RecognaSumm: A Novel Brazilian Summarization Dataset

Pedro Henrique Paiola[1], **Gabriel Lino Garcia**[1], **Danilo Samuel Jodas**[1], **João Vitor Mariano Correia**[1], **Luis Claudio Sugi Afonso**[1], **João Paulo Papa**[1]
[1]School of Sciences, São Paulo State University, Bauru, Brazil
{pedro.paiola, gabriel.lino, mariano.correia, luis.afonso, joao.papa}@unesp.br
danilo.jodas@gmail.com

Abstract

Research in the field of automatic summarization, particularly in abstractive summarization, for the Portuguese language still faces a significant challenge due to the limited availability of datasets with annotated summaries. Although existing datasets enable research, they are comparatively smaller than those available for the English language, thereby impeding the attainment of more robust results. This paper introduces RecognaSumm, a novel Portuguese dataset comprising a diverse set of journalistic texts annotated with summaries. With a total of 135,272 samples, it stands as the largest known Portuguese summarization dataset to date, to the best of our knowledge. Additionally, this work introduces an abstractive summarization model trained on this dataset[1], offering a baseline for future studies.

1 Introduction

The increasing availability of information in the digital age has generated an unprecedented demand for Natural Language Processing (NLP) systems capable of analyzing, comprehending, and summarizing large volumes of text. One of the most notable applications of this technology is automatic text summarization, which aims to extract the essential content from extensive documents in a concise and readable manner. Text summarization plays a pivotal role in various domains, including academic research, journalism, data analysis, and information retrieval.

Summarization methods can be classified in various ways. In particular, the most common classification is one that distinguishes between extractive methods, which seek the most important sentences from the original text to compose the summary, and abstractive methods, which, in contrast to the former, are capable of generating their own sentences to compose the summary (Nenkova et al., 2011).

In the Brazilian context, despite the growing interest in the field of NLP, there has been a limited availability of suitable databases for text summarization tasks. For instance, the TeMário (Pardo and Rino, 2003) and CSTNews (Cardoso et al., 2011) datasets are considered traditional resources in the domain of automatic summarization in Portuguese. However, when compared to datasets in English, they contain a significantly smaller number of samples. This deficiency has posed a challenge for researchers and developers aiming to create effective summarization models in the Portuguese language. To address this gap, this article introduces RecognaSumm[2], a novel and comprehensive database specifically designed for the task of automatic text summarization in Portuguese.

RecognaSumm stands out due to its diverse origin, composed of news collected from a variety of information sources, including agencies and online news portals. The database was constructed using web scraping techniques and careful curation, resulting in a rich and representative collection of documents covering various topics and journalistic styles. The creation of RecognaSumm aims to fill a significant void in Portuguese language summarization research, providing a training and evaluation foundation that can be used for the development and enhancement of automated summarization models.

In this article, we present in detail the methodology for constructing the RecognaSumm database, its features, and evaluation metrics. Furthermore, we demonstrate the practical utility of the dataset through the application of various summarization models, highlighting its potential in various natural language processing applications. The availability of RecognaSumm to the research and development community is a pivotal step in driving innovation in the field of automatic summarization in the Por-

[1]Model available at: https://huggingface.co/recogna-nlp/ptt5-base-summ

[2]Dataset available at: https://huggingface.co/datasets/recogna-nlp/recognasumm

tuguese language and facilitating significant advancements in this domain.

In summary, this work represents a significant milestone in the creation of essential resources to advance research in text summarization in Brazil and offers a valuable contribution to the academic community and industry interested in NLP and applications related to text content analysis in the Portuguese language.

The remainder of this paper is organized as follows: Section 2 provides a review of related works, and Section 3 introduces the proposed dataset. Sections 4 and 5 present the experimental setup and results of a toy-evaluation performed over the proposed dataset, respectively. Finally, Section 6 states conclusions.

2 Related Works

This section presents and describes the primary datasets concerning to summarization in Brazilian Portuguese.

- CSTNews (Leixo et al., 2008) (Cardoso et al., 2011): comprises 140 news articles, categorized into various subjects and sources, namely: Folha de São Paulo, Estadão, O Globo, Jornal do Brasil, and Gazeta do Povo;
- RulingBR (de Vargas Feijó and Moreira, 2018): developed for the summarization of legal texts in Portuguese, containing $10,623$ decisions from the Brazilian Supreme Federal Court;
- Temário (TExtos com suMÁRIOs) (Pardo and Rino, 2003): a dataset composed of 100 news articles, distinguished by having its summaries authored by a professional summarizer, in addition to a teacher and a journalism expert. This corpus has also been expanded from 100 to 251 news articles (Maziero et al., 2007);
- WikiLingua (Ladhak et al., 2020): encompasses 18 languages, comprising a total of $141,457$ articles extracted from the WikiHow website, of which $81,695$ are in Portuguese;
- XL-Sum (Hasan et al., 2021): covers 44 languages and contains a total of $301,444$ samples, with $71,752$ samples in the Portuguese language sourced from the British Broadcasting Corporation (BBC) news extractions.

3 RecognaSumm Dataset

RecognaSumm was designed for tailoring and leveraging research involving abstractive summarization in the context of the Portuguese language.

3.1 News source selection

The first step involves selecting the most respected and influential news agencies in Brazil for the news article collection and analysis. The selection process hinges on the prominence and influence of such news agencies in the Brazilian journalistic scenario, thus ensuring data diversity and trust according to the public's interest. Table 1 summarizes the Brazilian news agency adopted for the dataset design.

Table 1: News agencies adopted in the process of the RecognaSumm creation.

Agency	# of news
BBC	$6,902$
CNN	$29,709$
Extra	$8,128$
G1	$51,061$
iG	$7,068$
O GLOBO	$5,812$
Olhar Digital	$9,078$
UOL	$15,795$
Total	$135,272$

The news collection was performed using web crawlers specifically designed for each news agency website, thus allowing a customized and accurate data composition. Each web crawler was developed to track the info category related to each news agency website. This process ensures diversity and a broad range of reports, as well as an extensive collection of topics including but not limited to politics, technology, and sports (Table 2).

3.2 Data preprocessing and organization

RecognaSumm is structured to support a wide range of research involving text summarization. Each news article includes the components presented in Table 3.

After the news articles are extracted by a web crawler, we proceed with a pre-processing phase, which includes standardization of terms, removing words or elements that could introduce distortion to the news content, such as tags, "None" values, advertising text, URLs, and the like.

Table 2: Categories for RecognaSumm.

Category	# of news
Brazil	14,131
Economy	12,613
Entertainment	5,337
Health	24,921
Policy	29,909
Science and Technology	15,135
Sports	2,915
Travel and Gastronomy	2,893
World	27,418
Total	135,272

Table 3: Metadata used to describe each sample.

Information	Description
Title	Title of article
Sub-title	Brief description of news
News	Information about the article
Category	News grouped according to your information
Author	Publication author
Date	Publication date
URL	Article web address
Reference summary	Combined title and subtitle

Upon completion of this pre-processing stage, we commence the utilization of summarization techniques on the news articles.

3.3 Abstractive summarization

RecognaSumm is designed to produce concise and accurate summaries by only taking advantage of the title and subtitle of each news article. This process aims to yield more informative and condensed summaries while refraining from utilizing a random selection of the specific parts of the original text. To better evaluate the characteristics of the reference summaries of this dataset, we adopted the compression and abstraction ratios.

The compression ratio is computed from the balance between the number of tokens within the prospect summary and the number of tokens in the original news article text. A value close to 1 means the summary size is close to the one of the original text. Table 4 exhibits a text reduction around $\frac{1}{4}$ of the original texts according to the compression ratio computed for each set assembled from the whole RecognaSumm dataset.

Conversely, the abstraction ratio seeks to find the frequency of the n-grams appearances within the candidate summary yielded by the abstractive summarization. The abstraction ratio is computed according to the following equation:

Table 4: RecognaSumm compression ratio.

Split	Compression ratio
Training set	24.31%
Validation set	23.48%
Test set	24.16%
Test set (candidate summaries)	24.65%

$$Abs_n(C_n, S_n) = 1 - \frac{|C_n \cap S_n|}{|C_n|}, \qquad (1)$$

where C_n stands for the n-grams set within the candidate summary, while S_n is the n-grams set of the original text, being n the number of connected strings in a single n-gram. A higher $Abs_n(C_n, S_n)$ value indicates greater similarity between the candidate summary and the original content. Table 5 displays the abstraction ratios for n-grams of sizes n=[1, 2, 3].

Table 5: RecognaSumm abstraction rate.

n-gram	split	percentage
1-gram	Training set	24.00%
1-gram	Validation set	23.05%
1-gram	Test set	24.06%
2-gram	Training set	60.05%
2-gram	Validation set	60.13%
2-gram	Test set	60.18%
3-gram	Training set	73.89%
3-gram	Validation set	73.97%
3-gram	Test set	74.02%

3.4 Datasets comparison

The proposed dataset stands out for its size and the substantial number of samples compared to other datasets available for text summarization in the context of the Portuguese language. This aspect is essential for training a broad range of summarization models, thus enabling more effective performance when generating summaries in Portuguese. Table 6 shows the number of samples for each dataset proposed for the Portuguese language summarization.

4 Experimental Setup

With the dataset already created, we followed the same methodology used by PTT5-Summ (Paiola et al., 2022), conducting fine-tuning of the PTT5 model (Carmo et al., 2020) using the RecognaSumm data. The goal was to obtain preliminary results for this dataset, which would serve as a baseline for future research endeavors.

Table 6: Comparison of samples among the baseline datasets.

Datasets	# of samples
Summ-it	50
TeMário	251
CSTNews	140
RulingBR	10,623
XL-Sum	71,752
WikiLingua	81,695
RecognaSumm	135,272

The training code was implemented in Python language, using PyTorch and Transformers. For the optimization, the Adam algorithm (Kingma and Ba, 2014) was used, with learning rate of 3×10^{-5}. The models were trained on an NVidia T4 GPU, with 15GB of VRAM, for 2 epochs. A maximum of 512 tokens were considered as input and 150 as output. Beam search algorithm was used to generate the candidate, with $k = 5$ as beam width.

The candidate summaries produced by the models were evaluated using the set of ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) metrics (Lin, 2004). ROUGE metrics are content-based measures aimed at indicating how much of the reference summary is preserved in the generated summary, calculated by counting the number of overlaps of *n*-grams between the candidate summary and the reference summary.

5 Experimental Results

Table 7 presents the evaluation metrics for the candidate summaries generated by the PTT5 model after fine-tuning with RecognaSumm.

Table 7: Evaluation of PTT5 fine-tuned with RecognaSumm dataset, according to the measures ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL)

Dataset	R1	R2	RL
RecognaSumm (PTT5)	38.45	17.19	28.19

These results should be viewed as preliminary, serving as baselines for future experiments. However, it is worth noting that they do not deviate significantly from the metrics obtained in other datasets. In particular, when considering datasets in Portuguese, the model proposed by the authors of XL-Sum, for instance, achieves values of 37, 17, 15, 90, and 28, 56 for ROUGE-1, ROUGE-2, and ROUGE-L metrics, respectively, for Portuguese texts in this dataset.

Through empirical evaluation of the results, it is also possible to observe that the model has indeed acquired the ability to summarize texts, although it may still be subject to inherent text generation issues, such as hallucinations. Below is an example of a news article, its reference summary, and the summary generated by the trained model.

Source text: A FromSoftware anunciou que a expansão DLC estava oficialmente em desenvolvimento em uma publicação no Twitter, embora a empresa não tenha revelado para quando o lançamento pode ser aguardado. [...] Elden Ring foi lançado dia 25 de fevereiro do ano passado, e até agora o jogo só recebeu patches de balanceamento, e uma atualização que permite um melhor PvP nos coliseus do jogo. [...]

Reference summary: Expansão de Elden Ring está oficialmente em desenvolvimento. A FromSoftware confirmou que está desenvolvendo um DLC de Elden Ring, um dos games de maior sucesso de 2022.

Candidate summary: FromSoftware anuncia expansão DLC de Elden Ring. O jogo foi lançado em fevereiro do ano passado, e até agora o jogo só recebeu patches de balanceamento e uma atualização que permite um melhor PvP.

6 Conclusions

This work aimed to produce a new Portuguese dataset incorporating text summaries by tailoring the abstractive text summarization to ensure a large corpus prioritizing quality, representativity, and diversity of the news articles collected from different news agency sources. The news collection, data organization, and abstractive summary generation were conducted to offer a novel and comprehensive information source that seeks to capitalize on research on Portuguese text summarization. RecognaSumm aims to shed insights and enhance the knowledge in Portuguese summarization, thus promoting opportunities for innovative algorithms and research progress in specific aspects of the Portuguese language in terms of the wide range of language processing tasks.

Future research will be conducted to expand the number of samples in the RecognaSumm datasets. In addition, further experiments are expected to explore novel language models and fine-tuning approaches to handle the nuances of the Portuguese texts.

References

Paula CF Cardoso, Erick G Maziero, Mara Luca Castro Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago AS Pardo. 2011. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting at the 8th Brazilian Symposium in Information and Human Language Technology*, pages 88–105, Cuiabá, Mato Grosso. NILC.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *ArXiv*, abs/2008.09144.

Diego de Vargas Feijó and Viviane Pereira Moreira. 2018. RulingBR: A summarization dataset for legal texts. In *Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 255–264, Canela, Rio Grande do Sul. Springer.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Priscila Leixo, Thiago Alexandre Salgueiro Pardo, et al. 2008. CSTNews: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Espanha. Association for Computational Linguistics.

Erick Galani Maziero, VR Uzêda, Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes. 2007. Temário 2006: Estendendo o córpus temário.

Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

Pedro H. Paiola, Gustavo H. de Rosa, and João P. Papa. 2022. Deep learning-based abstractive summarization for brazilian portuguese texts. In *BRACIS 2022: Intelligent Systems*, pages 479–493, Cham. Springer International Publishing.

Thiago Alexandre Salgueiro Pardo and Lucia Helena Machado Rino. 2003. Temário: Um corpus para sumarização automática de textos.