

Automatic Simplification of Legal Texts in Portuguese Using Machine Learning

Alexandre ALVES^a, Péricles MIRANDA^b, Rafael MELLO^{a,b} and
André NASCIMENTO^{a,b,1}

^a *Cesar School, Recife, Brazil*

^b *UFRPE, Recife, Brazil*

Abstract. Texts produced by the Brazilian judiciary have a complex and technical vocabulary, with elaborate use of the Portuguese language and many legal terms difficult to be understood, generating a barrier in communication between the judiciary and the population. In this sense, the Automatic Text Simplification (ATS), activity of the Natural Language Processing (NLP) area, can be applied to improve the readability of these types of text using specialized algorithms, and promote scalability in simplifying them, in view of the great demand in the courts. In this context, this article presents an evaluation of four methods of state of the art in text simplification, evaluated according to readability metrics, to improve the quality of existing texts in the judicial summaries, dataset containing 100 summaries of the Federal Regional Court of the 5th Region (TRF5) and another 100 of the Federal Supreme Court (STF). The methods MUSS(EN), MUSS(Pt), Transformers and NMT + Attention were tested, and the results of the simplifications exceeded the FRE readability index of the original texts, making them more readable.

Keywords. automatic text simplification, machine learning, natural language processing

1. Introduction

Legal documents in Brazilian judicial proceedings, such as judgments, appeals, documents, orders, decisions, sentences among others, use elaborate vocabulary in the Portuguese language and heavy legal jargon. This makes them difficult to read and understand for the average citizen [1]. In Brazil, according to data from the National Literacy Indicator (INAF), 88% of the population between 15 and 64 years old have some problem of reading comprehension [2]. It should be noted that the National Council of Justice (CNJ) has sought to strengthen the relations of the judiciary with society, e.g., Resolution N° 325, of June 29 (2020)² that "refers to the adoption of communication strategies and objective, agile and easy-to-understand procedures, aiming at transparency and strengthening the judiciary as an institution that guarantees rights". In addition, the large volume of documents produced daily in Brazilian courts, demands automatic and scalable solutions.

¹Corresponding Author: André Nascimento, UFRPE, Brazil; E-mail: andre.camara@ufrpe.br.

²<https://atos.cnj.jus.br/files/original182343202006305efb832f79875.pdf>

Automatic Text Simplification (ATS) comprise solutions that aims to produce a simpler version of the original text through lexical and syntactic transformations, without altering the original meaning of the text, and consequently improve communication and easy understanding of the texts [3]. In this context, the text simplification process is configured as an alternative to the Brazilian courts in the implementation of more inclusive and effective communication, making it more accessible, specially for people with aphasia, dyslexia and autism [4,5,6,7,8]. In addition, shorter sentences are better suited for people with cognitive disabilities and also for non-native speakers [9].

This article performs a practical evaluation of four different ATS methods in a dataset extracted from two major Brazilian Federal Courts (TRF5 and STF). The datasets comprise 200 summarized documents (100 for each court), addressing topics from different legal domains. The evaluated methods were: two unsupervised models, i.e., MUSS (English and Portuguese versions), and two supervised approaches, namely Transformers and NMT + Attention [10,11,12]. Readability metrics were employed in order to assess the quality of the produced, such as Flesch Reading Ease (FRE). Thus, this work intends to investigate the following research questions, **RQ1: How to increase the average readability score of simplified texts, in relation to the originals, on a set of summaries, so that they are simpler to read and understand? RQ2: How to generate a parallel corpus in Portuguese and in the legal environment to be used in the training of models that use supervised learning?**

2. Related Work

ATS is an active research topic, which many different approaches proposed in the legal domain. In [13], ATS was evaluated on South Korean legislation texts, while [1] used in his research the simplification of texts contained in contracts covering legal matters obtained through the LEDGAR dataset. Both works employed word substitution (lexical simplification methods). [14] performed rule based simplifications of documents available in the Senate and the Philippine Supreme Court. In [15] authors conducted research on simplifying legal texts obtained through the United States Supreme Court, through deep learning, e.g., the Legal-BERT model. Most of the previous work on ATS in the legal domain focuses on the English language. [3] demonstrated interesting results on the usage of Neural Machine Translation methods on ATS of Portuguese documents.

Another form of making legal documents simpler, is by the application of text summarization methods. [16] summarized 100,000 court rulings from the German court using the German language, while [17] proposed LegalSumm, a system to perform automatic summarization of court decisions in Portuguese. The authors made their dataset public, becoming an important benchmark for ML tasks in the legal domain. However, even though automatic text summarization and simplification may have the same objectives, they comprise different methods and evaluation metrics, since the objective of the first is to produce a brief summary of the main ideas of the text, while the latter is intended to reduce the linguistic complexity of the text while maintaining its original meaning [18].

3. Methods

This section covers the methods, which use models with algorithms specialized in ATS and NLP used in this research.

3.1. ATS based on Transformer architecture and attention mechanisms

The Transformer architecture, introduced by Vaswani et al. [19], has revolutionized natural language processing tasks, including text simplification. The key innovation of Transformers lies in their attention mechanisms, which allow the model to focus on different parts of the input sequence when making predictions. This attention mechanism enables the model to capture complex dependencies and relationships within the text, making it well-suited for tasks like text simplification. Several studies have leveraged the power of Transformers for text simplification tasks, achieving notable success [10,9]. By employing self-attention mechanisms, these models can effectively identify and rewrite complex structures or expressions to produce simplified, more accessible versions of the original text. In our experiments, we used the “bertimbaulaw-base-Portuguese-cased” pre-trained model, with a training on the parallel corpus of 13,288 pairs of sentences, for 10 epochs, with batch size 8.

3.2. ATS based on Multilingual Unsupervised Sentence Simplification by Mining Paraphrases (MUSS)

The Multilingual Unsupervised Sentence Simplification by Mining Paraphrases (MUSS) approach, proposed by Martin et al. [10], employs unsupervised learning to identify and generate simplified versions of sentences across multiple languages. By mining paraphrases from large corpora, MUSS provides a versatile framework for simplifying text without the need for parallel datasets or complex training procedures. To generate simplification and not just paraphrase sentences, MUSS allows to control attributes such as the compression rate based on sentence size, number of paraphrases to be used and linguistic complexity. The controllable form of the model allows the simplification process to be made more flexible through adjustments of parameters such as lexical complexity, syntactic and Levenshtein similarity. The ability to operate in an unsupervised manner also makes MUSS highly adaptable to different domains and applications, further enhancing its practical utility in real-world scenarios. In this work, we used the “muss_en_wikilarge_mined” and the “muss_pt_mined” pre-trained models. As the datasets are written in Portuguese, prior and after the use of the MUSS(EN) models, it was necessary to use two translation models, respectively: “Helsinki-NLP/opus-mt-Mul-en” and “Helsinki-NLP/opus-mt-Tc-big-en-pt” both available in Huggingface and used for the translation of sentences in Portuguese/English and English/Portuguese. To perform the simplification tests of the summaries, the environment used was a virtual machine (Virtual Box 6.1.38), with Ubuntu 22.04.1 64x, 16GB of RAM and Python 3.10.6.

3.3. ATS based on Neural Machine Translation (NMT) with attention mechanisms

The conversion of a complex sentence to a simpler one, can be modelled as a translation task, making Neural Machine Translation (NMT) [20] a good fit for the problem. NMT with attention mechanisms, selectively focus on different parts of the input sentence dur-

ing the translation process [12]. By doing so, the model can give appropriate emphasis to crucial elements, resulting in simplified sentences that preserve the original meaning. This attention-driven approach enhances the quality of the generated simplifications, making them more coherent and contextually accurate. Researchers have demonstrated the effectiveness of ATS based on attention mechanisms across various languages and domains, showcasing its versatility and potential for widespread applications [3]. The simplification model in this work was trained from scratch with a parallel corpus of 132,879 pairs of sentences, for 100 epochs and batch size 128.

4. Materials

Two datasets were created, one with 100 TRF5 summaries obtained from existing judgments in the query application “Julia — Easy Search”³ and another with 100 of the STF extracted from existing judgments in the dataset RulingBR [17]. Only sentences with a maximum length of 512 characters were used to reduce computational complexity. In addition, in order to evaluate methods that required a parallel corpus, three more datasets were created in this work. A sample of 10,574 STF summaries from the RulingBR dataset, were submitted to the pre-trained MUSS(Pt) model, resulting in 132,879 pairs of sentences. A subsample (10%) of these sentences was extracted, with 13,288 pairs. On average, TRF5 summaries have 424 tokens (words), while STF summaries have an average of 191 tokens. Datasets were preprocessed (sentence splitting) with Punkt⁴ [21], and a manually created dictionary of legal acronyms and abbreviations was employed.

5. Evaluation

The test dataset was extracted from the original data, totalling 200 sentences (100 for each, STF and TRF5 datasets). Readability was evaluated with the Flesch Reading Ease (FRE) method, given by: $FleschPT = 248.835 - (1.015 \times ASL) - (84.6 \times ASW)$, where, *ASL*: Average length of sentences and *ASW*: Average number of syllables per word. The scores obtained are in the range of 0 to 100, and the lower values indicate that the text is more difficult to read, while the higher values consider the texts more readable [22]. Besides, auxiliary metrics were also employed, i.e., “Number of existing sentences in the text”, “Average length of sentences” and “Average number of syllables per words”, as they help to understand the simplification process.

6. Results

The readability metrics of the evaluation are presented in Table 1. The original TRF5 texts presented 54.86 points in the FRE index, being considered relatively difficult to read, while the STF summaries were 45.09 points, classified as difficult to read. After simplifying the texts, the Transformers method obtained 64.71 points leaving the normal TRF5 summaries to read. The MUSS(EN) method increased the STF summaries score

³<https://julia.trf5.jus.br/julia/entrar>

⁴<https://www.nltk.org/modules/nltk/tokenize/punkt.html>

Dataset	Method	Number of Sentences	ASL	ASW	Readability
TRF5	Original	28,92 \pm 8,43	17,52 \pm 3,45	2,11 \pm 0,10	54,86 \pm 9,66
	MUSS(EN)	27,44 \pm 18,26	19,02 \pm 5,70	2,00 \pm 0,10	62,82 \pm 8,90
	MUSS(PT)	29,20 \pm 8,39	15,37 \pm 3,08	2,11 \pm 0,12	56,64 \pm 10,28
	TRANSFORMERS	30,19 \pm 8,27	13,16 \pm 2,75	2,05 \pm 0,12	64,71 \pm11,13
	NMT+ATTENTION	26,61 \pm 7,82	12,93 \pm 4,07	2,57 \pm 0,12	19,85 \pm 10,24
STF	Original	15,41 \pm 9,56	11,25 \pm 6,78	2,29 \pm 0,20	45,09 \pm 18,79
	MUSS(EN)	13,22 \pm 8,32	12,63 \pm 6,18	2,10 \pm 0,19	60,02 \pm17,21
	MUSS(PT)	15,32 \pm 9,15	10,28 \pm 5,49	2,30 \pm 0,22	45,23 \pm 19,82
	TRANSFORMERS	15,63 \pm 9,14	10,30 \pm 4,40	2,29 \pm 0,23	46,12 \pm 20,41
	NMT+ATTENTION	14,66 \pm 8,70	10,09 \pm 4,55	2,74 \pm 0,22	7,96 \pm 18,41

Table 1. Readability metrics on the test dataset

to 60.02 points, making them normal for reading, and although the Transformers used a reduced parallel corpus the fact of having used a pre-trained model adjusted with legal terms helped him in this process. Except for the NMT+Attention method that did not do well in the tests, all other methods improved the readability of the texts.

It can also be noted that the original STF summaries have an FRE readability index of 45.09 points, being classified as “Difficult” to read, while those of TRF5 are “Relatively Difficult” to read with 54.86 points, thus, one can infer that STF sentences are more difficult to simplify. It should also be noted that the training process of this method had problems and was not completed, even using a reduced parallel corpus, due to the limited resources of the computational environment used for training, this probably impacted the results of the supervised methods. On the other hand, the MUSS(EN) method proved to be very competitive, obtaining good scores both in the TRF5 and STF summaries, differently from what happened with the MUSS(PT) method. This is probably impacted by the translation process, since English translated texts tends to be more objective than in Portuguese [23]. The MUSS(PT) method was evaluated directly in Portuguese sentences, eliminating the translation step, which can be seen as a pre-simplification of the texts.

In addition to the quantitative analysis performed, a survey with 16 people was made to assess the quality of the simplifications. Overall, 69% of the respondents believe that the models were able to simplify the texts, with the majority of votes assigned to MUSS(PT) (36%), followed by MUSS(EN) 29%, Transformers (24%) and NMT+Attention (11%).

7. Conclusion

This work have evaluated different ATS methods in the task of legal text simplification. The experiments demonstrated that ATS with pre-trained models can improve readability of legal documents, achieving more compact sentences with shorter words. A survey with 16 people corroborated with the quantitative results, indicating the potencial of ATS to be incorporated in the legal domain. As future work, a parallel corpus specifically created for legal texts in Portuguese, adjusted and revised in quantity and quality of sentence pairs, can be used for training of task specific (supervised) model. Another important point to highlight as future work is the possibility to evaluate ATS based on Large Language Models (LLM), as Generative Pre-Training Transformer (GPT) and other models of this domain.

References

- [1] A. Garimella, A. Sancheti, V. Aggarwal, A. Ganesh, N. Chhaya and N. Kambhatla, Text Simplification for Legal Domain: Insights and Challenges (2022).
- [2] E. Motta, JUDICIAL SENTENCES AND SIMPLE LANGUAGE: a possible and necessary meeting (2022).
- [3] T. Lima, A. Nascimento, G. Valença, P. Miranda, R. Mello and T. Si, Portuguese Neural Text Simplification Using Machine Translation (2021).
- [4] S.M. Aluísio and C. Gasperin, Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts, in: *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, Association for Computational Linguistics, 2010, pp. 46–53.
- [5] J. Carroll, G. Minnen, Y. Canning, S. Devlin and J. Tait, Practical simplification of English newspaper text to assist aphasic readers, in: *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998, pp. 7–10.
- [6] R. Evans, C. Orasan and I. Dornescu, An evaluation of syntactic simplification rules for people with autism, in: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 2014, pp. 131–140.
- [7] G.H. Paetzold and L. Specia, Unsupervised lexical simplification for non-native speakers, in: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [8] L. Rello, R. Baeza-Yates, S. Bott and H. Saggion, Simplify or help?: text simplification strategies for people with dyslexia, in: *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, ACM, 2013, p. 15.
- [9] L. Martin, E. Clergerie, A. Bordes and B. Sagot, Controllable Sentence Simplification (2020).
- [10] L. Martin, A. Fan, E. Clergerie, A. Bordes and B. Sagot, MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases (2021).
- [11] C. Schröer, F. Kruse and J. Gómez, A Systematic Literature Review on Applying CRISP-DM Process Model (2021).
- [12] S. Yang, Y. Wang and X. Chu, A Survey of Deep Learning Techniques for Neural Machine Translation (2020).
- [13] A. Muralidharan, Automatic English Text Simplification for Statutes (2021).
- [14] M. Collantes, M. Hipe, J.L. Sorilla, L. Tolentino and B. Samson, Simpatico: a text simplification system for senate and house bills, in: *Proceedings of the 11th National Natural Language Processing Research Symposium*, 2015, pp. 26–32.
- [15] M. Cemri, T. Çukur and A. Koç, Unsupervised Simplification of Legal Texts (2022).
- [16] I. Glaser, S. Moser and F. Matthes, Summarization of German Court Rulings (2021).
- [17] D. Feijó, Summarizing Legal Rulings (2021).
- [18] S. Ma and X. Sun, A Semantic Relevance Based Neural Network for Text Summarization and Text Simplification (2017).
- [19] A. Vaswani, N. Shazzer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser and I. Polosukhin, Attention Is All You Need (2017).
- [20] M.T. Luong, H. Phan and C.D. Manning, Effective Approaches to Attention-based Neural Machine Translation, 2015.
- [21] R. López and T. Pardo, Experiments on Sentence Boundary Detection in User-Generated Web Content, 2015.
- [22] W. DuBay, *Smart Language: Readers, Readability, and the Grading of Text*, Impact Information, 2006.
- [23] V.C. Lígia, After all, is English easier or more difficult than Portuguese? (2017), Accessed on 15 June 2023.