PublicHearingBR: A Dataset of Public Hearing Transcripts for Summarization of Long Documents in Portuguese

Leandro Carísio Fernandes[1][0000−0002−4114−2334] and Guilherme Zeferino Rodrigues Dobins[2][0009−0002−4336−2143]

[1] Tribunal de Contas da União
[2] Unicamp

Abstract. This paper introduces PublicHearingBR, a new dataset designed for summarizing long documents in Portuguese. The dataset consists of transcriptions of public hearings held by the Brazilian Chamber of Deputies, which are paired with corresponding journalistic articles and associated metadata (who are the individuals and what they support or said). These resources aim to facilitate the development and evaluation of summarization models of long documents in Portuguese. The contributions of this work include the PublicHearingBR, a prompt for a hybrid summarization model to establish a baseline for future studies, and a discussion on evaluation metrics for summarization involving large language models, with a focus on addressing hallucination in the generated summaries.

Keywords: Long Documents Summarization · Dataset · Portuguese

1 Introduction

Text summarization is a task in natural language processing that condenses a long text into a shorter version while retaining its main ideas. This task is important for various applications, including information retrieval, content management, automatically writing news article from documents, and providing quick insights into lengthy documents, among others.

Summarization methods can be categorized into three types: abstractive, extractive, and hybrid [7]. Abstractive summarization generates new sentences that capture the essence of the original text. Extractive summarization selects and concatenates segments directly from the source text. Hybrid approaches combine elements of both methods.

The type of documents involved in summarization tasks can vary, ranging from short news articles to long research papers, and collections of multiple documents. Each type presents distinct challenges. Summarizing short documents (such as a news article) is easier to perform and evaluate. Summarizing multiple documents typically depends on a prior retrieval stage on a specific topic within those documents. Summarizing a long document usually involves relatively high

compression rates, which means it is necessary to assess what is the most relevant informations in the document.

The availability of datasets is fundamental to advancing research and development in summarization [4]. Datasets provide the necessary training and evaluation resources for developing models that can handle various summarization tasks. However, most available datasets are in English, creating a gap for other languages. This limitation hinders the development of summarization tools for non-English speakers.

Portuguese, one of the world's most spoken languages[3], exemplifies this issue. Despite its wide usage, there is a scarcity of summarization datasets in Portuguese. Existing datasets usually focus on short documents like news articles, leaving a gap in resources for long document summarization.

Addressing this gap, this paper introduces PublicHearingBR, an open dataset designed for the summarization of long documents in Portuguese. The dataset consists of transcriptions of public hearings held by the Brazilian Chamber of Deputies. These transcriptions are lengthy and cover diverse topics, providing a source of data for developing and testing summarization models.

The contributions of this paper are: (1) PublicHearingBR, the dataset for summarization of long documents in Portuguese; (2) a prompt for a hybrid summarization model, which provides a baseline for future studies; (3) a discussion on evaluation metrics for summarization involving large language models (LLMs), including a discussion on hallucination on the summary.

2 Related Work

One of the first summarization datasets available in Portuguese is the TeMário [9], which contains 100 pairs of journalistic articles extracted from two newspapers and summaries written by a specialist. On average, the texts contain approximately 600 words, and the summaries about 200 words. When writing the summary, the specialist kept the length of the summary as 25% to 30% of the original article. The dataset was later expanded with an additional 150 samples [6]. The new samples are slightly larger, with an average of 1 200 words per text.

Other datasets in Portuguese for news summarization are the CSTNews [1] and the RecognaSumm [8]. CSTNews has manual annotation, it contains 140 samples of short texts, and all the texts together sum to about 47 000 words, that is, just over 330 words per entry. On the other hand, RecognaSumm has about 130 000 entries and consider that the summary is the title and subtitle of the news.

News articles are, in general, short texts. A Portuguese dataset with slightly larger documents is the RulingBR [2]. This dataset contains more than 10 000 rulings from the Brazilian Supreme Federal Court. Each document contains four sections: *Ementa*, *Relatório*, *Voto*, and *Acórdão*. The *Ementa* section is considered the reference summary of the decision. In this database, the full content of

[3] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

the document contains an average of 1 200 words, reaching up to 2 000 words. The summary contains up to 150 words, with an average of about 90 words, making the summary about 7% of the total document size. This type of dataset focuses only on summarizing the document to find what the decision is.

There are also multilingual summarization datasets, notably WikiLingua [5] and XL-Sum [3]. The former deals with summarizing "How To" texts, and the latter with summarizing news articles. Both contain text pairs in English and summaries in one or more languages (up to 18 languages for WikiLingua and up to 44 languages for XL-Sum). Generally, the texts are short. For example, the average length of WikiLingua articles is 391 tokens, and the summaries are 39 tokens long.

There is a lack of datasets for summarizing long documents in Portuguese. Recently, the dataset Cem Mil Podcasts was made available for non-commercial research purposes. The database contains transcriptions and some metadata for about 114 000 podcasts in Brazilian Portuguese (pt-BR) and another 8 000 episodes in European Portuguese (pt-PT). On average, the transcriptions have about 9 500 words (with a median of approximately 6 700 words). The longest transcription contains about 205 000 words. The task it to propose a description for the episode using the transcript as input and compare it to the the one given by the creator of the podcast. To access the dataset, one must request permission from the maintainers. However, according to the official website[4], as of December 2023, due to shifting priorities, they no longer take requests to access it.

Despite the availability of summarization datasets in Portuguese, there is a lack of open datasets for summarizing long documents. This is a gap that the dataset PublicHearingBR aims to fill. The dataset includes 206 long documents and their summaries (metadata indicating the individuals and what they said or support, and a related nes article about the public hearing using the same metadata). The documents are transcriptions of public hearings of the Brazilian Chamber of Deputies on various topics. The task is to extract the relevant individuals and ther opinions, which can be used to write a news article about the event. The dataset is open and can be accessed freely without needing any authorization [5].

3 Dataset creation

3.1 Concept and purpose of the dataset

The committees of the Chamber of Deputies hold public hearings and debates with civil society entities on various topics of national interest. The complete transcripts of these hearings are made available by the Chamber of Deputies as public documents, freely accessible on their official website[6]. These transcripts

[4] https://podcastsdataset.byspotify.com/

[5] https://github.com/carisio/IA024_deep_learning_nlp/tree/main/projeto%20final/--%20arquivos%20finais%20--/o%20dataset

[6] https://www.camara.leg.br

are extensive documents, containing dozens and often more than a hundred pages, reflecting the complexity and richness of the discussions held.

The Agência Câmara de Notícias[7] frequently publishes news articles about these public hearings. These articles are written by journalists who follow the daily activities of the Chamber of Deputies and who, after listening to the hearings, synthesize the discussions into journalistic articles. These journalists consider not only the content of the hearings but also the social and political context at the time they occurred to create summaries that highlight the most relevant points of the discussions.

The article can be seen as a summarization of a long document, the transcript. Thus, we can use the article (a short document) as a proxy for the transcript. From the article, we can extract a set of metadata indicating the main subject, the individuals involved in the debates, and their opinions.

By associating all this information (available in Portuguese), we can create a dataset containing tuples with (a) the transcript (long document); (b) an article (summary: short document); (c) and a set of metadata indicating the main subject, the individuals and their opinions (a structed summary). Because the article was written by a specialized team, we can consider its text and the metadata extracted from it as ground truth (gold standard).

This dataset can be used for the following tasks:

1. Summarizing long documents into topics: The transcript can be used to extract the main individuals of the public hearing and their opinions. The extracted data can be compared with the metadata in the dataset.
2. Writing a journalistic article from a set of metadata: In this case, the metadata can be used as input for a pipeline to write an article, which should be compared with the article available in the dataset.
3. Writing a journalistic article from a long document: The transcript can be used in a pipeline to write a journalistic article, whose result can be compared with the article available in the dataset.

In this context, this dataset aims to use the articles as examples of summarization of long documents, the transcripts, allowing the development and testing of automatic summarization algorithms in Portuguese. The proposed dataset has been named PublicHearingBR.

3.2 Step-by-step process for creating PublicHearingBR

To create the dataset, the process began with a manual selection of URLs of journalistic articles written by Agência Câmara de Notícias that were related to public hearings. During this initial step, we selected 206 articles about public hearings that had occurred in recent years. Then, a manual search was conducted on the websites of the commissions that held the hearings to find the URLs of the corresponding transcripts. This ensured that each article had a matching transcript.

[7] https://www.camara.leg.br/noticias

The next step involved an automated download of HTML files for both the selected articles and the corresponding transcripts, which were parsed to extract the text contents. This involved cleaning the HTML structure and isolating the main text (videos and figures of the articles were removed).

Using the extracted text from the articles as inputs, we used GPT-4o with the prompt shown in Fig. 1 to extract the metadata of the articles. The metadata included key information such as the main subject of the hearing, the individuals involved, and their opinions.

To ensure the accuracy and reliability of the generated metadata, we manually compare all of them with the news articles and correct them.

[System]
Você é um assistente que analisa matérias escritas pela Agência Câmara, da Câmara dos Deputados. Seu papel é identificar na matéria os seguintes itens:

- Tópico principal que está sendo tratado
- O nome das pessoas envolvidas
- O que cada pessoa defende
- O que cada pessoa disse (em caso de existir citação direta)

Desconsidere o nome dos jornalistas ou editores da matéria. As únicas pessoas que interessam são as que estão no corpo da matéria.

O retorno deve ser no formato JSON, com duas propriedades:

– "assunto": uma string que indica o assunto principal da matéria
– "envolvidos": uma lista de objetos que indica as pessoas envolvidas na matéria. O objeto deve ter três propriedades:
  – "nome": string, indica nome da pessoa
  – "cargo": string, indica cargo que a pessoa ocupa, juntamente com o órgão, a entidade ou a empresa em que ela trabalha, se estiver disponível
  – "opinioes": lista de string indicando todas as opiniões que a pessoa defendeu e que estão indicadas no texto. As opiniões devem ser listadas de forma detalhada. Se for uma citação direta, o texto indicado na lista DEVE OBRIGATORIAMENTE ser idêntico ao contido na matéria, incluindo as aspas.

[User]
{ARTICLE_TEXT}

Fig. 1. Prompt used to extract the metadata of the news articles used in PublicHearingBR.

3.3 Characteristics of PublicHearingBR

PublicHearingBR contains 206 entries. Each entry includes a public hearing transcription, a journalistic article, and a set of metadata that associates opinions with individuals as presented in the article. Table 1 and Fig. 2 show the statistics of the word counts[8] in the transcripts and articles, and the number of individuals and opinions in the metadata.

Table 1. Statistics of PublicHearingBR.

Statistic	Transcript (words)	Article (words)	Relative size of article compared to the transcription (%)	Individuals (count)	Opinions (count)
Mean	18 102	627	4.20	5.2	10.7
Std	12 137	160	2.06	1.8	4.5
Min	4 437	288	0.75	2	3
Q (5%)	8 495	399	1.76	3	5
Q (25%)	12 079	522	2.75	4	8
Median	16 424	607	3,97	5	10
Q (75%)	20 858	706	5.24	6	13
Q (95%)	30 761	932	7.21	8.75	18.75
Max	147 728	12 15	15.07	12	31

On average, the transcripts have about 18 000 words. Ninety percent of them contain between approximately 8 500 and 30 500 words. There are two transcripts with over 50 000 words, and the longest transcript available in the dataset has nearly 150 000 words.

The articles are more uniform in length, averaging around 600 words, with the longest article being approximately twice this length. The low variability in length is characteristic of journalistic articles.

Articles are typically about 4% of the length of their corresponding transcript, which translates to a compression rate of about 96%. Due to the relatively consistent length of articles and varying transcription lengths, an article can be up to 15% of the length of a transcription.

On average, each article in the dataset presents the opinions of approximately five individuals, and there are about ten opinions per article. The article with the most opinions includes 31 opinions.

[8] This article considers words as the result of a simple split of a text.

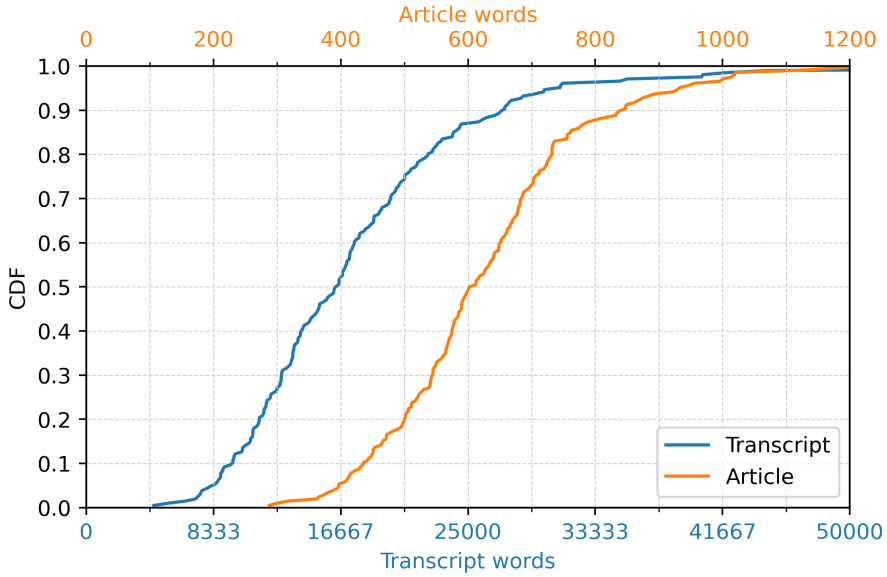

Fig. 2. Cumulative distribution function (CDF) of the words in the transcripts and articles of PublicHearingBR.

4 Experiment - Summarization of public hearing transcripts using ChatGPT

4.1 Description of the experiment

This experiment considers the primary use case for the dataset: extracting a set of relevant opinions from individuals based on a public hearing transcript.

As shown in section 3.3, the transcripts in this dataset are extensive text. Directly inserting the text into a LLM can, depending on the model, exceed its context window limit. Due to this limitation, we opted to test ChatGPT instead of GPT. ChatGPT is more capable of handling larger inputs, making it a better choice for working with these contents. This also simulates the perspective of an end user of the tool.

We created a Custom GPT with a specialized instruction prompt, as shown in Fig. 3. This prompt provides detailed instructions on how the transcripts should be read and interpreted. It shows the format of the data and instructs the language model to extract the transcript from an input file.

4.2 Results and discussion

In this experiment, ChatGPT is used to generate a metadata structure similar to what the dataset provides, which includes individuals and their opinions. The primary goal of this section is to evaluate three aspects: (a) recall - to determine the percentage of relevant opinions that the experiment was able to return; (b) precision - to check the percentage of returned opinions that are relevant; (c) hallucination - to ensure that all returned opinions are present in the transcript, even if they are not in the metadata.

To calculate recall and precision, a three-stage pipeline was used:

- The first stage involves mapping the individuals mentioned in the transcript to those in the metadata available in PublicHearingBR. Since names might differ (due to abbreviations, for example), a language model is initially asked to perform this mapping. This was done using GPT-4o with the prompt shown in Figure 4.
- The second stage involves mapping, for each identified individual, the opinions extracted from the transcript to those contained in PublicHearingBR. This is necessary because, as these are natural language texts, it is important to analyze whether the opinions are semantically the same. This mapping is also performed using GPT-4o with the prompt shown in Figure 5.
- Once the mapping is complete, it is possible to verify the number of opinions from the dataset that were returned and, thereby, calculate recall and precision.

In this type of problem, the most relevant metric is recall, as it aims to determine if the opinions returned by human evaluators are also returned by the summarization system. A low precision is not necessarily a bad outcome. When writing an article, a journalist may choose to omit other relevant opinions due to space constraints, for example. This is a summarization problem where the goal is to produce a relevant news article among many possible ones.

Despite this, it is important to evaluate whether the prompt or tested system is hallucinating in conjunction with precision. A low precision associated with hallucinations undermines the credibility of the system. On the other hand, a low precision without hallucinations is not harmful: as long as the system has sufficiently summarized the transcript, it provides the user with a broad range of opinions to write a journalistic article.

The issue of hallucination is critical, because the objective is writing a news article and in this situation false information is not acceptable. Another issue is that, as it is a debate, it is essential to ensure that the opinions expressed are correctly attributed to the right person. To check for hallucination, we propose a pipeline with the following steps:

- Separate each person's speeches.
- For each person, divide the speeches into small chunks.
- For each opinion returned in the summarization, search the related chunks of that person.

– Send the opinion returned in the summarization and the chunks found in the previous step to an LLM to check if the generated information can be inferred from the chunks.

Table 2 shows the results using the proposed methodology. The average recall was 44.9%, indicating that the custom GPT used recovered almost half of the relevant opinions indicated in the dataset. On the other hand, the average precision was 25%, indicating that for each relevant opinion, the custom GPT also brought three other opinions. On average, 8% of the returned opinions are hallucinations. In this type of system, it is desirable that this number be 0.

In general, hallucinations occurred when opinions from one person were attributed to another. In a purely hybrid summarization system, this can be resolved by summarizing in parts, where the system only summarizes the speeches of a single person.

The data in the table also show a situation where there was 100% hallucination. This occurred due to an error in the mapping done in the hallucination calculation. The transcription refers to an event where the participants used Libras (Brazilian Sign Language), and therefore, the transcription also indicated the interpreter, which was not considered by the automatic evaluation system.

Table 2. Results of the Experiment Using ChatGPT.

Statistic	Recall (%)	Precision (%)	Hallucination (%)
Mean	44.9	24.8	8.2
Std	22.5	13.1	10.3
Min	0.0	0.0	0.0
Q (5%)	11.5	5.5	0.0
Q (25%)	28.6	15.2	0.0
Median	44.4	23.9	5.5
Q (75%)	60.0	33.0	12.5
Q (95%)	87.1	47.5	22.2
Max	100.0	71.4	100.0

O usuário irá enviar uma transcrição de uma audiência pública realizada em alguma Comissão da Câmara dos Deputados. Seu papel é ler o arquivo que será enviado e identificar os seguintes itens:

– Tópico principal da audiência pública
– O nome das pessoas envolvidas
– O que cada pessoa defende ou comenta

Os itens identificados deve possibilitar que o usuário redija uma matéria jornalística com início, meio e fim.

A transcrição contém a fala exata dita pelos participantes. Inicialmente o participante é identificado. Todo o texto que se segue até a identificação de uma nova pessoa é a fala daquele participante. O texto possui o seguinte formato:

```
[[O(A) SR.(SRA.) PESSOA 1]]
[[UM OU MAIS PARÁGRAFOS CONTENDO TODA A FALA DA PESSOA 1]]

[[O(A) SR.(SRA.) PESSOA 2]]
[[UM OU MAIS PARÁGRAFOS CONTENDO TODA A FALA DA PESSOA 2]]

...
...

[[O(A) SR.(SRA.) PESSOA N]]
[[PARÁGRAFOS CONTENDO TODA A FALA DA PESSOA N]]
```

Após ler e analisar o documento enviado, você deverá dar a sua resposta no formato JSON com três propriedades:

– "assunto": uma string que indica o assunto principal da audiência pública. Essa informação normalmente está na primeira fala do primeiro participante, logo na abertura da audiência.
– "envolvidos": uma lista de objetos que indica as pessoas envolvidas no debate. O objeto deve ter três propriedades:
  – "nome": string, indica nome da pessoa
  – "cargo": string, indica cargo que a pessoa ocupa, juntamente com o órgão, a entidade ou a empresa em que ela trabalha, se estiver disponível
  – "opinioes": lista de string indicando TODAS as opiniões relevantes ao assunto que a pessoa defendeu e que estão indicadas no texto. As opiniões devem ser listadas de forma detalhada
– "tl_dr": um resumo que possibilitará ao usuário escrever uma matéria jornalista sobre a audiência pública usando os dados (nome e opiniões) dos envolvidos extraídos da transcrição

Fig. 3. Instruct prompt of the custom GPT used to extract the metadata of the transcripts in PublicHearingBR.

[System]
Você receberá duas listas de pessoas, com seus nomes e cargos.
Uma mesma pessoa pode aparecer nas duas listas, mas de maneira diferente (pequenas alterações no nome, variação do cargo...)
Seu objetivo é criar uma correspondência entre as duas listas, no formato de um dicionário, em que as chaves são o nome da pessoa na lista 1, e os valores são os nomes correspondentes da lista 2. Anote apenas os nomes, e não os cargos. E se não houver uma correspondência, escreva 'Não identificado'.
Sua resposta final deve ser apenas o json, não escreva nada além disso.

Exemplo
Lista 1:
Nome: Carlos Eduardo Souza. Cargo: Engenheiro
Nome: Roberta da Silva Borges. Cargo: Jornalista investigativa
Nome: Maurício Dornelles. Cargo: Assessor de imprensa

Lista 2:
Nome: Pablo Mariano. Cargo: Psicólogo
Nome: Carlos Eduardo de Souza. Cargo: Engenheiro de Produção
Nome: Roberta Borges. Cargo: Jornalista

Resposta:
```
{
    "Carlos Eduardo Souza": "Carlos Eduardo de Souza",
    "Roberta da Silva Borges": "Roberta Borges",
    "Maurício Dornelles": "Não identificado"
}
```

Fig. 4. Prompt to map individuals in two lists.

[System]
Você receberá duas listas contendo assuntos tratados em uma audiência pública na Câmara dos Deputados.
Um mesmo assunto pode aparecer nas duas listas, mas escrito de formas diferentes. Seu objetivo é avaliar o conteúdo dos assuntos e fazer um mapeamento entre cada item da primeira lista com itens da segunda, indicando quais são similares.
Para serem considerados similares, os assuntos devem necessariamente representar uma mesma informação, mas escrita de diferentes formas.
Os assuntos de cada lista serão numerados, e você deve criar um dicionário com o mapeamento dos assuntos de cada lista.
Um mesmo assunto da segunda lista pode ser mapeado para mais de um assunto da lista 1, e vice-versa.
Os assuntos sem um correspondente devem ser marcados com "Não identificado".
Sua resposta deve ser apenas o json, não escreva nada além disso.

Exemplo:
Lista 1:
1. Acusou o deputado Marcos de publicar mentiras em redes sociais, e defendeu que o mesmo seja punido por isso.
2. Destacou a importância de haver verificações de veracidade de publicações em todas as redes sociais.
3. Defendeu o bloqueio de contas que publiquem informações falsas.

Lista 2:
1. Citou o Projeto de Lei 2630/20, conhecido como PL das Fake News.
2. Fez uma acusação contra o deputado Marcos, por publicar fake news em seu facebook.
3. Insinuou que o deputado Marcos seja punido.
4. Insinuou que contas que publiquem fake news devem ser derrubadas das redes sociais.
5. Citou o nome de envolvidos no Projeto de Lei.

Resposta:
```
{       "1": "2, 3",
      "2": "Não identificado",
      "3": "4"
}
```

Fig. 5. Prompt to map opinions.

5 Conclusion

We introduced PublicHearingBR, a new dataset for the summarization of long documents (transcripts of public hearings) in Portuguese. It also presented, as a baseline, a custom GPT to summarize these transcripts, extracting the main individuals of the hearing and their opinions. We also discussed evaluation metrics for this task.

There is a lack of datasets for this task in Portuguese, and the existing datasets for summarizing large documents do not include a summary of the main discussions carried out in the document.

In the summarization task for extracting people's opinions, the suggested custom GPT returned, on average, almost 45% of the opinions indicated in the dataset (recall). Additionally, the average precision was approximately 25%, indicating that it brought three times more opinions that were not in the reference dataset. On average, about 8% of the generated opinions were hallucinations, which mainly occurred due to opinions being attributed to the wrong people.

In addition to this task, the dataset also enables two other tasks: generating a news article from a set of metadata and generating a news article from a public hearing transcription. Future work should focus on creating baselines for these two tasks of the dataset.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to check the grammar and semantics of the human written text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Paula Christina Figueira Cardoso, Erick Galani Maziero, Mara Elena Lucia, R. Castro Jorge, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças, Volpe Nunes, Thiago Alexandre Salgueiro Pardo, and Rodovia Washington Luís. CST-News - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. 2011.
2. Diego de Vargas Feijó and Viviane Pereira Moreira. RulingBR: A Summarization Dataset for Legal Texts. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold, editors, *Computational Processing of the Portuguese Language*, pages 255–264, Cham, 2018. Springer International Publishing.
3. Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei

Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August 2021. Association for Computational Linguistics.

4. Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Computing Surveys*, 55(8):1–35, December 2022.
5. Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November 2020. Association for Computational Linguistics.
6. Erick Galani Maziero, Vinícius Rodrigues de Uzêda, Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes. TeMário 2006: Estendendo o Córpus TeMário. Technical report, Universidade de São Paulo, Universidade Federal de São Carlos, Universidade Estadual Paulista, August 2007. NILC-TR-07-06.
7. Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.
8. Pedro Henrique Paiola, Gabriel Lino Garcia, Danilo Samuel Jodas, João Vitor Mariano Correia, Luis Afonso Sugi, and João Paulo Papa. RecognaSumm: A novel Brazilian summarization dataset. In Pablo Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira, and Raquel Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 575–579, Santiago de Compostela, Galicia/Spain, March 2024. Association for Computational Lingustics.
9. Thiago Alexandre Salgueiro Pardo and Lucia Helena Machado Rino. TeMário: Um Corpus para Sumarização Automática de Textos. Technical report, Universidade de São Paulo, Universidade Federal de São Carlos, Universidade Estadual Paulista, October 2003. NILC-TR-03-09.