all of the above characteristics contribute to greater complexity of legal documents summarization models (Kanapala; Jannu; Pamula, 2019; Jain; Borah; Biswas, 2021). Especially, the length and quantity of legal documents from a single legal case harm the performance of SOTA (State-Of-The-Art) models for ATS (e.g. encoder-decoder based models), given the limitation of possible tokens to be processed.

ALDS has a multitude of applications, from simplifying the work of lawyers, who need to search a huge set of legal documents, to supporting judges in their judicial decisions (Anand and Wagh, 2019; Jain; Borah; Biswas, 2021). In practice, legal documents and processes are still summarized manually by legal experts (Jain; Borah; Biswas, 2021). In the Brazilian Legal System, thousands of cases are received per year. According to the CNJ (National Council of Justice)'s 2023 "Justice in Numbers" report, Brazil has 81.4 million cases in progress, and each court case can contain hundreds of documents with dozens of pages. In this scenario, there is an urgent need for good models to automate the process of summarizing legal documents, as it makes it possible to optimize work and increase the productivity of specialists and, consequently, improve the efficiency of the courts (Bhattacharya et al., 2019).

SOTA models for ATS use Deep Learning in the automatic abstractive summarization of texts, mainly those based on encoder-decoder or transformer. For the English language, SimCLS (Liu and Liu, 2021) stands out for general domain documents, which applies a Contrastive Learning (CL) approach with the reference-free evaluation technique. For legal documents, especially in Portuguese, LegalSumm (Feijó and Moreira, 2021) applies CL but through the technique of generating false examples. More recently, with the popularization of LLMs (Large Language Model) with satisfactory performance in several NLP tasks, including text summarization (Adams et al., 2023), there is an urgent need to evaluate such models for summarization of legal documents in Portuguese Language.

In this context, this work presents CLSJUR.BR, a Contrastive Learning model for automatic summarization of legal documents in Portuguese language, that applies the reference-free evaluation technique aiming to improve this very important task for Legal AI (Legal Artificial Intelligence) systems. The research questions that guided the development of this work were:

RQ1 – Is the Contrastive Learning approach with the reference-free evaluation technique more effective for ALDS?

RQ2 – Does the use of language-specific language models improve the performance of an ALDS system for the Portuguese Language?

RQ3 – How much do general LLMs improve the performance of an ALDS system for the Portuguese Language?

To evaluate CLSJUR.BR, the Ruling.BR corpus, composed of judicial decisions from the Supreme Federal Court of Brazil, and several models were used in the experiments. The models were the multilingual models BERT (Devlin et al., 2019) and mBART (Liu et al., 2020); the model refined for the Portuguese language - Bertimbau (Souza; Nogueira; Lotufo, 2020); and a specific language model for the legal domain in Portuguese Language - LegalBert-PT (Silveira et al., 2023). The results of the proposed model were compared with baseline systems, with SOTA systems for ALDS in Portuguese language and with LLMs (GPT3.5, GPT4 (OpenAI, 2023) and Llama 2 (Touvron et al., 2023)). CLSJUR.BR presented results that surpassed, among others, LegalSumm and the LLMs models, when dealing with legal documents in Portuguese, indicating their good applicability to the task of summarizing legal documents.

2 Related Works

Traditionally, sequence-to-sequence neural models - Seq2Seq (Sutskever et al., 2014) have been widely used in text generation tasks, such as abstractive summarization and machine translation. These models are generally trained under the Maximum Likelihood Estimation (MLE) structure, which, in practice, adopts teacher-forcing (Williams and Zipser, 1989), which maximizes the probability of each token, given the current state of the model. Nevertheless, this approach has some problems. The first arises during inference (testing phase), the legitimate passed target tokens are not available and are therefore replaced by tokens generated by the model itself, generating a discrepancy between the way the model is used in training and how it is used in testing, introducing a gap between training and testing called exposure bias by Ranzato et al. (2016). The second problem encountered is the gap between the objective function or loss function (Liu and Liu, 2021; Bengio et al., 2015). This is and the evaluation metrics, as the objective function is based on local token-level predictions, while the evaluation metrics (e.g. ROUGE (Lin, 2004) metrics) compare the similarity holistic between the golden standard references and the system outputs (Liu and Liu, 2021).

Minimum Risk Training, as an alternative to resolve this gap between training and testing, has also been used in language generation tasks (Shen et al., 2016; Wieting et al., 2019). However, the estimated loss accuracy is limited by the number of sampled outputs. Paulus et al. (2018) and Li et al. (2019) propose the use of the Reinforcement Learning (RL) paradigm to mitigate the gap between training and testing. Although RL training makes it possible to train the model with rewards based on global predictions and closely related to the evaluation metrics, it presents the challenges inherent to RL such as the problem of noise in gradient estimation (Greensmith et al., 2004), which, often, makes training unstable and sensitive to hyperparameters (Liu and Liu, 2021). In order to overcome the challenging and complex optimization process of RL-based methods, the work of Liu and Liu (2021), inspired by Zhong et al. (2020) and Liu, Dou and Liu (2021), proposed SimCLS to generalize the Contrastive Learning (CL) paradigm (Chopra et al., 2005) through the reference-free evaluator technique, introducing an abstractive summarization approach that directly optimizes the model with the corresponding evaluation metrics, thus mitigating the gaps between the training and testing stages. Even though some related works, such as that of Lee et al. (2021) and Pan et al. (2021), proposed the introduction of contrastive loss as an addition to MLE training, Liu and Liu (2021) chose to disentangle the contrastive loss and MLE loss functions, introducing them in different parts of the structure of their framework (Liu and Liu, 2021). SimCLS was evaluated on the CNNDM (Hermann et al., 2015; Nallapati et al., 2016) and XSUM (Narayan et al., 2018) corpus and obtained better results than the approaches that used BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020a).

For the ALDS task in Portuguese language, LegalSumm (Feijó and Moreira, 2021) applies Contrastive Learning, but through the generation of false examples, which aims to force the model to learn to distinguish true and false chunk-summary pairs. The author evaluated this model based on the Ruling.BR corpus and obtained better results than the BertSumExt (Liu and Lapata, 2019), BertSumAbs (Liu and Lapata, 2019) and BART approaches. These models are subject to the inherent limitation of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) in processing input texts with a length of up to 512 tokens. Therefore, this length must be divided to compose the source text and the summary in summarization tasks. In the case of LegalSumm, 400 tokens remained to be used as source text, failing to include a large part of the texts in the summary.

3 A Golden Collection for ALDS

In this work, the Ruling.BR (Feijó and Moreira, 2018) was used as a Golden Collection (GC) for ALDS, which is a corpus in Portuguese composed of 10,623 judicial sentences from the Federal Supreme Court, the highest body of the Brazilian judiciary, dated between 2012 and 2018. The Ruling.BR's judicial sentences are structured into the following topics: Summary, Report, Vote and Judgment.

The National Council of Justice (CNJ) of Brazil defines guidelines for preparing summaries. According to this document, the topic "Summary" of a judgment summarizes and discloses the content of judicial decisions, summarizing the legal reasons and the factual consequences relating to the *res judicata*. It is a summary of the main points discussed in each case and how the judges decided. Therefore, the topic "Summary" is used as the reference summary in the evaluation of ALDS models. The topic "Judgment", as defined by the Superior Electoral Court (TSE) Portal, is the manifestation of a collegial judicial body that reveals a legal position, based on arguments about the application of a certain right to a specific factual situation. The topic "Report", in turn, contains the narration of the facts of the process and the law in question. It is in the Report that the principles of fact and law are established, serving as the basis for judgment. Finally, the topic "Vote" is the manifestation of each member of the panel's understanding of the case being judged. This topic is the largest part, corresponding to 69% of the complete judicial sentence.

A descriptive analysis of the tokens of each part of the judicial sentences contained in this GC was carried out. The judicial sentence tokens were identified using the Bertimbau tokenizer (Souza; Nogueira; Lotufo, 2020). Table 1 presents the total number of tokens for each topic, the average number of tokens, the standard deviation (std) and the distribution of tokens by quartile. For example, the summaries have an average of 363 tokens, with 75% of them containing up to 424 tokens. In line with Table 1, Figure 1 illustrates that the number of tokens in the summaries in the first, second and third quartiles are approximate, however, in the fourth quartile we have observations reaching up to 776 tokens, above that we have the outliers that represent 7.3% of summaries.

	Summary	Report	Vote	Judgment
Average	363	956	3,111	93
Std	300	1,336	5,329	50
Min	29	70	89	44
25%	188	275	1,240	75
50%	288	622	1,970	81
75%	424	1,206	3,307	94
Max	4,842	62,806	125,856	1838
Total	3,855,614	10,154,195	33,044,092	989,175

Table 1: Golden Collection Ruling.BR Descriptive Statistics

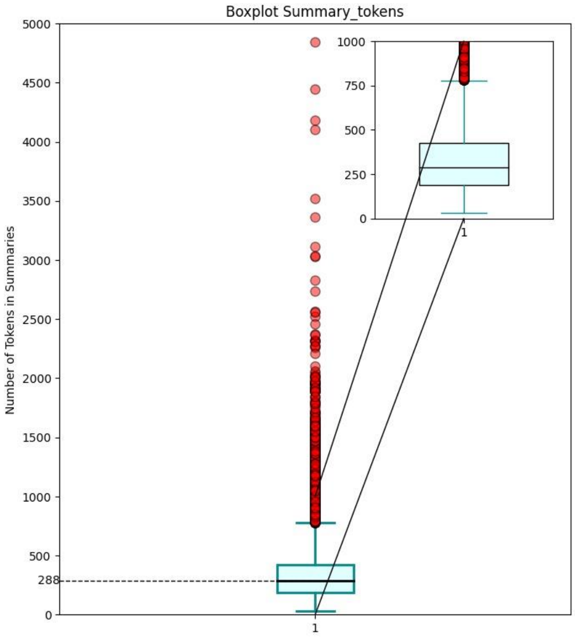

Figure 1: Boxplot chart of Summary tokens.

Furthermore, a *n-gram* analysis was carried out in the GC and their overlap between the summaries and other topics of the judicial sentences. Table 2 show the percentages of common bigrams between the summary, report, vote and judgment. The topic "Vote" is the one that contains the most bigrams in common with the Summary topic (52.52%). From this analysis, we can state that, on average, 41.06% of the words in the topic "Summary" do not appear in other parts of the judicial sentence.

% of / is contained in	%Summary	%Report	%Vote	%Judgment
Report	26.07%	-	-	-
Vote	52.52%	-	-	-
Judgment	5.84%	-	-	-
-	41.06%	-	-	-
Summary	-	11.93%	11.89%	7.74%

Table 2: Percentages of common bigrams between summaries and other topics of the judicial sentences.

4 CLSJUR.BR – A Model for Abstractive Summarization of Legal Documents in Portuguese language based on Contrastive Learning

In this work, we propose CLSJUR.BR, a model for abstractive summarization of legal documents in Portuguese language, based on Contrastive Learning. Inspired by SimCLS (Liu and Liu, 2021), the CLSJUR.BR architecture is divided into three stages: Pre-processing, Generation of Candidate Summaries and Evaluation of Summaries and Election of the Final Summary. (see Figure 2)

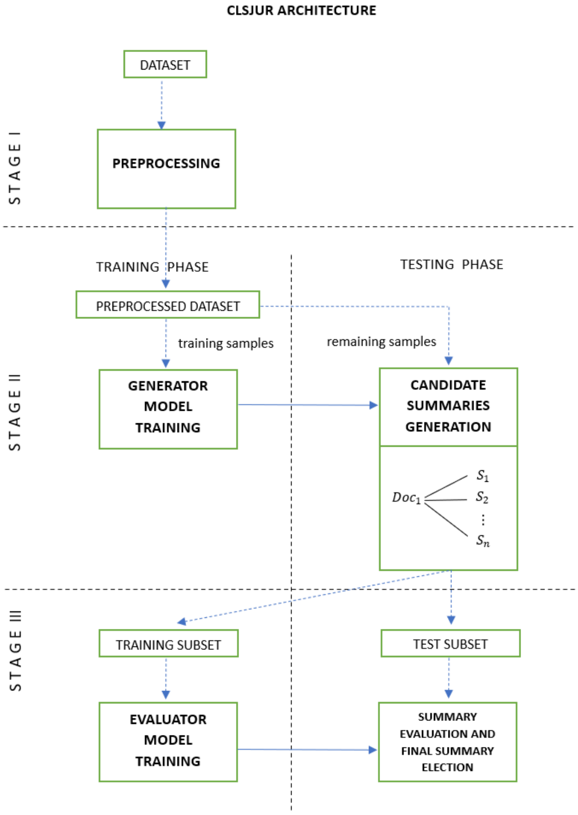

Figure 2: CLSJUR.BR Architecture.

Stage I – Pre-processing:

At this stage, adjustments are made to the documents in order to make them compatible with the model: removal of special characters (e.g. quotation marks); adjustment to the nested structure of the file; union of the topics of each document in order to compose a single text to be summarized (bearing in mind that judicial sentences are divided into several topics); and distribution of examples into different subsets, according to the following scheme –

- Training/Validation and Test sets for the Generator model (Stage II).
- Training/Validation and Test sets for the Evaluator model (Stage III).

Stage II – Generation of Candidate Summaries:

At this stage there is the training phase and the summarization phase. In the training phase, a pre-trained Seq2Seq model, for example in Portuguese (mBART and T5 (Raffel et al., 2020)), is refined (fine-tuned) using pairs of input (legal document) and output (summary) sequences and learns to generate multiple candidates for summaries.

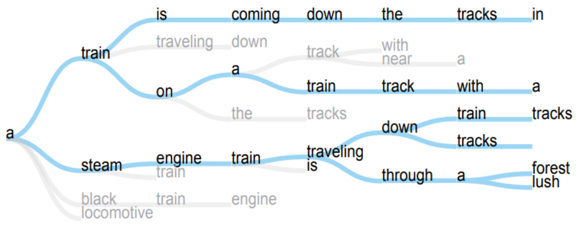

Figure 3: Example of sequence diversification provided by a Sampling Strategy (Vijayakumar et al., 2016).

In the summarization phase, the Generator model produces *n* candidate summaries for each text in the test dataset. This sampling strategy allows the Generator model to predict more than one word per token, according to the probability distribution, and produces different sentences. In the end, *n* variations of summaries are generated for each text, increasing the chance of producing a final summary closer to the ideal summary (Freitag and Al-Onaizan, 2017). Figure 3 illustrates the diversification of generated word sequences. For example, from the "a train" tokens, the sentence "a train is coming down the tracks in..." and the sentence "a train on a train track with a" can be generated.

At the end of this stage, two subsets of examples are made available with their respective candidate summaries that will serve as inputs for the next stage: Training Subset and Test Subset for the Stage III.

Stage III – Evaluation of Summaries and Election of the Final Summary:

At this stage, the model evaluates the candidate summaries generated in the previous stage, assigning each one a score and choosing the best scored as the final summary.

In the training phase, the Evaluator model is fine-tuned using the Stage III training subset and through a variation of the Contrastive Learning technique, called reference-free evaluator. In this case, a ranking loss, L, is introduced for the evaluation function h (·), which has the following formula:

$$L = \sum_{i} \max(0, h(D, \check{S}_i) - h(D, \hat{S})) \\ + \sum_{i} \sum_{j>i} \max(0, h(D, \check{S}_j) - h(D, \check{S}_i) + \lambda_{ij}),$$

where Ŝ is the reference summary, Š1, . . ., Šn is the list of candidate summaries descendingly sorted by M (Ši, Ŝ), M is the ROUGE automated evaluation metric, λij = (j -i) ∗ λ is the corresponding margin defined according to Zhong et al. (2020), and λ is a hyperparameter. The function h (·), which its formula is below, is calculated by instantiating a pre-trained classifier model that encodes sometimes Ši and D vectors and sometimes Ŝ and D vectors, separately, and applies the cosine similarity between the two, obtaining a score.

$$h(S_i, D) = \frac{\sum_{j=1}^{n} S_{ij} \cdot D_j}{\sqrt{\sum_{j=1}^{n} S_{ij}^{2}} \cdot \sqrt{\sum_{j=1}^{n} D_j^{2}}}$$

where,

n is the size of bigger vector;
S_{ij} is the *j*-term *tf-idf* weight of S_i ;
D_j is the *j*-term *tf-idf* weight of D;

After training, the summaries are evaluated by the function h (·), responsible for assigning different scores to them, based only on the similarity between the source document (D) and the candidate summary (Si). Then, the candidate with the highest score is selected to compose the final summary (S), according to the formula below.

$$S = \underset{S_i}{\operatorname{argmax}}\, h(S_i, D).$$

The diagram in Figure 4 illustrates the operation of CLSJUR.BR in Stage III and its training and testing steps.

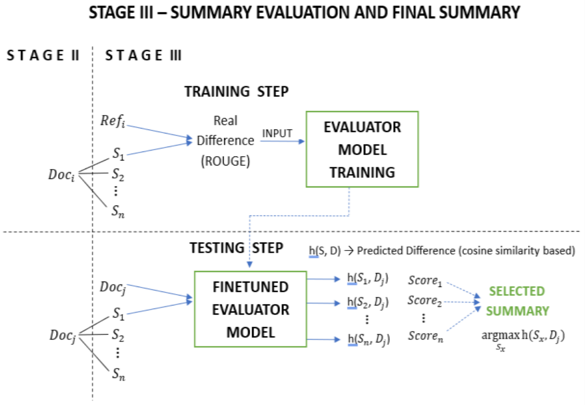

Figure 4: Stage III – Evaluation of Summaries and Election of the Final Summary.

5 Experimental Evaluation

5.1 Methodology

1. Dataset preparation

At this stage, the Ruling.BR (in Json format) was pre-processed as follows. First, the “quote” type characters are removed and the nested structure of the Json file is adjusted (removal of an unnecessary object at the first level). Then, the 10,623 examples are distributed into the following datasets, the same Ruling.BR examples adopted in LegalSumm (SOTA system):

- Training/Validation of Stage II: 6,998 examples (65.88%);
- Testing of Stage II: 3.625 examples, where:
    - 1,500 examples (14.12%) to Stage III Training/Validation;
    - 2,125 examples (20%) to Stage III Testing.

Finally, the topics of a court ruling were united in a single document, in the following order: Report, Vote and Judgment, following the proposal in Feijó (2021). To validate this design decision, test experiments were carried out alternating the order of topics and the best results indicated this as the best joining order (Report, Vote and Judgment).

2. Definition of models and parameters

For the Summary Generator model (Stage II), mBART (Liu et al., 2020) was used, a multilingual version of BART (Lewis et al., 2020) that includes the Portuguese Language, and the Diverse sampling strategy Beam Search (Vijayakumar et al., 2016). For Stage III, as Summary Evaluator models, several models were used in the experiments, these are: BERT (Devlin et al., 2019), Bertimbau (Souza; Nogueira; Lotufo, 2020), LegalBert-PT (Silveira et al., 2023) and mBART (Liu et al., 2020), one for each CLSJUR.BR evaluation scenario.

It is noted that both models, Summary Generator and Evaluator, are trained in 5 epochs and use the k-fold cross validation technique (k=5). The following parameters are defined in each evaluation scenario: maximum input token size (TME), maximum output token size (TMS) and beam number (number of candidate summaries).

3. Definition of Evaluation Scenarios

Four evaluation scenarios were defined to validate the Summary Evaluator Model:

- EXP 1 - BERT, multilingual version (with TME = 512);
- EXP 2 - Bertimbau, pre-trained in Portuguese (with TME = 512);
- EXP 3 – mBART, multilingual (with TME = 1024);
- EXP 4 - LegalBert-PT, a refined language model for legal documents in Portuguese (with TME = 512).

It is noteworthy that TMS = 256 was adopted in all experiments, following that adopted in Feijó and Moreira (2021), and beam number = 16, following Liu and Liu (2021).

With the advent of LLMs (GPT-3.5, GPT4 (OpenAI, 2023), and Llama 2 (Touvron et al., 2023)), the following evaluation scenarios were created for comparison purposes with the CLSJUR.BR, proposed here. They are:

- EXP 5 – in this scenario the LLM “gpt-3.5-turbo” from the GPT-3.5 series was used, limited to 4,096 tokens. The prompt used to generate the summaries followed a zero-shot learning approach as follows “Generate a summary of a maximum of 256 tokens from the following text: <Report> <Vote> <Judgment>”;

- EXP 6 – For financial cost reasons, in this scenario, 100 examples were selected from the test set, specifically the 50 best and 50 worst test cases, based on the ROUGE-2 metric, because it presented the smallest difference between winning system and SOTA system. The model used was GPT4 with 8,192 limitation tokens. The instruction and input were limited to 7.800 tokens, to ensure that the total input and output (generated summary) remain within the maximum token limit. The prompt also followed a zero-shot learning approach as follows “You are a legal professional and will receive the report, vote and judgment on a judicial decision. The summary is a resume of the content of the court decision. Make a summary based on the data presented: <Report> <Vote> <Judgment>”;
- EXP 7 – in this scenario the LLama2 model was used, with the same set of texts and input instructions as EXP 6. The limit of input tokens used was 1,524 and the number of output tokens was set at 512;
- EXP 8 – in this scenario the GPT4 model was used with the 100 examples from EXP 6, but in a few-shot prompt approach, based on Brasil (2021). The example in the prompt was composed by <Report> <Vote> <Judgment> followed by the <summary>”. The instruction and input were limited to 7.800 tokens, to ensure that the total input and output (generated summary) remain within the maximum token limit.

5.2 Results and Discussion

Table 3 presents the results obtained from experiments with CLSJUR.BR, using the Test dataset of the Stage III with 2,125 examples, compared to baseline and optimal approaches.

The baseline and optimal reference systems implement only Stages I and II (Summary Generation) and select summaries based on their ROUGE scores. The Oracle Max system consists of selecting the summary with the highest score, being considered an optimal system and represents an upper limit for ALDS systems. The Oracle Average system selects the summary ROUGE score closest to the average calculated across candidates. The Oracle Random system chooses a summary randomly among the candidates.

Considering *RQ2* (*Does the use of language-specific language models improve the performance of an ALDS system for the Portuguese Language?*), it appears that refined models in the Portuguese language and in the legal documents (EXP2 and EXP4) present better results than the BERT multilingual model (EXP 1). However, the mBART model (EXP 3), which supports a greater number of input tokens with TME = 1024, despite not being a pre-trained model exclusively in Portuguese, outperformed all other models, due to its greater text coverage. It is worth mentioning that, among the 2,125 examples, 104 examples have less than 1,024 tokens. Considering only this subset of the test dataset, CLSJUR.BR LegalBert-PT version (EXP 4) achieved ROUGE-1 = 0.5605, supplanting CLSJUR.BR mBart version (EXP 3) with ROUGE-1 = 0.5455, indicating that, in smaller texts, the LegalBert-PT is better and the token limitation of this model (512 tokens) impacted its performance.

In relation to the reference approaches, CLSJUR.BR did not surpass the optimal Oracle Max upper limit, in the same way as Liu and Liu (2021) but presented better results than the baseline systems (random selection or by the average of candidates – Oracle Random and Oracle Average, respectively).

Evaluation Scenario	**ROUGE-1 (F1)**	**ROUGE-2 (F1)**	**ROUGE-L (F1)**
EXP 1 - CLSJUR.BR - Bert 512 tks	0.4773	0.2882	0.4614
EXP 2 - CLSJUR.BR - Bertimbau 512 tks	0.4856	0.2982	0.4694
EXP 3 CLSJUR.BR - mBart 1024 tks	**0.4955**	**0.3066**	**0.4789**
EXP 4 CLSJUR.BR - LegalBert-PT 512 tks	**0.4863**	**0.2991**	**0.4699**
EXP 5 GPT 3.5 – 4096 tks	0.3150	0.1276	0.2984
Oracle Max 512/1024 tks (optimal)	0.5485 / 0.5688	0.3669 / 0.3883	0.5332 / 0.5526
Oracle Average 512/1024 tokens (baseline)	0.4016 / 0.4171	0.2242 / 0.2362	0.3860 / 0.4005
Oracle Random 512/1024 tokens (baseline)	0.3997 / 0.4200	0.2235 / 0.2359	0.3846 / 0.4029

Table 3: Results of the Evaluation Scenarios using CLSJUR.BR and of the baseline and optimal systems.

Table 4 compares the best CLSJUR.BR Evaluator models (mBart – EXP 3 and LegalBert-PT – EXP 4), with ALDS SOTA systems for Portuguese language - LegalSumm (abstractive summarization) and LetSum (extractive summarization) (Farzindar and Lapalme, 2004). It is noted that all systems were tested on the same test subset - 2,125 examples.

SYSTEM	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)
CLSJUR.BR - mBART)	0.4955	0.3066	0.4789
CLSJUR.BR - LegalBert-PT	0,4863	0,2991	0,4699
LegalSumm (SOTA)	0.43	0.27	0.35
LetSum (SOTA)	0.2338	0.0950	0.2136

Table 4: Comparison between CLSJUR.BR Results and SOTA systems - LegalSumm and LetSum.

The results in Table 4 allow for some analyzes in order to answer *RQ1* (*Is the contrastive learning approach with the reference-free evaluation technique more effective for ALDS?*). The best CLSJUR.BR models (EXP 3 and EXP 4) supplanted SOTA LegalSumm, improving the best abstractive summarization approach for Portuguese in Ruling.BR GC. Thus, there is an advantage of using the “free-reference evaluation” technique over the “generation of false examples” technique in the context of ALDS in Portuguese, answering *RQ1*.

To answer *RQ3* (*How much do general LLMs improve the performance of an ALDS system for the Portuguese Language?*), in addition to EXP5 of Table 3, Tables 5 and 6 present the results of scenarios EXP3, EXP6, EXP7 and EXP8, considering the 50 best and 50 worst test cases, based on the ROUGE-2 metric. For the 50 best cases analyzed, the LLMs GPT4 and Llama2 present much lower performance than the CLSJUR.BR-mBart (EXP 3) (see table 5). On the contrary, for the 50 worst cases analyzed, the GPT4 model, in both zero-shot and few-shot learning approaches, shows an improvement in relation to the CLSJUR.BR-mBart model (EXP 3) (see table 6). Analyzing the 100 cases of these experiments, it is known that the average number of tokens in the 50 worst cases is 6,449 tokens, much higher than the average number of tokens in the 50 best cases (1,746 tokens), indicating that the CLSJUR.BR model has difficulty in summarizing long texts. It is important to note that for the 50 worst cases (table 6), with the highest average number of tokens, EXP6 (zero-shot) obtained better results than EXP8 (few-shot), contrary to what occurred in the 50 best cases. This can also be explained by the fact that few-shot prompting has a greater number of tokens due to the example sent in the request to the LLM.

Evaluation Scenario	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)
CLSJUR.BR-mBART (EXP 3	0.9475	0.9307	0.9473
EXP6 - GPT4 (zero-shot learning)	0.3793	0.1520	0.2358
EXP7 - Llama2 (zero-shot learning)	0.1706	0.0672	0.1251
EXP8 - GPT4 (few-shot learning)	0.3967	0.1801	0.2498

Table 5: Comparison of the top 50 ROUGE-2 results between the GPT4, Llama2 and CLSJUR.BR best model.

Evaluation Scenario	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)
EXP3 CLSJUR.BR-mBART)	0.2241	0.0385	0.2028
EXP6 - GPT4 (zero-shot learning)	0.2783	0.0968	0.1666
EXP7 - Llama2 (zero-shot learning)	0.1543	0.0400	0.1069
EXP8 - GPT4 (few-shot learning)	0.2464	0.0791	0.1554

Table 6: Comparison of the 50 worst ROUGE-2 results between the GPT4, Llama2 and CLSJUR.BR best model.

Furthermore, we have included the test set examples with their respective generated summaries, in the following repository folder: *https://github.com/duchuchebu/CLSJURBR*.

6 Conclusion and Future Works

This work proposes CLSJUR.BR - a model for automatic abstractive summarization of legal documents in Portuguese language, which applies the Contrastive Learning approach in two stages: “Generation of Candidate Summaries” and “Evaluation of Summaries and Election of the Final Summary”. CLSJUR.BR was trained and evaluated based on a data set composed of judicial decisions on cases from a court of last instance in the Brazilian Legal System. The results showed that, within the scope of legal summarization for the Brazilian Legal System, the model's characteristic of generating several candidate summaries for each document, through the sampling generation strategy, made it possible to obtain better summaries than just generating a single summary. Furthermore, it was found that the evaluation technique used by the model, free-reference evaluation, allowed the selection of summaries closer to the optimum, in relation to other strategies tried. Finally, refining models for Portuguese language and legal documents enables better results in the ALDS task. As an extension of this work, it is important to evaluate large language models (LLMs) for the ALDS task with other prompting learning strategies (e.g., dense prompts), as well as evaluate whether a refinement process with legal documents would improve the performance of such models. For future works, it is suggested that factuality and named entities (NER) be considered when training and refining the proposed model, so that the model can learn the importance of facts and entities in relation to summaries, especially those related to legal norms and case law. Furthermore, it is suggested that examples containing outliers be pruned relative to the number of tokens in the summary topic and the full document.

References

Adams, G., Fabbri, A., Ladhak, F., Lehman, E., and Elhadad, N. (2023). From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting. ArXiv. Retrieved from https://arxiv.org/abs/2309.04269.

Ambedkar Kanapala, Srikanth Jannu, Rajendra Pamula. Summarization of legal judgments using gravitational search algorithm. Neural Computing and Applications, Springer Nature 2019. 2019.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. CoRR, abs/1610.02424.

Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, Izzat Alsmadi. Deep reinforcement and transfer learning for abstractive text summarization: A review. Computer Speech & Language, Volume 71, 2022, 101276, ISSN 0885-2308.

Brasil, Conselho Nacional De Justiça; UERJ REG. Diretrizes para a elaboração de ementas. Brasília: CNJ, 2021. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2021/09/diretrizes-elaboracao-ementas-uerj-reg-cnj-v28092021.pdf. Acesso em: 6 nov. 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 1, Article 140 (January 2020), 67 pages.

Deepa Anand, Rupali Wagh, Effective deep learning approaches for summarization of legal texts, Journal of King Saud University - Computer and Information Sciences, 2019, <http://www.sciencedirect.com/science/article/pii/S1319157819301259>.

Diego de Vargas Feijó; Viviane P. Moreira. Improving abstractive summarization of legal rulings through textual entailment. Artificial Intelligence and Law (2021). https://doi.org/10.1007/s10506-021-09305-4

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. Journal of Machine Learning Research, 5(9).

Farzindar, A.; Lapalme, G.: Letsum, an automatic legal text summarizing system. Legal knowledge and information systems, JURIX, pp. 11–18 (2004).

Feijó, Diego de Vargas. Summarizing Legal Rulings. Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação. Porto Alegre, 2021

Feijó, Diego de Vargas; Moreira, Viviane Pereira. 2018. Rulingbr: A summarization dataset for legal texts. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonc¸alo Oliveira, and Gustavo Henrique Paetzold, editors, Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pages 255–264.

Freitag, M., Al-Onaizan, Y. (2017). Beam search strategies for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pp. 56–60, Vancouver. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

J. Zhang, Y. Zhao, M. Saleh, P.J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. Proceedings of the 37th International Conference on Machine Learning. (2020), pp. 11328-11339

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL)

Jain, Deepali; Borah, Malaya Dutta; Biswas, Anupam. Summarization of legal documents: Where are we now and the way forward. Department of Computer Science and Engineering, National Institute of Technology Silchar, Assam, 788010, India. 2021.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Toma´s Ko ˇ cisk ˇ y, Edward Grefen- ´ stette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Lin, Chin-Yew. Rouge: A package for automatic evaluation of summaries ACL, in: Proceedings of Workshop on Text Summarization Branches Out Post Conference Workshop of ACL, 2004, pp. 2017–05.

Liu, Y.; Lapata, M. 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). [S.l.: s.n.], 2019. p. 3721–3731.

Liu, Yinhan & Gu, Jiatao & Goyal, Naman & Li, Xian & Edunov, Sergey & Ghazvininejad, Marjan & Lewis, Mike & Zettlemoyer, Luke. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics. 8. 726-742. 10.1162/tacl_a_00343.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6197–6208, Online. Association for Computational Linguistics.

OpenAI. GPT-4 Technical Report. arXiv arXiv:2303.08774, 2023.

Paheli Bhattacharya, Kaustubh Hiware1, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh . A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. Springer Nature Switzerland AG 2019 L. Azzopardi et al. (Eds.): ECIR 2019, LNCS 11437, pp. 413–428, 2019.

R. J. Williams and D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," in Neural Computation, vol. 1, no. 2, pp. 270-280, June 1989, doi: 10.1162/neco.1989.1.2.270.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, C¸ aglar Gulc¸ehre, and Bing Xiang. 2016. ˘ Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ranzato, MA., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence level training with recurrent neural networks. Paper presented at 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. ICLR .

S. G. Jindal and A. Kaur, "Automatic Keyword and Sentence-Based Text Summarization for Software Bug Reports," in IEEE Access, vol. 8, pp. 65352-65370, 2020.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent Neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 1171–1179.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In International Conference on Learning Representations.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., Furtado, V. (2023). LegalBert-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain. In: Naldi, M.C., Bianchi, R.A.C. (eds) Intelligent Systems. BRACIS 2023. Lecture Notes in Computer Science(), vol 14197. Springer, Cham. https://doi.org/10.1007/978-3-031-45392-2_18

Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. 2019. Deep reinforcement learning with distributional semantic rewards for abstractive summarization. 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 6038–6044, Hong Kong, China. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Souza, F., Nogueira, R., Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science(), vol 12319. Springer, Cham. https://doi.org/10.1007/978-3-030-61377-8_28

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 539–546. IEEE.

Sutskever, Ilya & Vinyals, Oriol & Le, Quoc. (2014). Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems. 4.

Turtle, H. Text retrieval in the legal world. Artificial Intelligence and Law, Springer, v. 3, n. 1, p. 5–54, 1995.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACLIJCNLP), Virtual.

Yixin Liu, Zi-Yi Dou, and Pengfei Liu. 2021. RefSum: Refactoring neural summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1437–1448, Online. Association for Computational Linguistics.