**UNIVERSIDADE ESTADUAL PAULISTA**
**"JÚLIO DE MESQUITA FILHO"**
Câmpus de Bauru

Pedro Henrique Paiola

# Sumarização Abstrativa de Textos em Português Utilizando Aprendizado de Máquina

Bauru
2022

Pedro Henrique Paiola

# Sumarização Abstrativa de Textos em Português Utilizando Aprendizado de Máquina

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de Bauru, como requisito parcial para à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. João Paulo Papa

Bauru
2022

P148s

Paiola, Pedro Henrique
Sumarização Abstrativa de Textos em Português Utilizando Aprendizado de Máquina / Pedro Henrique Paiola. -- Bauru, 2022
104 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru
Orientador: João Paulo Papa

1. Processamento de Linguagem Natural. 2. Aprendizado de Máquina. 3. Sumarização. 4. Sumarização Abstrativa. 5. Português Brasileiro. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

**ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE PEDRO HENRIQUE PAIOLA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS.**

Aos 09 dias do mês de setembro do ano de 2022, às 14:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de PEDRO HENRIQUE PAIOLA, intitulada **Sumarização Abstrativa de Textos em Português Utilizando Aprendizado de Máquina**. A Comissão Examinadora foi constituida pelos seguintes membros: Prof. Dr. JOAO PAULO PAPA (Orientador(a) - Participação Virtual) do(a) FC / UNESP/Bauru (SP), Profa. Dra. HELENA DE MEDEIROS CASELI (Participação Virtual) do(a) Departamento de Computação / Universidade Federal de São Carlos, Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Participação Virtual) do(a) FC / UNESP/Bauru (SP). Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final:_ _ _APROVADO_ _ . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

João Paulo Papa

Prof. Dr. JOAO PAULO PAPA

*Dedico este trabalho à Deus, que me deu forças para seguir esta jornada, e a minha família, que me acompanhou durante toda a caminhada.*

# Agradecimentos

Agradeço primeiramente a minha família, especialmente aos meus pais, que foram fundamentais na minha formação e na minha dedicação aos estudos; e à minha namorada e companheira, que me apoiou e me deu forças para concluir este projeto;

Agradeço a todos os professores que me instruíram ao longo da minha vida, pois eles me forneceram a bagagem que me permitiu chegar aqui.

Agradeço, em especial, ao meu orientador, pela oportunidade de realizar este trabalho e por me guiar durante toda sua realização.

Por fim, agradeço aos meus colegas de pós-graduação. Por diversas vezes compartilhamos experiências e nos ajudamos mutuamente a prosseguir nesta jornada.

*"O perigo real não é de computadores começarem a pensar como homens, mas de homens começarem a pensar como computadores."*

(Sydney J. Harris)

# Resumo

A sumarização automática consiste no processo de capturar as informações mais relevantes de um texto e condensá-las em um texto compreensível em linguagem natural. Este processo pode ser classificado como sumarização extrativa, quando identifica as sentenças mais importantes do texto de origem para compor o sumário utilizando as mesmas sentenças, ou sumarização abstrativa, quando gera novas sentenças baseadas nas informações mais relevantes do texto de origem. Pesquisas em sumarização automática abstrativa para o português brasileiro ainda são escassas, especialmente para sumarização abstrativa baseada em aprendizado em profundidade. Por este motivo, este consiste no foco desta pesquisa. Nesta dissertação são apresentados experimentos com modelos pré-treinados, ajustados para as bases TeMário, CSTNews e para os textos em português da WikiLingua e XL-Sum. Os resultados apresentados por estes experimentos são relativamente satisfatórios, ainda apresentando problemas, dos quais a maioria são comuns em sumarização abstrativa, mas que podem servir como ponto de partida para futuras pesquisas.

**Palavras-chave:** Processamento de Linguagem Natural. Aprendizado de Máquina. Sumarização. Sumarização Abstrativa. Português Brasileiro.

# Abstract

Automatic summarization captures the most relevant information in a text and condenses it into an understandable text in natural language. This process can be classified as extractive summarization, which identifies the most important sentences from the source text and composes the summary using that very same sentences, or abstractive summarization, which generates new sentences based on the most relevant information from the source text. Research on Brazilian Portuguese-based abstractive summarization is still scarce, especially for deep learning-based abstractive summarization. For this reason, this is the focus of this research. This master thesis presents experiments with pre-trained models, fine-tuned for the TeMário and CSTNews databases and for the texts in Portuguese from WikiLingua and XL-Sum. The results presented by these experiments are relatively satisfactory, still presenting problems, most of which are common in abstractive summarization, but can serve as a starting point for future research.

**Keywords:** Natural Language Processing. Machine Learning. Summarization. Abstractive Summarization. Brazilian Portuguese.

# Lista de ilustrações


Figura 1 – Um exemplo de sentença e sua AMR correspondente. Elaborado pelo autor. 25
Figura 2 – Arquitetura base da PEGASUS. Adaptado de (ZHANG et al., 2020) . . . . 28
Figura 3 – Procedimento para o cálculo da similaridade entre duas sentenças através do BERTScore. Extraído de (ZHANG et al., 2019) . . . . . 34
Figura 4 – Comparação do MoverScore com o BERTScore em relação a correspondência de palavras. Extraído de (ZHAO et al., 2019) . . . . . 36
Figura 5 – Fluxograma proposto para a sumarização abstrativa em português a partir de arquiteturas baseadas em inglês. . . . . 43
Figura 6 – Fluxograma proposto para a sumarização extrativa em português a partir de arquiteturas baseadas em inglês. . . . . 43
Figura 7 – Fluxograma proposto para o ajuste-fino monolíngue do modelo PTT5. . . . 44
Figura 8 – Fluxograma proposto para o ajuste-fino multilíngue do modelo mT5. . . . . 45
Figura 9 – Fluxograma proposto para o ajuste-fino entre línguas do modelo T5. . . . . 46
Figura 10 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos treinados na base TeMário. . . . . 57
Figura 11 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos treinados na base CSTNews. . . . . 59
Figura 12 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos mT5 treinados na base TeMário. . . . . 68
Figura 13 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos mT5 treinados na base CSTNews. . . . . 70

# Lista de tabelas


Tabela 1 – Resultados da comparação entre métodos de sumarização extrativa multi-documento. Adaptado de (SODRÉ; OLIVEIRA, 2018) . . . . . 39
Tabela 2 – Quantidade de amostras nos conjuntos de treinamento, validação e teste. . 41
Tabela 3 – Quantidade de palavras nas bases de dados avaliadas. . . . . 42
Tabela 4 – Resultados dos experimentos com sumarizadores extrativos aplicados à base TeMário, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS), MoverScore (MS) e grau de compressão (Comp) . . . . . 48
Tabela 5 – Resultados dos experimentos com sumarizadores extrativos aplicados à base CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS), MoverScore (MS) e grau de compressão (Comp) . . . . . 48
Tabela 6 – Resultados dos experimentos com sumarizadores abstrativos aplicados à base TeMário, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 49
Tabela 7 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos sumários abstrativos gerados para a base TeMário. . . . . 49
Tabela 8 – Resultados dos experimentos com sumarizadores abstrativos aplicados à base CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2) e ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 50
Tabela 9 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos sumários abstrativos gerados para a base CSTNews. . . . . 50
Tabela 10 – Resultados do ajuste-fino do modelo T5 (Base) aplicados as bases TeMário e CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 51
Tabela 11 – Avaliação dos modelos treinados com as bases WikiLingua e XL-Sum, conforme as medidas ROUGE-1 (R1), ROUGE-2 (R2) e ROUGE-L (RL) . 52
Tabela 12 – Resultados dos modelos treinados na base TeMário, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 55
Tabela 13 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos treinados na base TeMário. . . . . 56

Tabela 14 – Resultados dos modelos treinados na base CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 58
Tabela 15 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos treinados na base CSTNews . . . . . 58
Tabela 16 – Resultados do modelo mT5 treinado na base TeMário, avaliado com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 67
Tabela 17 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos multilíngues treinados na base TeMário . . . . . 68
Tabela 18 – Resultados do modelo mT5 treinado na base CSTNews, avaliado com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS) . . . . . 69
Tabela 19 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos multilíngues treinados na base CSTNews . . . . . 70
Tabela 20 – Avaliações dos sumários produzidos para a base TeMário pelos diferentes modelos experimentados neste trabalho. . . . . 75
Tabela 21 – Comparação das avaliações dos sumários produzidos pelos sistemas em inglês e os sistemas treinados na base CSTNews. . . . . 75

# Lista de abreviaturas e siglas

| | |
|---|---|
| AMR | *Abstract Meaning Representation* |
| BERT | *Bidirectional Encoder Representations from Transformers* |
| CNN | *Convolutional Neural Networks* |
| LSTM | *Long Short-Term Memory* |
| PLN | Processamento de Linguagem Natural |
| ROUGE | *Recall-Oriented Understudy for Gisting Evaluation* |
| RNN | *Recurrent Neural Networks* |
| RST | *Rethorical Structure Theory* |
| SCU | *Summarization Content Units* |
| TF-IDF | *Term frequency-inverse document frequency* |
| TF-ISF | *Term frequency-inverse sentence frequency* |
| WMD | *Word Mover's Distance* |

# Sumário


**1 INTRODUÇÃO . . . . . 16**
**1.1 Perguntas de Pesquisa . . . . . 17**
**1.2 Hipótese de Pesquisa . . . . . 18**
**1.3 Organização do Trabalho . . . . . 18**

**2 REVISÃO BIBLIOGRÁFICA . . . . . 19**
**2.1 Sumarização Extrativa . . . . . 21**
**2.2 Sumarização Abstrativa . . . . . 23**
2.2.1 Abordagens baseadas em estrutura . . . . . 23
2.2.2 Abordagens baseadas em semântica . . . . . 25
2.2.3 Abordagens baseadas em aprendizado em profundidade . . . . . 26
2.2.3.1 PEGASUS . . . . . 28
2.2.3.2 GSum . . . . . 29
2.2.3.3 Controle de cópia . . . . . 29
**2.3 Sumarização entre línguas e multilíngue . . . . . 30**
**2.4 Medidas de avaliação . . . . . 31**
**2.5 Sumarização de textos em português . . . . . 37**
2.5.1 Bases de dados . . . . . 37
2.5.2 Principais trabalhos . . . . . 38

**3 METODOLOGIA . . . . . 41**
**3.1 Bases de dados . . . . . 41**
**3.2 Abordagem proposta . . . . . 42**
3.2.1 Experimentos com sistemas treinados com bases em inglês . . . . . 42
3.2.2 Treinamento de modelos com bases em português . . . . . 44
3.2.2.1 Treinamento monolíngue . . . . . 44
3.2.2.2 Treinamento multilíngue . . . . . 45
3.2.2.3 Treinamento entre línguas . . . . . 46
3.2.3 Avaliação dos sumários candidatos . . . . . 46

**4 RESULTADOS . . . . . 48**
**4.1 Experimentos com sistemas treinados em inglês . . . . . 48**
4.1.1 Treinamento entre línguas . . . . . 51
**4.2 Modelos treinados com bases em português . . . . . 52**
4.2.1 WikiLingua e XL-Sum . . . . . 52
4.2.1.1 Exemplo de sumarização na base WikiLingua . . . . . 53

4.2.1.2 Exemplo de sumarização na base XL-Sum . . . 54
4.2.2 TeMário e CSTNews . . . 55
4.2.2.1 Treinamento monolíngue . . . 55
4.2.2.1.1 Exemplo de sumarização na base TeMário . . . 60
4.2.2.1.2 Exemplo de sumarização na base CSTNews . . . 64
4.2.2.2 Treinamento multilíngue . . . 67
4.2.2.2.1 Exemplo de sumarização na base TeMário . . . 71
4.2.2.2.2 Exemplo de sumarização na base CSTNews . . . 72
**4.3 Comparação dos experimentos . . . 74**

**5 CONCLUSÕES E TRABALHOS FUTUROS . . . 77**

**REFERÊNCIAS . . . 79**

**APÊNDICES 88**

**APÊNDICE A – SUMÁRIOS GERADOS A PARTIR DE UM TEXTO DO CORPUS TEMÁRIO . . . 89**
**A.1 Texto fonte . . . 89**
**A.2 Sumário de referência . . . 90**
**A.3 Sumários candidatos . . . 91**
A.3.1 GistSumm . . . 91
A.3.2 (MILLER, 2019) . . . 91
A.3.3 (LIU; LAPATA, 2019) . . . 91
A.3.4 (SONG et al., 2020) . . . 91
A.3.5 PEGASUS (Large) . . . 92
A.3.6 PEGASUS (MultiNews) . . . 92
A.3.7 PEGASUS (Newsroom) . . . 92
A.3.8 PEGASUS (XSum) . . . 92
A.3.9 T5 (Base) . . . 93
A.3.10 T5 (Large) . . . 93
A.3.11 Fine-tuned T5 (Base) . . . 93
A.3.12 PTT5-A . . . 93
A.3.13 PTT5-B . . . 94
A.3.14 PTT5-C (WikiLingua) . . . 94
A.3.15 PTT5-D (XL-Sum) . . . 94
A.3.16 mT5-A . . . 95
A.3.17 mT5-B (WikiLingua) . . . 95
A.3.18 mT5-C (WikiLingua) . . . 95

A.3.19 mT5-D (XL-Sum) . . . 96
A.3.20 mT5-E (XL-Sum) . . . 96

**APÊNDICE B – SUMÁRIOS GERADOS A PARTIR DE UM TEXTO DO CORPUS CSTNEWS . . . 97**
**B.1 Texto fonte . . . 97**
**B.2 Sumário de referência . . . 98**
**B.3 Sumários candidatos . . . 98**
B.3.1 GistSumm . . . 98
B.3.2 (MILLER, 2019) . . . 98
B.3.3 (LIU; LAPATA, 2019) . . . 99
B.3.4 (SONG et al., 2020) . . . 99
B.3.5 PEGASUS (Large) . . . 99
B.3.6 PEGASUS (MultiNews) . . . 99
B.3.7 PEGASUS (Newsroom) . . . 100
B.3.8 PEGASUS (XSum) . . . 100
B.3.9 T5 (Base) . . . 100
B.3.10 T5 (Large) . . . 100
B.3.11 Fine-tuned T5 (Base) . . . 100
B.3.12 PTT5-E . . . 101
B.3.13 PTT5-F . . . 101
B.3.14 PTT5-G (WikiLingua) . . . 101
B.3.15 PTT5-H (XL-Sum) . . . 101
B.3.16 mT5-F . . . 102
B.3.17 mT5-G (WikiLingua) . . . 102
B.3.18 mT5-H (WikiLingua) . . . 102
B.3.19 mT5-I (XL-Sum) . . . 102
B.3.20 mT5-J (XL-Sum) . . . 103

# 1 Introdução

A sumarização de textos é um processo bastante comum e particularmente útil no cotidiano das pessoas, permitindo economizar tempo ao se buscar uma determinada informação. Um sumário pode apenas indicar qual o assunto do texto-fonte, de forma que o leitor pode decidir se é do seu interesse ler o texto completo ou ainda pode servir como um substituto do mesmo, trazendo todas as informações principais.

Atualmente, a sumarização se torna cada vez mais útil e necessária com o grande crescimento na quantidade de informações no ambiente *Web*, visando reduzir o esforço e tempo necessários para se encontrar informações relevantes considerando uma determinada consulta. Em particular, cresce o interesse na pesquisa por métodos de sumarização automática dentro da área de Processamento de Linguagem Natural (PLN), uma vez que, devido a este crescimento dos dados, se torna inviável a sumarização humana de todo este conteúdo.

A sumarização automática não é uma tarefa simples, visto que os métodos devem ser, de alguma forma, capazes de capturar as informações de um texto, perceber quais são as informações mais relevantes e por fim condensá-las em um texto compreensível em linguagem natural. Outro fator que complica esta tarefa é o fato de não se possuir uma resposta única. Considerando a sumarização humana, por exemplo, duas pessoas podem produzir sumários escritos de forma completamente diferente para um mesmo texto.

Métodos de sumarização podem ser classificados de diferentes maneiras. Em especial, a classificação mais comum é a que separa os métodos extrativos, que buscam as sentenças mais importantes do texto original para compor o sumário, dos métodos abstrativos, que diferentemente dos anteriores são capazes de gerar suas próprias sentenças para compor o sumário. Evidentemente, métodos de sumarização abstrativa são mais complexos e constituem uma tarefa mais árdua, porém podem atingir resultados mais próximos ao nível humano.

Os métodos de sumarização também apresentam diferentes abordagens, sendo uma das mais populares a baseada em aprendizado de máquina, em especial, em aprendizado em profundidade, devido aos resultados que estas abordagens vêm obtendo em diversas aplicações em PLN. Particularmente, pode-se destacar o modelo *Encoder-Decoder*, muito útil em aplicações de PLN, inclusive para sumarização. Além disso, com a ascensão de modelos pré-treinados como o BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2018), as pesquisam conseguem obter resultados cada vez melhores, mesmo em cenários em que os dados para treinamento são escassos. Ainda assim, estes métodos estão longe do nível humano, enfrentando diversos desafios.

O foco deste trabalho é a aplicação de métodos de sumarização automática abstrativa em textos na língua portuguesa. A grande maioria dos trabalhos voltados para a língua

portuguesa se encontram em sumarização extrativa multidocumento. Já para sumarização abstrativa os trabalhos ainda são escassos, principalmente aqueles baseados em aprendizado em profundidade. Uma possível justificativa para isso é que as bases de dados de sumarização mais tradicionais, na língua portuguesa, apresentam poucas amostras, o que dificulta o aprendizado dos modelos.

Neste sentido, realizou-se uma análise aprofundada desta área como um todo, em especial dos métodos de sumarização abstrativa baseados em aprendizado em profundidade, destacando o estado da arte, quais são os modelos mais utilizados e que atingem melhores resultados, avaliando-se a possibilidade de aplicar estes modelos para textos em português. Em especial, os experimentos foram realizados com as bases TeMário e CSTNews, duas das bases mais tradicionais nas pesquisas de sumarização em português, mas também foram realizados experimentos utilizando as bases WikiLingua e XL-Sum, que consistem em corpora multilíngues e incluindo uma quantidade de amostras de textos em português significativamente maior do que as duas bases anteriores. Dentre os experimentos realizados, buscou-se utilizar os resultados nestas últimas bases para impulsionar o desempenho dos modelos na TeMário e CSTNews, como será tratado melhor ao decorrer desta dissertação.

A partir da revisão bibliográfica e considerando os recursos e tempo disponíveis para a realização deste projeto, foram realizados os experimentos de sumarização abstrativa em português. A primeira linha de experimentos se deu a partir da utilização de sistemas de sumarização treinados com sistemas em inglês, aplicados às bases de dados em português com auxílio de um sistema de tradução. O principal objetivo destes experimentos foi obter alguma base de comparação para futuros resultados. Por fim, os últimos experimentos consistiram no treinamentos de modelos com as bases em português. Seus resultados foram analisados sob as devidas medidas de avaliação e comparados com os resultados anteriores.

## 1.1 Perguntas de Pesquisa

Nesta seção são apresentadas as perguntas de pesquisa que pretende-se considerar nesta dissertação de mestrado. Como não foram encontrados outros trabalhos de sumarização abstrativa em português para as bases TeMário e CSTNews, inicialmente foram realizados experimentos utilizando sistemas treinados em inglês, como citado anteriormente, buscando obter uma base de comparação para a avaliação das seguintes questões.

1. O treinamento de modelos de sumarização abstrativa baseados em aprendizado em profundidade, a partir do uso de *transformers* e modelos pré-treinados, utilizando as bases de dados disponíveis em língua portuguesa obtém resultados melhores que os obtidos com sistemas treinados em inglês?

2. O treinamento de modelos de sumarização abstrativa em bases de dados em português

com um maior número de amostras anotadas, seguida por ajuste-fino em bases com poucos recursos melhora o desempenho do modelo para estas bases?

3. O ajuste-fino de modelos de sumarização em inglês para as bases disponíveis em língua portuguesa, traduzidas para o inglês, melhora o desempenho do modelo?
4. O treinamento de modelos de sumarização abstrativa em bases multilíngue, utilizando textos de outras línguas além do português, melhora o desempenho do modelo?
5. O ajuste-fino do modelos treinados em bases multilíngue para bases com textos apenas em português melhora o desempenho do modelo para estas bases?

## 1.2 Hipótese de Pesquisa

A partir do estudo realizado e das perguntas de pesquisa, apresentadas anteriormente, a hipótese principal desta pesquisa é de que é possível utilizar modelos de aprendizado em profundidade, principalmente considerando a utilização de *transformers* e modelos pré-treinados, para produzir sumários abstrativos minimamente satisfatórios em português, mesmo em cenários com poucos recursos, em termos de quantidade de dados.

## 1.3 Organização do Trabalho

Este trabalho está dividido em mais 4 capítulos, além desta introdução. O Capítulo 2 apresenta uma revisão bibliográfica sobre o tema, tratando sobre diferentes classificações e abordagens de sumarização. Na revisão bibliográfica o problema da sumarização automática é tratado de forma geral, incluindo diversos trabalhos voltados para a língua inglesa, por exemplo, mas também destaca-se o cenário da sumarização automática para a língua portuguesa.

A seguir, o Capítulo 3 descreve a metodologia aplicada nesta pesquisa, apresentando as bases de dados utilizadas e as abordagens empregadas. No Capítulo 4 os resultados dos experimentos são apresentados e discutidos, comparando os diferentes modelos entre si, através das medidas de avaliação obtidas para cada um. E por fim, no Capítulo 5 são apresentadas as considerações finais desta dissertação.

# 2 Revisão Bibliográfica

Sumários podem ser classificados de diversas formas. Em especial, ressalta-se a classificação pela forma como o sumário é produzido: por sumarização extrativa ou sumarização abstrativa (NENKOVA; MCKEOWN, 2011), como citado na introdução. A sumarização extrativa gera resumos através da seleção e da extração das sentenças consideradas mais importantes do texto. Essas sentenças não sofrem qualquer tipo de alteração em relação ao material original. Já na sumarização abstrativa, novas sentenças serão escritas para criação de um sumário que transmita as principais informações do texto de entrada. Estas sentenças serão escritas a partir do uso de frases ou cláusulas do documento original. Este modelo de sumarização pode obter resultados melhores, porém necessita de métodos que sejam capazes de analisar o texto mais profundamente, que consigam representar conhecimento e que sejam capazes de gerar textos em linguagem natural de forma satisfatória (MOHAN et al., 2016; MOIRANGTHEM; LEE, 2020).

Um sumário também pode ser classificado com relação ao número de documentos utilizados no processo. Ele pode ser um sumário monodocumento (*single-document*), quando ele é gerado a partir de um único texto; ou multidocumento (*multi-document*), quando diversos documentos são utilizados na sumarização. Métodos multidocumento emergiram com os avanços na pesquisa, tendo como grande motivação o contexto da *Web*, onde pode-se encontrar um grande montante de páginas relacionadas a um mesmo tópico ou evento, que poderiam ter seus conteúdos sumarizados em um único texto (NENKOVA; MCKEOWN, 2011).

Considerando o conteúdo de um sumário, este pode ser classificado como indicativo, quando seu foco é dizer sobre o que o texto se trata, introduzindo seu assunto, ou ainda pode ser classificado como informativo, fornecendo as principais informações do texto, sendo considerado como um substituto do documento original.

Um sumário ainda pode ser classificado como genérico, quando ele não é destinado para um público específico. Sendo assim, todo o conteúdo do texto é considerado para a sumarização. Uma sumarização genérica também não realiza nenhum tipo de suposição sobre o domínio dos conteúdos que precisam ser sumarizados. Por outro lado, tem-se a sumarização focada em perguntas (*query-focused*), em que só são sumarizadas as informações relevantes para o usuário, baseado em suas consultas.

De forma geral, um sistema de sumarização é aquele que, a partir de um ou mais documentos de entrada, produz um sumário fluente e conciso das informações mais importantes da entrada. E para isso, o sistema requer uma capacidade de identificar, reorganizar, modificar e fundir informações expressas em diferentes sentenças. Devido a dificuldade deste processo, as abordagens mais tradicionais lidavam exclusivamente com a tarefa de identificar as sentenças

mais importantes do texto, constituindo métodos de sumarização extrativa (NENKOVA; MCKEOWN, 2011).

O primeiro trabalho de sumarização automática foi apresentado na década de 1950, por Luhn (1958). A ideia básica de Luhn, e que também seria adotada e refinada em diversos trabalhos futuros, era identificar as palavras mais significativas em um documento, e considerar que as sentenças mais importantes do texto são aquelas que apresentam mais destas palavras significativas próximas uma das outras. Neste método, a significância de uma palavra é determinada a partir da frequência com que ela ocorre no texto. Porém, o próprio Luhn aponta que algumas das palavras mais comuns em um documento podem não ser muito significativas para o conteúdo do texto. Artigos, preposições e pronomes, por exemplo, podem ser muito frequentes em um documento, mas não ajudam a descrever seu conteúdo. Para resolver isto, o método utiliza uma lista pré-definida com muitos destes termos, aos quais ele chama de *stop words*. Outro problema são as palavras que são muito comuns em um determinado domínio, mas que não ajudam a indicar o principal tópico do texto. A palavra "computador", por exemplo, pode aparecer com bastante frequência em um artigo sobre computação, mas sem explicitar claramente qual o assunto tratado no texto. Por este motivo, o método de Luhn estabelece limiares empíricos de frequência superior e inferior. Sendo assim, as palavras consideradas mais significativas são aquelas que possuem frequência no intervalo destes limiares.

O método de Luhn apresenta diversos problemas que podem ser facilmente percebidos. Por exemplo, um mesmo conceito ou entidade pode ser referido diversas vezes ao longo do texto utilizando palavras diferentes, através de sinônimos e pronomes. Palavras com significados distintos também podem indicar um tópico específico quando aparecem juntas, por exemplo, "furacão", "dano" e "vítimas" apontam para a descrição de um desastre natural. O método de Luhn não é sensível a estas situações. Uma mesma palavra também pode aparecer com diferentes variações morfológicas, como no caso de verbos apresentados em diferentes formas conjugais. Este último caso o próprio Luhn tenta evitar, através da unificação de palavras que são similares exceto pelos últimos seis caracteres (NENKOVA; MCKEOWN, 2011).

Outros trabalhos nesta área surgiram durante a década de 1960, com destaque para o método apresentado por Edmundson (1969), que buscava determinar quais as sentenças mais importantes considerando diversos atributos, tanto ao nível da palavra como da sentença. Os atributos considerados eram: número de ocorrências de uma determinada palavra, o número de palavras na sentença que também aparecem no título, a posição da sentença no texto, e a quantidade de expressões pertencentes a uma lista pré-compiladas de expressões-chaves como "em suma".

Durante as próximas duas décadas houve poucos progressos nessa área, sendo retomada a partir dos anos 1990, e se beneficiando dos avanços na área de PLN como um todo, especialmente nos últimos anos, permitindo inclusive a ascensão de métodos de sumarização abstrativa (MORENO, 2014). A seguir, são apresentadas com mais detalhes as principais

abordagens utilizadas atualmente para realizar sumarizações extrativas e abstrativas, assim como os trabalhos mais recentes que constituem o atual estado da arte nesta área.

## 2.1 Sumarização Extrativa

O processo de sumarização extrativa consiste, basicamente, em selecionar e concatenar os fragmentos, como sentenças ou parágrafos, mais representativos do texto original. Sendo assim, a peça chave para métodos deste tipo consiste em como determinar quais são os fragmentos mais relevantes de um documento.

Na prática, a estimativa de importância de uma sentença em um texto está fortemente baseada em atributos estatísticos e linguísticos. Normalmente são considerados os atributos ao nível da palavra e ao nível da sentença. Os atributos mais utilizados são os seguintes (MORATANCH; CHITRAKALA, 2017):

Atributos ao nível da palavra:

- Palavras-chave: partindo das ideias iniciais de Luhn (1958), encontrar as palavras mais significativas do texto ajudam a determinar quais as sentenças mais importantes. Sentenças que contenham mais palavras-chaves tendem a ser consideradas mais relevantes. No geral, as palavras-chaves são determinadas a partir da frequência com que ela ocorrem no texto, em especial costuma-se utilizar o valor TF-IDF (*Term frequency-inverse document frequency*) (JONES, 1972; JONES, 2004), que é calculado a partir da frequência da palavra no texto e pelo inverso da frequência nos documentos do corpus.
- Palavras do título: sentenças que contenham termos presentes no título possuem uma grande chance de contribuir positivamente para o sumário final, pois estes termos tendem a indicar o tema central do texto.
- Expressões-chave: algumas expressões servem de pistas sobre a estrutura do documento e auxiliam a determinar sentenças importantes. Expressões como "em suma", "porque" e "esta informação", por exemplo, costumam ser incluídas no sumário final.
- Palavras de tendência: são palavras normalmente relacionadas a um domínio específico. Elas podem ser consideradas importantes pois ajudam a descrever o tema do documento.
- Palavras em caixa alta: palavras escritas em caixa alta costumam ser palavras importantes, podendo se referir a entidades centrais no conteúdo do texto.

Atributos ao nível da sentença:

- Localização da sentença: documentos normalmente possuem uma estrutura hierárquica em que as sentenças mais importantes se concentram no início e no final dos parágrafos.

- Comprimento da sentença: o comprimento das sentenças ajudam a identificar quais as sentenças mais relevantes. Frases curtas normalmente não acrescentam muita informação, e sentenças muito longas também devem ser evitadas em um sumário. No geral, é considerado o comprimento normalizado, a partir do cálculo da razão do número de palavras na sentença pelo número de palavras da sentença de maior comprimento.
- Localização do parágrafo: de forma similar a localização da sentença, a localização do parágrafo no qual a sentença pertence ajuda a indicar sua relevância. Normalmente as sentenças mais relevantes se encontram nos parágrafos do início e do fim do documento.
- Coesão entre sentenças: este é um atributo importante porém mais difícil de obter, em comparação aos outros. A ideia é calcular o nível de coesão entre cada par de sentenças, de forma a determinar se uma sentença pode proceder a outra em um sumário e ainda manter a coesão do texto.

As abordagens adotadas para sumarização extrativa são tradicionalmente divididas em supervisionadas e não-supervisionadas. Os métodos não-supervisionados não precisam de sumários humanos para decidir quais são os atributos importantes em um documento. Por isso, estes métodos vão precisar de um algoritmo mais sofisticado, que compense a falta deste conhecimento. As abordagens típicas destes métodos são as baseadas em grafo, lógica nebulosa, conceitos ou análise de semântica latente. Por outro lado, os métodos supervisionados necessitam de uma grande quantidade de dados rotulados, que neste caso consiste em possuir sumários construídos manualmente para um conjunto de textos. Em especial, destacam-se as abordagens baseadas na regra de Bayes, redes neurais e campos aleatórios condicionais.

O primeiro trabalho a utilizar aprendizado de máquina para sumarização foi apresentado por Kupiec, Pedersen e Chen (1995). Utilizando diversos atributos, tanto ao nível da palavra quanto da sentença, inspirado por estudos anteriores, os autores acreditavam que a análise estatística de um corpus de sumarização, com documentos anotados com sumários de referência, revelaria quais atributos são mais relevantes para determinar a importância de cada sentença, e como ponderar estes atributos para calcular esta importância. O modelo de aprendizado utilizado foi um classificador *Naive Bayes*, a partir de um corpus com 188 documentos.

Mais recentemente, técnicas baseados em aprendizado em profundidade vêm se tornando cada vez mais populares e atingindo melhores resultados. O primeiro trabalho utilizando redes neurais foi em 2014, com a aplicação de um *unfolding recursive auto-encoder* (RAE) na sumarização extrativa de documentos (KÅGEBÄCK et al., 2014), atingindo o estado da arte na base de dados Opnosis (GANESAN; ZHAI; HAN, 2010) na época.

Como exemplos mais recentes de métodos baseados em redes neurais, destacam-se os recentes trabalhos baseados no modelo BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2018). Em (LIU, 2019), os autores propõem uma variante deste modelo, a qual batizam de BERTSUM, para permitir a realização da sumarização extrativa

de um documento. Em (LIU; LAPATA, 2019) este modelo é aplicado não apenas para a sumarização extrativa, mas também para a abstrativa. Os resultados obtidos por este trabalho atingiram o estado da arte para a base de dados CNN/Dailyman.

Mais recentemente, esses resultados foram impulsionados em (ZHONG et al., 2020), que também se baseia no BERT, criando uma variação chamada Siamese-BERT, porém com o foco de garantir um emparelhamento semântico entre o texto original e o sumário, ou seja, garantir que o sumário gerado preserve o significado do documento. Este objetivo é atingido justamente por adaptar o BERT em uma rede siamesa. Esta rede é treinada de forma a minimizar a distância dos *embeddings* dos documentos e os sumários associados, e maximizar a distância aos sumários referentes a outros documentos.

## 2.2 Sumarização Abstrativa

Como sumarizadores extrativos não são capazes de gerar novas sentenças, eles possuem certas limitações implícitas. Por este motivo, a maioria destes sumarizadores enfrenta problemas de coesão, legibilidade e de outros fatores de qualidade. A sumarização abstrativa, capaz de gerar novas sentenças a partir do texto original, pode se mostrar como um caminho para vencer estas barreiras. Porém, em contrapartida, se mostra uma tarefa muito mais árdua e complexa.

Enquanto a sumarização extrativa consiste basicamente em identificar e selecionar as sentenças mais relevantes de um certo documento, a sumarização abstrativa pode ser dividida em três sub-tarefas: extração de informação, seleção de conteúdo e geração de sentenças (LIN; NG, 2019).

Diferentes abordagens podem ser empregadas para realizar cada uma das três sub-tarefas do *pipeline* de sumarização abstrativa. A seguir, são apresentadas as abordagens mais comuns atualmente.

### 2.2.1 Abordagens baseadas em estrutura

Os métodos que seguem uma abordagem baseada em estrutura buscam as informações mais importantes em um documento e as representam por meio de estruturas pré-definidas, como grafos, árvores ou ontologias, para criar sumários abstrativos. Podemos subdividir os métodos desta abordagem em métodos baseados em: árvores, *templates*, ontologias, *lead and body*, grafos e regras (GUPTA; GUPTA, 2019).

Nos métodos baseados em árvores, primeiro são extraídos os trechos considerados mais importantes do texto original. Então, sentenças similares são identificadas a partir de um analisador e posteriormente são utilizadas para preencher estruturas de dados do tipo árvore. As árvores geradas são então linearizadas, sendo convertidas em texto, para a obtenção do sumário. No geral, utilizam-se árvores de dependências para a representação do texto. Em

(KURISINKEL; ZHANG; VARMA, 2017), por exemplo, é apresentado um método de abstração multi-documento baseado em árvores de dependência parciais, recombinação e em linearização sintática baseada em transições. A utilização de sistemas de geração de linguagem natural auxilia na criação de sumários gramaticalmente corretos, mas os métodos desse tipo ainda falham em considerar o contexto ao determinar quais frases devem ser incluídas no sumário. O foco maior é na sintaxe, e não na semântica (GUPTA; GUPTA, 2019).

Por outro lado, os métodos baseados em *templates* extraem fragmentos do texto original a partir do uso de palavras-chaves ou pistas, os quais são utilizados para compor os *templates* que formarão o sumário final. Como sua estrutura é pré-definida, ela permite criar sumários concisos e coerentes, funcionando bem quando o texto é estruturado de alguma maneira. Mas, como as regras e padrões dos *templates* devem ser definidos manualmente, métodos desta categoria consomem bastante tempo e requerem muito esforço manual.

Existem, também, os métodos baseados em ontologias. Ontologias podem ser definidas como um conjunto de entidades e as relações entre elas. Classes são os elementos principais em uma ontologia, sendo responsáveis por representar conceitos. Um conjunto de ontologias constitui uma base de conhecimento. Considerando o contexto do PLN e da sumarização automática, ontologias são normalmente utilizadas para extrair conceitos e relações de um texto, e a partir destas informações pode-se criar um sumário. Ainda que estes métodos sejam caracterizados como baseados em estrutura, eles auxiliam inclusive na representação semântica de um texto, especialmente quando há problemas de ambiguidade, com palavras que podem assumir diferentes significados (GUPTA; GUPTA, 2019; MOHAN et al., 2016).

Dentre os métodos baseados em grafos, normalmente se utiliza a estrutura *Word Graph* para representar os textos. A partir destes grafos, é realizada uma fusão de sentenças similares para a criação de um sumário, visando a remoção de redundâncias. Porém, é necessário assumir que existem diversas sentenças similares no texto e, ainda assim, não é fácil determinar quais são essas sentenças similares (GUPTA; GUPTA, 2019; BANERJEE; MITRA; SUGIYAMA, 2015).

Por fim, os métodos baseados em regras são voltados para sumarizações baseadas em perguntas. Inicialmente, é introduzido um conjunto de regras e categorias no sistema, pelas quais o texto do documento será categorizado visando encontrar as informações mais relevantes para a sumarização. Em seguida, questões são formuladas de acordo com o domínio do texto e as respostas são extraídas do documento, de modo a ajudar a compor o sumário final. Desta forma, como os baseados em *templates*, estes métodos demandam um grande trabalho manual devido à necessidade da especificação das regras e categorias a serem inseridas no sistema (GUPTA; GUPTA, 2019).

### 2.2.2 Abordagens baseadas em semântica

Abordagens baseadas em semântica consistem em obter uma representação semântica do texto e, a partir disso, gerar um sumário. A seguir, são apresentados alguns tipos de métodos que visam capturar a semântica de um texto.

Os chamados métodos baseados em sujeito-predicado buscam obter uma representação semântica da sentença a partir da estrutura sintática da mesma, através da identificação dos verbos, sujeitos, objetos e outros possíveis componentes. A partir disso, busca-se encontrar sentenças semelhantes que possam ser fundidas para a geração do sumário. Uma forma de encontrar sentenças semelhantes é utilizando algoritmos de agrupamento, como o *K-means* (GUPTA; GUPTA, 2019; ALSHAINA; JOHN; NATH, 2017).

Há também os métodos baseados em grafos semânticos. Diferentemente dos métodos baseados em grafos apresentados no item anterior, que visam capturar a estrutura sintática de um texto, o foco aqui é na estrutura semântica, onde o grafo representará as relações semânticas entre as sentenças do texto. Dentro das relações semânticas, também estão inclusas relações ontológicas e relações sintáticas entre as palavras (GUPTA; GUPTA, 2019).

Um exemplo de grafo semântico bastante utilizado nesses métodos é o grafo AMR (*Abstract Meaning Representation*) (BANARESCU et al., 2013), em que os vértices representam entidades e as arestas as relações entre elas. Na Figura 1 é apresentado um exemplo de um grafo AMR construído para uma determinada sentença. Nesta figura, percebe-se que o verbo "querer", por exemplo, possui dois argumentos, representados pelas arestas $ARG0$ (quem quer?) e $ARG1$ (o que quer?). É importante ressaltar o foco da AMR é na estrutura semântica, e não sintática, de modo que o verbo "desejar" na frase original não precisa ser mantido na AMR, desde que o significado da sentença se mantenha o mesmo. Dessa forma, uma determinada sentença pode admitir diversas AMRs, assim como uma certa AMR pode ser convertida em diferentes sentenças.

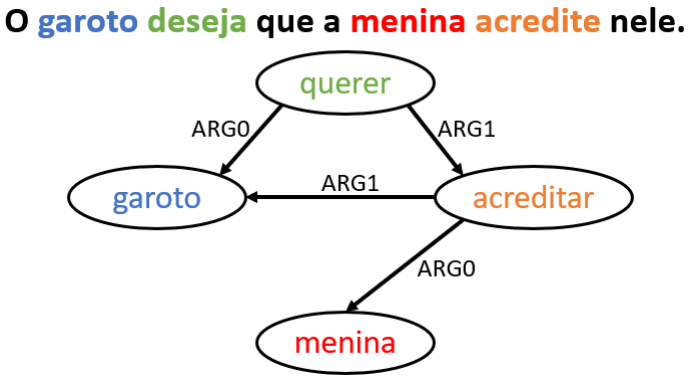


Figura 1 – Um exemplo de sentença e sua AMR correspondente. Elaborado pelo autor.

Em (DOHARE; GUPTA; KARNICK, 2018), é apresentado um sistema de sumarização abstrativa baseado em grafos AMR. A abordagem dos autores se inspira no modo com que os

humanos resumem um texto. Sendo assim, tendo gerado uma AMR para o texto de entrada, são identificadas as entidades e eventos mais importantes no texto e, após isso, as relações chaves entre essas entidades mais importantes. Posteriormente, são capturadas informações de contexto para as relações selecionadas. Estas entidades e relações chaves compõem um subgrafo da AMR do texto, e pode ser interpretado como uma sumarização do texto original, e do qual se obtém o texto sumarizado.

Os resultados ainda se encontram longe do nível humano, mas como os próprios Dohare, Gupta e Karnick (2018) sugerem, melhorias nos passos do *pipeline* de sumarização devem beneficiar diretamente os resultados. E após a publicação do artigo, houve avanços nesta área, como o uso do modelo BERT no processo *text-to-AMR* (CAI; LAM, 2020), que obtiveram resultados melhores do que os métodos utilizados neste trabalho.

### 2.2.3 Abordagens baseadas em aprendizado em profundidade

Atualmente, uma das abordagens mais populares é a baseada em aprendizado em profundidade, devido ao desempenho obtido em diversas aplicações de PLN. Esta abordagem permitiu a popularização dos métodos de sumarização abstrativa, atingindo resultados cada vez mais satisfatórios.

Diferente dos métodos clássicos, que precisavam lidar diretamente com as sub-tarefas de extração de informação, seleção de conteúdo e geração de sentenças, os métodos de aprendizado em profundidade constituem uma abordagem de ponta-a-ponta, aprendendo diretamente como abstrair as principais informações do documento de entrada e gerar o sumário correspondente. Uma desvantagem em relação aos métodos clássicos é que se possui muito menos controle neste processo, em especial, é difícil identificar exatamente o que o modelo aprendeu e como ele está extraindo e codificando as informações do texto (LIN; NG, 2019).

A maioria dos métodos desta abordagem utilizam o modelo *Encoder-Decoder*, que é especialmente útil quando tanto a entrada quanto a saída são sequências de palavras, problema conhecido como *Seq2Seq*. Basicamente, o *encoder* e o *decoder* são redes neurais, sendo a primeira responsável por converter as palavras em representações vetoriais, que auxiliam também a capturar o contexto em que ela se insere, e o *decoder* prevê a próxima palavra a ser inserida no sumário, com base nas anteriores (GUPTA; GUPTA, 2019). O primeiro trabalho de sumarização abstrativa a utilizar esta abordagem foi de Rush, Chopra e Weston (2015), baseado em mecanismos de atenção, o qual foi batizado de *Attention-Based Summarization* (ABS).

Neste modelo *Encoder-Decoder*, o *encoder* tem um objetivo similar à tarefa de extração de informação dos métodos clássicos, buscando capturar e codificar as informações relevantes do texto para a geração do sumário. O *encoder* deve receber representações vetoriais das palavras do texto, chamadas de *embeddings*. Os *embeddings* podem ser obtidos por modelos

pré-treinados em grandes bases de dados através do *word2vec* (MIKOLOV et al., 2013), GloVe (PENNINGTON; SOCHER; MANNING, 2014) ou o BERT (DEVLIN et al., 2018), por exemplo, ou também podem ser aprendidos durante o treinamento em si. Por fim, a seleção da arquitetura de rede neural a ser utilizada é um passo crucial, visando obter a melhor representação abstrata possível do texto e controlar quais informações serão enviadas para o *decoder*. Em especial, ainda é um desafio lidar com textos longos (LIN; NG, 2019). Algumas das arquiteturas normalmente aplicadas são as Redes Neurais Convolucionais (CNN - *Convolutional Neural Network*), Redes Neurais Recorrentes (RNN - *Recurrent Neural Networks*) e as *Long Short-Term Memory* (LSTM).

O *decoder*, por sua vez, é comumente implementado usando uma RNN. A cada passo da execução, o *decoder* recebe dois vetores de entrada: a representação do texto de entrada gerada pelo *encoder* e uma representação das palavras que já foram geradas nos passos anteriores. A partir disso é produzido um vetor correspondente ao tamanho do vocabulário, que será transformado em uma distribuição de probabilidades a partir de uma camada *softmax*. Dada esta distribuição, ou a palavra mais provável é gerada como saída, ou os $k$ melhores caminhos até este momento são identificados por um algoritmo de busca em feixe (*beam search*) em que $k$ é o tamanho do feixe (LIN; NG, 2019).

A utilização de modelos pré-treinados e, de modo mais geral, de aprendizado por transferência, vêm se tornando comum em trabalhos de PLN, incluindo a sumarização automática. Um exemplo de trabalho que busca se beneficiar de modelos baseados em aprendizado por transferência é a *framework* T5 (*Text-to-Text Transfer Transformer*), proposta por Raffel et al. (2020), que cobre as tarefas de sumarização, *question answering*, classificação de textos, tradução, entre outras.

Há também uma série de trabalhos que buscam aplicar técnicas de aprendizado não supervisionado ou semi-supervisionado para construir modelos de sumarização. A principal vantagem destes modelos é dispensar, ou minimizar a necessidade de sumários de referência, reduzindo o esforço humano de coletar e anotar grandes conjuntos de textos. Zhou e Rush (2019), por exemplo, propõem um modelo que utiliza dois modelos de linguagem: um modelo genérico pré-treinado e outro que é treinado para uma tarefa específica, neste caso, com o objetivo de produzir textos com sentenças mais curtas. Dessa forma, o principal objetivo do treinamento será produzir textos mais curtos, a partir do texto-fonte, enquanto mantém uma alta similaridade contextual com a mesma. Modelos mais recentes, como os *transformers*, também impulsionam estes resultados. Yang et al. (2020) propõem a utilização de um *transformer* pré-treinado em uma base de dados de notícias, seguido por um ajuste fino com o objetivo de maximizar a similaridade semântica entre o sumário candidato e o texto original, enquanto utiliza um *denoising autoencoder* para evitar que o modelo aprenda meramente a copiar o texto de entrada.

Outra forma de utilizar técnicas de aprendizado não supervisionado é a partir do uso

de métodos de avaliação extrínsecos, avaliando um sumário candidato a partir do quão ele auxilia uma determina tarefa. Scialom et al. (2019), por exemplo, calculam a qualidade de um sumário se baseando na capacidade de se responder uma série de questões sobre o texto de origem a partir deste sumário. Estas questões são geradas automaticamente, dispensando a necessidade de intervenção humana.

A seguir são apresentadas alguns dos últimos trabalhos que atingiram o estado da arte em sumarização abstrativa.

#### 2.2.3.1 PEGASUS

Assim como acontece com os métodos extrativos, a sumarização abstrativa se beneficiou com o surgimento de novos modelos, como o BERT. O sistema PEGASUS (ZHANG et al., 2020) alcançou o estado da arte em diversas bases de dados a partir da aplicação deste modelo.

A principal estratégia do PEGASUS é realizar uma etapa de pré-treinamento que se assemelhe ao máximo a tarefa de sumarização propriamente dita, tendo como hipótese que isso irá permitir a realização de um ajuste fino (*fine-tuning*) melhor e mais rápido. Na prática, durante o pré-treinamento do modelo, a técnica de mascarar palavras, utilizada no BERT, é aplicada também para sentenças completas, de forma que o modelo aprenda a identificar estas lacunas, como pode ser observado no diagrama da Figura 2.

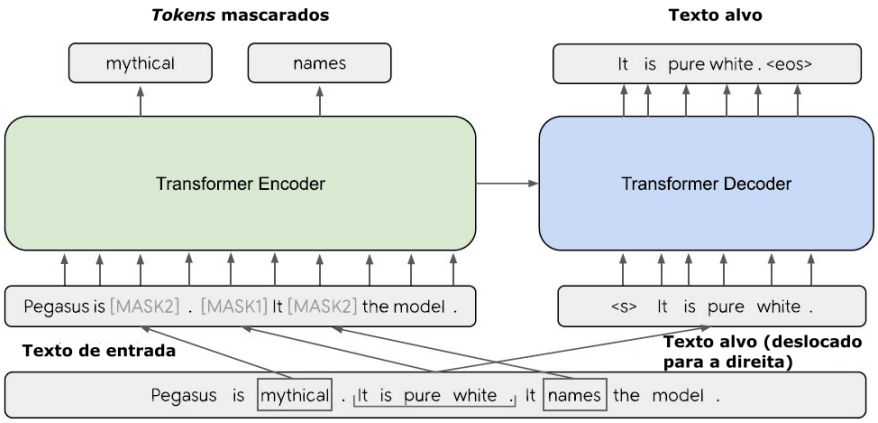


Figura 2 – Arquitetura base da PEGASUS. Adaptado de (ZHANG et al., 2020)

Com o modelo pré-treinado, é realizado o ajuste fino utilizando bases de dados de sumarização. Em seis bases diferentes esta técnica atingiu o estado da arte utilizando somente 1000 amostras. Este grande ganho de desempenho não é interessante apenas pela possibilidade de se adaptar a novas bases mais rapidamente, mas também pelo fato de dispensar a necessidade de bases muito grandes, facilitando aplicações em diversos cenários do mundo real.

#### 2.2.3.2 GSum

Os métodos de sumarização abstrativa têm a vantagem de serem mais flexíveis, podendo gerar sumários mais fluentes, coerentes, próximos do nível humano, em comparação à sumarização extrativa. Porém, isso também gera alguns problemas, como apontado por Dou et al. (2021). Um primeiro problema é a possibilidade de gerar sumários não fiéis ao documento original, contendo erros factuais ou até conteúdos que sequer foram mencionados no texto. Outro problema é a dificuldade em controlar o conteúdo dos sumários, sendo difícil definir com antecedência quais informações do documento original devem ser abordadas.

Buscando mitigar estes problemas, os autores propõe a *framework* GSum, que permite a criação de métodos de sumarização abstrativa guiados, que utilizam diversas pistas ou sinais de orientação para buscar restringir o conteúdo do sumário, de forma que ele não se distancie do texto original, e permita um certo controle sobre o conteúdo a partir de entradas especificadas pelo usuário. As pistas de orientação definidas no GSum são: sentenças destacadas no documento de origem, palavras-chave, relações e sumários de documentos similares.

Durante a fase de teste, as pistas de orientação podem ser extraídas automaticamente ou especificadas pelo usuário. Já na fase de treinamento é utilizado um sistema "oráculo" para obter estas pistas do texto ou do corpus. Para gerar as sentenças destacadas no documento de origem, por exemplo, é utilizado um método de sumarização extrativa, buscando obter as sentenças mais relevantes do texto.

Este trabalho atingiu o estado da arte em diversas bases de dados. Além disso, os sumários gerados foram avaliados por humanos, constatando que de fato há uma melhoria no controle das informações contidas no sumário final, através da utilização dos sinais de orientação, e também que os sumários produzidos são mais fiéis ao conteúdo do texto original, em comparação com os gerados por um modelo similar, mas sem a presença das pistas de orientação.

Esta abordagem ainda apresenta diversas possibilidades de exploração, como a inclusão de novas pistas de orientação, ou a incorporação de outras técnicas e estratégias utilizadas em outros trabalhos.

#### 2.2.3.3 Controle de cópia

Como já apontado pelo trabalho anterior, um dos maiores problemas na elaboração de métodos de sumarização abstrativa baseados em aprendizado em profundidade é a dificuldade em controlar o conteúdo dos sumários. Uma estratégia para mitigar este problema é desenvolver sumarizadores que permitam gerar sumários com diferentes taxas de conteúdo reutilizado do texto de entrada, ou seja, com diferentes taxas de cópia do documento original. De forma geral, sumários que contenham uma quantidade apropriada de conteúdo copiado são mais desejáveis do que os altamente abstrativos, pois tendem a preservar melhor o conteúdo do texto original.

Diversos trabalhos se basearam nesta estratégia de controlar a taxa de cópia dos sumários gerados (SEE; LIU; MANNING, 2017; CHEN; BANSAL, 2018; GEHRMANN; DENG; RUSH, 2018; WEBER et al., 2018; KRYŚCIŃSKI et al., 2018), porém todos eles foram treinados apenas com um sumário de referência por amostra para produzir saídas com uma taxa de cópia fixa, relativa a taxa apresentada pelo corpus. Neste sentido, o trabalho apresentado em (SONG et al., 2020) implementa uma *framework* que permita gerar sumários abstrativos com diferentes taxas de cópia, treinando com dados de corpus que também possuam apenas um sumário de referência por texto.

Para isso, os autores argumentam que um sistema de sumarização abstrativa não precisa, necessariamente, ser treinado com todas as palavras dos sumários de referência, de forma que algumas palavras podem ser mascaradas, e o modelo aprende a prevê-las. A partir disso, dividi-se as palavras dos sumários em duas classes principais: as palavras vistas, que aparecem no texto-fonte, e as não vistas, que não aparecerem. Se o sumarizador é treinado para prever apenas as palavras vistas, ele aprende a copiar os termos do texto de origem, produzindo sumários extrativos. Por outro lado, quanto mais palavras não vistas são utilizadas durante o treinamento, o modelo tende a aprender não apenas a copiar trechos do documento de entrada, mas também gerar novas palavras. Controlar a taxa de máscaras aplicadas a palavras vistas e não vistas permite que sumarizador gere sumários com mais ou menos cópia.

Os resultados obtidos alcançaram o estado da arte, considerando a base de dados Gigaword. Além disso, os sumários gerados por este sistema e por outros, frutos de trabalhos anteriores, foram avaliados por humanos em relação a gramática dos textos, o grau de informação (o quão o sumário cobriu o conteúdo relevante do texto original) e o grau de fidelidade (o quão o sumário preservou o sentido do texto original). Os resultados do sistema apresentado pelos autores foi superior aos dos outros métodos, e até mesmo do que os sumários produzidos por humanos, que também são imperfeitos, sendo sujeitos a erros.

## 2.3 Sumarização entre línguas e multilíngue

Como apresentado anteriormente, a abordagem mais popular para a realização de sumarização automática é a baseada em aprendizado em profundidade, em especial com a utilização de modelos *seq2seq*. Porém, estes modelos são fortemente orientado a dados, ou seja, é necessário uma grande quantidade de textos anotados com sumários para que estes modelos possam ser treinados de forma eficaz. Por este motivo, a maior parte dos trabalhos de sumarização são realizados com textos na língua inglesa, devido a disponibilidade de grandes bases de dados (HASAN et al., 2021). Neste sentido, a sumarização entre línguas (*cross-lingual summarization*) e a sumarização multilíngue (*multilingual summarization*) podem fornecer meios promissores para explorar a realização de sumarização de textos em outras línguas, em que a quantidade de dados disponíveis seja escassa.

O processo de sumarização entre línguas consiste em gerar um sumário em uma língua diferente do texto-fonte. Essa tarefa pode ser desempenhada através da combinação de um sistema de sumarização monolíngue com um sistema de tradução automática, porém ambos estes processos consistem em problemas complexos e não completamente solucionados, tornando a sumarização entre línguas ainda mais desafiadora. Com este *pipeline* em mente, Wan, Li e Xiao (2010) descrevem duas abordagens possíveis: sumarizar-e-traduzir (*summarize-then-translate*) e traduzir-e-sumarizar (*translate-then-summarize*). A primeira abordagem é preferível, uma vez que a tradução é aplicada ao sumário, ou seja, a um número menor de sentenças, evitando um maior índice de erros resultantes da tradução e demandando menor esforço computacional. Porém, esta abordagem só é aplicável quando há disponível alguma base de dados de sumarização com muitos recursos para a língua do texto-fonte. Se este não for o caso, a abordagem traduzir-e-sumarizar se torna a única opção viável (OUYANG; SONG; MCKEOWN, 2019). Wang et al. (2022b) trazem uma terceira classificação possível, que é a abordagem de ponta-a-ponta (*end-to-end*). Nos últimos anos, a maioria dos trabalhos publicados nessa área se concentraram nas abordagens *translate-then-summarize* (ZHANG; ZHOU; ZONG, 2016; PONTES et al., 2018; WAN et al., 2019; OUYANG; SONG; MCKEOWN, 2019) e *end-to-end* (NGUYEN; LUU, 2022; JIANG et al., 2022; LIANG et al., 2022; WANG et al., 2022a).

A sumarização multilíngue, por sua vez, acontece quando o modelo é treinado a partir de amostras de múltiplos idiomas, de forma que, idealmente, um único modelo se torna capaz de sumarizar textos de diferentes línguas. A vantagem deste tipo de sumarização é que idiomas que possuam similaridades morfológicas podem ganhar uma certa vantagem pela transferência de conhecimento durante o treino multilíngue, o que não ocorreria em um treino monolíngue tradicional. Segundo esta abordagem, Hasan et al. (2021), trazem uma base de dados contendo texto anotados com sumários abstrativos em 44 línguas, e também propõe um modelo de sumarização multilíngue. Entre os resultados, os autores conseguem um desempenho considerável para línguas com poucos recursos, para as quais não haviam resultados prévios conhecidos.

## 2.4 Medidas de avaliação

Um problema chave da sumarização automática é como avaliar um determinado método, como determinar a qualidade dos sumários produzidos por ele. As técnicas de avaliação de sumários podem ser divididas em qualitativas e quantitativas.

Técnicas qualitativas buscam avaliar satisfação do usuário com relação aos sumários gerados. Essas técnicas normalmente consistem na avaliação humana dos sumários. Por outro lado, as técnicas quantitativas buscam determinar uma medida de qualidade do sumário baseado em seu conteúdo e no conteúdo do texto original (GUPTA; GUPTA, 2019). Em particular,

as técnicas quantitativas são interessantes pois, normalmente, não dependem da avaliação humana, podendo ser aplicadas de forma automática, o que é essencial para métodos baseados em aprendizado de máquina, por exemplo.

Além disso, uma medida pode estar relacionada a diferentes aspectos do texto. No caso da sumarização, normalmente são consideradas medidas baseadas em conteúdo, que buscam avaliar o quanto o conteúdo do sumário corresponde ao texto original ou ao sumário de referência. Mas também pode ser importante avaliar aspectos de qualidade textuais como a coerência e estrutura do texto, corretude gramatical, ausência de redundâncias e clareza de referências na utilização de substantivos e pronomes. Medidas de qualidade de texto como essas dificilmente podem ser feitas automaticamente, normalmente necessitando de avaliações qualitativas manuais (STEINBERGER; JEŽEK, 2012).

Dentre as medidas quantitativas utilizadas, as mais comuns são as pertencentes ao pacote ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (LIN, 2004), que consiste em um conjunto de medidas baseadas em conteúdo. A princípio, era utilizada principalmente a medida de cobertura deste pacote, que pode ser denotada por ROUGE$_{recall}$. Esta medida busca indicar o quanto do sumário de referência permanece no sumário gerado, calculando o número de sobreposições de *n-grams* entre o sumário gerado ($SG$) e um sumário de referência ($SR$), considerando uma base de dados anotados. Dessa forma, a pontuação $ROUGE - n_{recall}$ para um certo sumário candidato é calculada da seguinte forma:

$$ROUGE - n_{recall} = \frac{\sum_{r \in SR} \sum_{g \in ng(r)} count_{r,s}(g)}{\sum_{r \in SR} \sum_{g \in ng(r)} count_r(g)}, \tag{1}$$

em que $ng(r)$ é o conjunto de *n-grams* para cada sentença do sumário de referência, $count_{r,s}(g)$ é o número máximo de coocorrências dos *n-grams* entre o sumário candidato e o sumário de referência e $count_r(g)$ é o número de *n-grams* no sumário de referência.

Além da cobertura também pode ser calculada a precisão (Equação 2), que indica o quanto do sumário gerado coincide com o sumário de referência, e a medida-F (Equação 3), que consiste na média harmônica entre as duas medidas anteriores (MARTSCHAT; MARKERT, 2017), sendo esta a mais utilizada na atualidade para a avaliação de sumários gerados. Por motivos de simplificação, a medida $ROUGE - n_{f-score}$ será referida apenas por $ROUGE - n$:

$$ROUGE - n_{precision} = \frac{\sum_{r \in SR} \sum_{g \in ng(SG)} count_{r,s}(g)}{|SR| \sum_{g \in ng(SG)} count_s(g)} \tag{2}$$

e

$$ROUGE - n_{f-score} = \frac{ROUGE - n_{precision} * ROUGE - n_{recall}}{2 * (ROUGE - n_{precision} + ROUGE - n_{recall})}. \tag{3}$$

Esta medida também possui diversas variações, como a ROUGE-L, que considera a maior subsequência comum entre os sumários para realizar a comparação (GUPTA; GUPTA, 2019; STEINBERGER; JEŽEK, 2012).

Ainda que a ROUGE seja uma das medidas mais utilizadas, ela possui uma série de limitações e problemas. Esta medida não considera, por exemplo, aspectos de qualidade de um texto, como os citados anteriormente. Além disso, como o cálculo da pontuação ROUGE é realizado a partir da sobreposição de palavras, ou *n-grams*, se um sumário apresentar o mesmo conteúdo do sumário de referência, contendo as mesmas ideias, mas utilizando expressões diferentes, a pontuação final não deve ser muito alta, apesar da qualidade do sumário. Sendo assim, ela sofre especialmente na avaliação de sumários abstrativos (GUPTA; GUPTA, 2019).

Uma métrica que busca avaliar um sumário a partir de um perspectiva semântica, o que é mais interessante para a sumarização abstrativa, é a *Pyramid Score*. Este método é baseado em dados anotados, pois ele calcula a pontuação de um sumário baseado na presença de unidades de conteúdo de sumarização (SCU - *Summarization Content Units*), que buscam representar partes do conteúdo presentes no texto. Para este método funcionar, é necessário uma base em que um texto possua vários sumários anotados por humanos, a partir da qual será calculado o peso de cada SCU, baseado na frequência em que ela ocorre nestes sumários (NENKOVA; PASSONNEAU; MCKEOWN, 2007). Segue-se um exemplo traduzido e adaptado de (NENKOVA; PASSONNEAU; MCKEOWN, 2007), em que é apresentado uma SCU e sua aparição em diferentes sumários de um mesmo texto:

**SCU**: Lopez deixou a GM pela VW

1. a contratação de Jose Ignacio Lopez, um empregado da GM ... pela VW
2. ele deixou a GM pela VW
3. Ele deixou a GM pela VW
4. recrutamento da GM ... Jose Ignacio Lopez
5. Agnacio Lopez De Arriortua, deixou seu trabalho ... na General Motor's Opel ... para se tornar diretor ... Volkswagen
6. Lopez ... GM ... foi para a VW

Um problema deste método é a necessidade de muita intervenção humana no processo, principalmente na identificação de presenças dos SCU no texto, considerando que cada unidade de conteúdo pode aparecer de diferentes maneiras em cada sumário. Houveram tentativas de automatizar ao máximo este processo, como o AutoPyramid (PASSONNEAU et al., 2013) e o PEAK (*Pyramid Evaluation via Automated Knowledge Extraction*) (YANG; PASSONNEAU; MELO, 2016), que de forma geral visam avaliar e identificar a presença de SCU de forma automática, porém, eles acabam falhando em diversos aspectos, como na identificação de paráfrases e de contradições.

Tentando resolver os problemas das métricas anteriores, (VADAPALLI et al., 2017) apresenta uma nova métrica batizada de SSAS (*Semantic Similarity for Abstractive Summarization*), que parte de diversas medidas de similaridade semântica e léxica e do aprendizado de um vetor de pesos, buscando calcular uma pontuação que maximize a correlação com as avaliações humanas dos sumários. Porém, este método demanda muito mais tempo de processamento do que as métricas anteriores.

Mais recentemente estão sendo propostas métricas para avaliar geração de textos, e que podem ser aplicadas para sumarização, baseado na utilização de *embeddings*, que tentar medir a similaridade entre sentenças no campo semântico, não se baseando apenas na correspondência exata de palavras ou *n-grams*, como no caso do ROUGE. O BERTScore (ZHANG et al., 2019), por exemplo, realiza um processo semelhante as métricas mais tradicionais, computando a similaridade de cada *token* da sentença candidata com cada *token* da sentença de referência, mas utilizando a similaridade entre os *embeddings* contextuais de cada *token* ao invés da contagem de correspondências exatas.

A Figura 3 ilustra o processo pelo qual a similaridade entre duas sentenças é calculada através do BERTScore. Dada a sequência de *tokens* da sentença de referência $y = \langle y_1, ..., y_k \rangle$, obtêm-se os *embeddings* contextuais $x = \langle x_1, ..., x_k \rangle$ através do modelo BERT. De forma análoga, para a sentença candidata $\hat{y} = \langle \hat{y}_1, ..., \hat{y}_m \rangle$ obtêm-se os *embeddings* $\hat{x} = \langle \hat{x}_1, ..., \hat{x}_m \rangle$. A partir disso, é calculada a similaridade de cosseno entre cada par de *tokens*. Como são utilizados vetores de *embeddings* pré-normalizados, o cálculo da similaridade entre um *token* de referência $y_i$ e um *token* candidato $\hat{y}_j$ consiste no produto interno $x_i^T \hat{x}_j$.

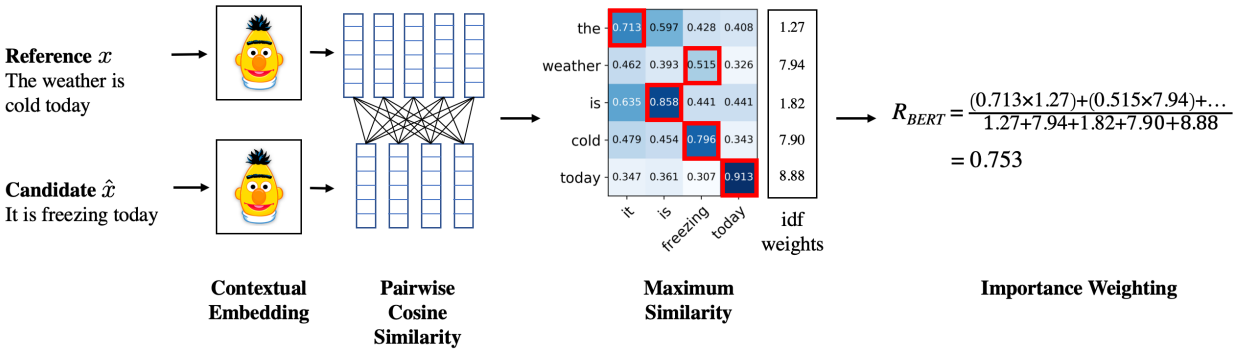


Figura 3 – Procedimento para o cálculo da similaridade entre duas sentenças através do BERTScore. Extraído de (ZHANG et al., 2019)

Então, a correspondência entre os *tokens* é feito de forma gananciosa, de forma que para cada *token* candidato $\hat{y}_j$ é marcado como correspondente ao *token* de referência $y_i$ para qual o valor de $x_i^T \hat{x}_j$ é máximo. Por fim, a precisão, a cobertura e a medida F1 do BERTScore podem ser calculados a partir das seguintes fórmulas:

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j, \tag{4}$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{5}$$

e

$$F_{BERT} = 2\frac{P_{BERT}.R_{BERT}}{P_{BERT} + R_{BERT}}. \tag{6}$$

Opcionalmente, as medidas de similaridade do BERTScore podem ser ponderadas a partir do valor de IDF (*inverse document frequency*) de cada palavra. Uma vantagem deste passo é que palavras raras ganham um peso maior do que as comuns, e conforme demonstrado em trabalhos anteriores (BANERJEE; LAVIE, 2005; VEDANTAM; ZITNICK; PARIKH, 2015), estas palavras tendem a ser mais indicativas para a similaridade de sentenças. Para o cálculo ponderado da cobertura do BERTScore, por exemplo, utiliza-se a Equação 7:

$$R_{BERT} = \frac{\sum\limits_{x_i \in x} idf(x_i) \max\limits_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j}{\sum\limits_{x_i \in x} idf(x_i)}. \tag{7}$$

Outra medida baseada em *embeddings* é o MoverScore (ZHAO et al., 2019). Este método é baseado na *Word Mover's Distance* (WMD) (KUSNER et al., 2015), que consiste em uma função de distância entre sentenças, derivada da distância do movedor de terra, ou métrica de Wasserstein. A ideia geral do WMD é medir a dissimilaridade entre duas sentenças a partir da distância mínima que os *embeddings* de um documento precisam "viajar" para alcançar os *embeddings* de outro documento, através da formulação de um problema de transporte de custo mínimo. De forma geral, o MoverScore estende as ideias do WMD para sequências de *n-grams* utilizando *embeddings* contextuais.

Uma vantagem do MoverScore sobre o BERTScore, por exemplo, é em relação a como eles realizam a correspondência entre as palavras. O BERTScore faz o alinhamento direto entre as palavras, com um *token* candidato correspondente a um único *token* de referência. Já o MoverScore alivia essa restrição, de forma que um *token* candidato pode ter correspondência com diversos *tokens* de referência que, idealmente, estejam semanticamente relacionados com ele. Esta diferença entre os dois métodos é ilustrada pela Figura 4. Além disso, Zhao et al. (2019) também destacam que o BERTScore pode ser definido como uma instância não otimizada, considerando o problema de transporte proposto, do MoverScore.

Tanto o BERTScore quanto o MoverScore apresentaram uma alta correlação com as avaliações humanas, em comparação com as métricas tradicionais, e são mais robustas na detecção de paráfrases.

Ainda existem os métodos de avaliação extrínsecos, que consistem em avaliar a qualidade de um sumário baseado em como ele auxilia uma determinada tarefa. Um exemplo são as avaliações a partir de sistemas de perguntas e respostas, em (EYAL; BAUMEL; ELHADAD, 2019), por exemplo, os autores calculam a qualidade de um sumário baseado no quão bem ele

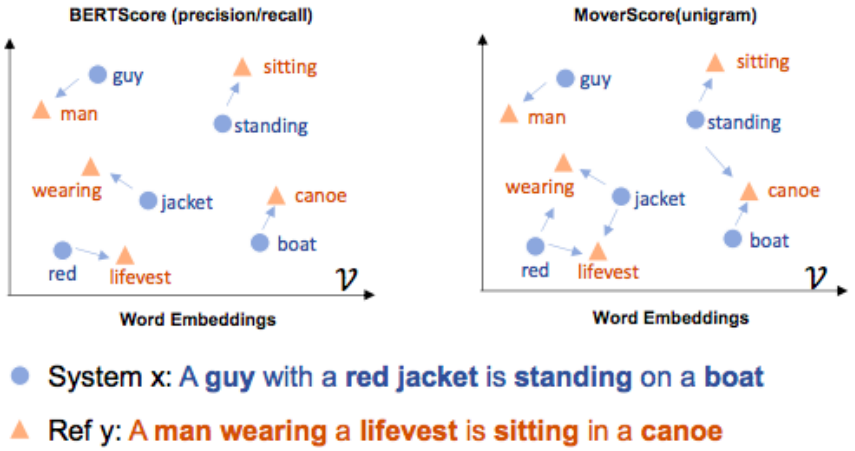


Figura 4 – Comparação do MoverScore com o BERTScore em relação a correspondência de palavras. Extraído de (ZHAO et al., 2019)

pode ser utilizado para obter respostas de questões sobre entidades centrais no texto de origem. Porém, para realizar isso as questões são geradas manualmente e são necessários sumários de referência. Já o método apresentado por (SCIALOM et al., 2019) se baseia na mesma ideia, porém dispensando a necessidade de sumários de referência e sendo capaz de gerar as questões automaticamente através do mascaramento de entidades nomeadas no documento de origem, evitando assim a necessidade de intervenção humana.

Apesar dos avanços obtidos neste sentido, com a criação de novas medidas, elas ainda são pouco utilizadas, e a ROUGE se mantém como a medida padrão. Fabbri et al. (2021) refletem sobre o assunto, e atribuem como principal fator para isso a falta de um protocolo de avaliação, de forma que cada trabalho que propõe uma nova medida faz isso em diferentes cenários, utilizando diferentes anotações humanas como referência e com diferentes metodologias. A falta de uma padronização nestas pesquisas dificulta a comparação entre elas, e a determinação de como cada métrica pode se comportar em cenários reais, diferente daqueles utilizados nas experimentações. Buscando promover um avanço nesta área, os autores tentam definir um protocolo mais completo para a avaliação de sumários, através da introdução do SummEval, que consiste em um conjunto de recursos para modelos de sumarização e para avaliação de sumários, contendo uma base de dados relativamente grande e variada anotada com a avaliação humana de diferentes sumários, assim como ferramentas para aplicação de diferentes medidas já criadas, desde as mais tradicionais até algumas mais recentes.

Em suma, o desenvolvimento de métricas de avaliação de sumários ainda é uma área de pesquisa em aberto e que enfrenta diversos desafios, mas que vem ganhando alguns avanços e que ainda devem ser impulsionados por trabalhos recentes.

## 2.5 Sumarização de textos em português

O principal foco deste trabalho é a sumarização automática de textos em português. Sendo assim, foi feito um levantamento dos trabalhos voltados para este mesmo objetivo. Comparado as estudos existentes na língua inglesa, as pesquisas em português ainda são escassas, com a maioria delas sendo voltadas para sumarização extrativa multidocumento. Em especial, foram encontrados apenas dois trabalhos de sumarização abstrativa em português (CONDORI; PARDO, 2017; INÁCIO, 2021), voltados para a sumarização de opiniões, uma subtarefa da sumarização multidocumento. A maioria dos trabalhos encontrados foram realizados pelo Núcleo Interinstitucional de Linguística Computacional (NILC).

Um dos motivos para a falta de trabalhos nesta área é o pequeno número de bases de dados anotadas com sumários em português, sendo que as disponíveis possuem poucas amostras quando comparadas com bases em inglês.

Nas próximas subseções são apresentadas as principais bases de dados em português disponíveis atualmente, assim como alguns dos principais trabalhos de sumarização automática em português.

### 2.5.1 Bases de dados

Um dos principais corpus em português voltado para sumarização é o TeMário (**TE**xtos com su**MÁRIO**s) (PARDO; RINO, 2003). Esse corpus é composto, basicamente, por textos jornalísticos e seus sumários em português. Os sumários foram escritos por um sumarizador profissional, professor e consultor de editoração de textos em português. Para construir o TeMário foram coletados 100 textos jornalísticos, com um total de 61.412 palavras. Os sumários produzidos para cada texto possuem, aproximadamente, 25-30% do tamanho do texto-fonte. Posteriormente este corpus foi estendido para o TeMário 2006 (MAZIERO et al., 2007), com o acréscimo de mais 150 textos.

Outro corpus voltado para pesquisas de sumarização automática é o Summ-it (COLLOVINI et al., 2007), anotado com informações discursivas, visando fornecer subsídios para enriquecer os modelos de sumarização automática, produzindo resultados mais coerentes e informativos. Para isto, os textos são anotados com informações de correferência e de relações retóricas descritas na RST (*Rethorical Structure Theory*) (MANN; THOMPSON, 1988). Este corpus não possui sumários de referência e é formado por 50 textos jornalísticos do caderno de Ciências da Folha de São Paulo. Em 2016, este corpus foi enriquecido, adicionando, por exemplo, anotações sobre entidades nomeadas e suas relações entre si, constituindo o corpus Summ-it++ (ANTONITISCH et al., 2016).

Existem também alguns corpus voltados para a geração de sumários multidocumentos, como é o caso do CSTNews (LEIXO; PARDO et al., 2008; CARDOSO et al., 2011), que possui 50 coleções de textos, sendo que cada coleção trata de um assunto diferente e tem em

média 4 documentos. Cada coleção é anotada com um sumário, além de outras informações baseadas em teorias do discurso. Cada texto, individualmente, também é anotado com um sumário de referência. Há também o CM2News (FELIPPO, 2016), que além de ser voltado para sumários multidocumentos, também trata da sumarização multilíngue, uma vez que as coleções de textos que formam o corpus possuem tantos textos em português quanto em inglês.

Com relação a sumarização multilíngue, há duas bases de dados recentes que vale destacar: a WikiLingua (LADHAK et al., 2020) e a XL-Sum (HASAN et al., 2021). A primeira inclui textos de 18 línguas, sendo 141.457 artigos em inglês e 81.695 em português, cada um correspondente a algum texto em inglês. Os textos da WikiLingua consistem em tutoriais extraídos da página WikiHow. A XL-Sum, por sua vez, possui textos de 44 línguas, com 301.444 amostras em inglês e 23.521 em português (posteriormente os autores disponibilizaram uma base apenas com textos em português, contendo 71.752 amostras), em que os textos consistem em notícias extraídas da página da British Broadcasting Corporation (BBC). Uma grande vantagem destas bases de dados, para estudos de sumarização em português, é quantidade de amostras anotadas com sumários, consideravelmente maior com relação aos corpora anteriores. Dessa forma, estas bases podem ser úteis para diversos estudos, mesmo que não sejam propriamente de sumarização multilíngue, mas restritos a um dos idiomas disponíveis no corpus.

Ainda há a possibilidade de serem construídos corpus para domínios específicos. O RulingBR (FEIJÓ; MOREIRA, 2018), por exemplo, é voltado para a sumarização de textos jurídicos em português, contendo cerca de 10 mil decisões do Supremo Tribunal Federal Brasileiro.

### 2.5.2 Principais trabalhos

A seguir são apresentados alguns dos principais trabalhos de sumarização em português encontrados. Idealmente os resultados obtidos por estes sistemas deveriam ser comparados, mas em muitos casos essa análise é dificultada pelo fato deles se basearem em corpus diferentes, tratados de maneiras diferentes. Dessa forma, comparações só serão feitas nos casos possíveis.

Um dos primeiros sistemas propostos para realizar sumarização automática extrativa em português foi o GistSumm (PARDO; RINO; NUNES, 2003). O objetivo deste método é identificar a ideia principal do texto, representada por uma sentença (sentença-*gist*), e as informações do texto que a complementem. Para determinar o grau de relevância de cada sentença o método fornece duas opções: o método de palavras-chave (BLACK; JOHNSON, 1988) e o TF-ISF (*Term Frequency-Inverse Sentence Frequence*) (NETO et al., 2000). No método das palavras-chave, cada palavra recebe uma pontuação relativa ao número de ocorrências no texto, e cada sentença tem como pontuação a soma dos valores de cada uma de suas palavras. Já pelo método TF-ISF, cada palavra recebe uma pontuação baseada no número de vezes que ocorreu na sentença e no inverso do número de sentenças em que a palavra ocorreu ao longo do texto.

Em um trabalho subsequente o GistSumm foi aprimorado e ganhou novas funcionalidades (PARDO, 2005). Ainda assim, é um método bastante simples e com resultados longe do nível humano, mas que comumente é citado em outros trabalhos e utilizados como base de comparação para métodos mais recentes.

Outro sistema que costuma ser citado é o SuPor (RINO; MÓDOLO, 2004), que utiliza um classificador Bayesiano para estimar a relevância das sentenças, baseado em uma série de atributos como: tamanho da sentença, frequência das palavras, localização da sentença ou do parágrafo e a ocorrência de substantivos próprios. Este sistema também utiliza os métodos de cadeias léxicas (BARZILAY; ELHADAD, 1999) e de mapa de relações (SALTON et al., 1997) para buscar manter a coesão do sumário.

Um trabalho mais recente apresenta o PragmaSUM (ROCHA, 2017), um método de sumarização extrativa que permite a personificação de sumários pelos usuários através do uso de palavras-chave, buscando aumentar sua precisão e melhorar seu desempenho. Para o treinamento deste modelo foi desenvolvido um corpus próprio, formado apenas por artigos científicos da área educacional.

Em (SODRÉ; OLIVEIRA, 2018) os autores utilizam algoritmos de regressão para, a partir de atributos ao nível da palavra e da sentença, estimar uma pontuação de importância para as sentenças de um texto. Para melhorar o desempenho do método são utilizadas algumas heurísticas, além de diversas técnicas de PLN durante a fase de pré-processamento, como etiquetagem de classes gramaticais, reconhecimento de entidades nomeadas e remoção de *stopwords*. Diversas técnicas de aprendizadas de máquina foram utilizadas nos experimentos, sendo que o algoritmo de regressão Bayesiana obteve os melhores resultados nas medidas do ROUGE-1. Diferente dos trabalhos anteriores, este trata da sumarização multidocumento, utilizando o corpus CSTNews para treinamento. Outros trabalhos que desempenham esta mesma tarefa e podem ser citados são os que propõem os métodos CSTSumm (CARDOSO et al., 2011) e o RC-4 (CARDOSO; PARDO, 2016). A Tabela 1 apresenta a comparação dos resultados obtidos por estes métodos, assim como pelo GistSumm aplicado ao corpus CSTNews.

Tabela 1 – Resultados da comparação entre métodos de sumarização extrativa multidocumento. Adaptado de (SODRÉ; OLIVEIRA, 2018)

| Sistemas | Cobertura | Precisão | Medida-F |
|---|---|---|---|
| (PARDO, 2005) | 0,5764 | 0,5471 | 0,5614 |
| (CARDOSO et al., 2011) | 0,5394 | 0,5597 | 0,5493 |
| (CARDOSO; PARDO, 2016) | 0,5910 | 0,6018 | 0,5964 |
| (SODRÉ; OLIVEIRA, 2018) | 0,6209 | 0,6012 | 0,6109 |

Com relação à sumarização abstrativa, foram encontrados dois trabalhos que abordam diretamente esta tarefa. Sendo o primeiro pesquisa que compara diferentes abordagens, extrativas e abstrativas, de sumarização de opinião (CONDORI; PARDO, 2017). O segundo

trabalho, por sua vez, propõe a utilização de AMR na tarefa de mineração de opinião, com foco na sumarização (INÁCIO, 2021). Ambos os trabalhos utilizam abordagens mais tradicionais de sumarização, como modelos baseados em templates, ou em AMR. Sendo assim, o único trabalho encontrado de sumarização abstrativa em português que utilize modelos baseados em aprendizado em profundidade foi o de (HASAN et al., 2021), já citado anteriormente, que não trata da língua portuguesa especificamente, mas ao propor um modelo de sumarização multilíngue traz resultados também para este idioma.

Como pesquisas análogas à sumarização abstrativa, que podem ser utilizadas de alguma forma para auxiliar a gerar sumários abstrativos, pode-se citar aquelas que tratam de compressão (NÓBREGA; PARDO, 2016; PARDO, 2020) ou fusão de sentenças (SENO; NUNES, 2009).

# 3 Metodologia

Este capítulo descreve a metodologia utilizada no trabalho para responder as perguntas de pesquisa levantadas anteriormente, assim como apresenta as bases de dados utilizadas. O foco é verificar se é possível treinar modelos de sumarização abstrativa baseados em aprendizado em profundidade utilizando bases de dados em português, obtendo resultados satisfatórios.

## 3.1 Bases de dados

Na revisão bibliográfica foram apresentadas as principais bases de dados em português voltadas para sumarização. Neste trabalho, as bases a serem utilizadas e analisadas nos experimentos serão a TeMário (PARDO; RINO, 2003), CSTNews (LEIXO; PARDO et al., 2008), WikiLingua (LADHAK et al., 2020) e XL-Sum (HASAN et al., 2021). Esta escolha deriva do fato de todas estas bases possuírem textos em português anotados com sumários de referências.

Para a realização do treinamento e da avaliação dos modelos nestas bases de dados, estas foram divididas em conjuntos de treinamento, validação e teste. No caso da WikiLingua e da XL-Sum, os próprios autores das bases já as disponibilizaram desta forma. A Tabela 2 apresenta a quantidade de textos de cada conjunto, referente a cada base.

Tabela 2 – Quantidade de amostras nos conjuntos de treinamento, validação e teste.

| Base de dados | Treinamento | Validação | Teste |
|---|---|---|---|
| TeMário | 175 | 26 | 50 |
| CSTNews | 93 | 16 | 31 |
| WikiLingua | 57.159 | 8.165 | 16.331 |
| XL-Sum | 57.402 | 7.175 | 7.175 |

Uma característica importante de se considerar para a realização do treinamento dos modelos e na geração dos sumários candidatos é o comprimento dos textos e dos sumários de referência. A Tabela 3 indica a quantidade média e máxima de palavras em cada uma das bases de dados utilizadas.

Tabela 3 – Quantidade de palavras nas bases de dados avaliadas.

| | Textos | | Sumários | |
|---|---|---|---|---|
| Base da dados | Média | Máxima | Média | Máxima |
| TeMário | 902 | 3.056 | 280 | 607 |
| CSTNews | 340 | 955 | 104 | 287 |
| WikiLingua | 352 | 2.957 | 37 | 463 |
| XL-Sum | 592 | 26.410 | 31 | 206 |

## 3.2 Abordagem proposta

Esta pesquisa iniciou-se com uma revisão bibliográfica detalhada, considerando tanto trabalhos sobre sumarização em português quanto em inglês. Nesta revisão foi possível levantar quais as principais pesquisas nesta área, as principais abordagens, e a partir disso, fundamentar a realização dos experimentos a serem realizados e a escrita deste trabalho.

Um dos desafios com relação à pesquisa sobre sumarização abstrativa em português é a escassez de trabalhos anteriores nessa área. É difícil comparar qualquer resultado até mesmo com métodos extrativos, devido a falta de uma configuração experimental de referência, que adotem medidas de avaliação padrão, como a ROUGE, e a ausência de códigos-fonte de muitos destes trabalhos.

Com isso em mente, foram realizados experimentos com sistemas de sumarização baseados em inglês aplicados a textos traduzidos em português para atingir algum grau de comparação para implementações futuras. A metodologia destes experimentos é apresentada na próxima subseção, e em seguida também é apresentada a metodologia utilizada para o treinamento dos modelos nas bases em português.

### 3.2.1 Experimentos com sistemas treinados com bases em inglês

Com as bases de dados TeMário e CSTNews, que possuem poucas amostras anotadas, foram realizados experimentos preliminares com modelos de sumarização em inglês, utilizando a tradução dos textos do português para o inglês e, posteriormente, dos sumários candidatos em inglês para o português. Nestes experimentos foram utilizados os modelos pré-treinados fornecidos pelos seus autores, sem a realização de um ajuste-fino nas bases TeMário ou CSTNews.

Também foi utilizado o método GistSumm, um dos mais citados trabalhos de sumarização extrativa em português. Além disso, outros modelos extrativos em inglês também foram considerados. Embora o foco deste trabalho seja a sumarização abstrativa, devido a ausência de uma base de comparação, os resultados destes modelos extrativos foram utilizados para

avaliar se os resultados obtidos pelos métodos abstrativos são minimamente satisfatórios.

Além disso, foi utilizado o sistema Lite-T5 Translation (LOPES et al., 2020) para realizar a tradução dos textos entre inglês e português. Este modelo possui uma performance próxima ao estado-da-arte, com a vantagem de requerer um *hardware* mais modesto. A etapa de tradução é fundamental para a abordagem destes experimentos, pois pode impactar diretamente na qualidade dos sumários gerados.

Para os experimentos em si, foram utilizadas duas abordagens, uma para os métodos abstrativos e outra para os extrativos, as quais estão ilustradas nas Figuras 5 e 6, respectivamente.

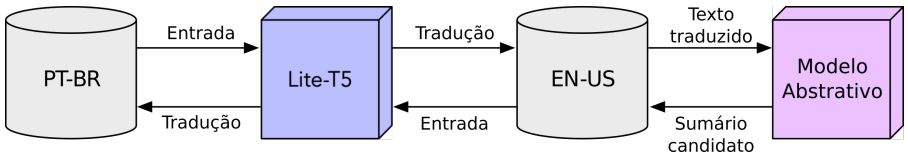


Figura 5 – Fluxograma proposto para a sumarização abstrativa em português a partir de arquiteturas baseadas em inglês.

Os sumarizadores abstrativos recebem os textos traduzidos e geram sumários candidatos, os quais são traduzidos de volta para o português. O processo é semelhante com os sumarizados de maneira extrativa, com a diferença que os sumários candidatos não são diretamente traduzidos para o português. Essencialmente, a ideia é selecionar as sentenças no texto original em português, que correspondem as sentenças selecionadas pelo sumarizadores para compor o sumário final, evitando a degradação do mesmo devido a problemas na etapa de tradução.

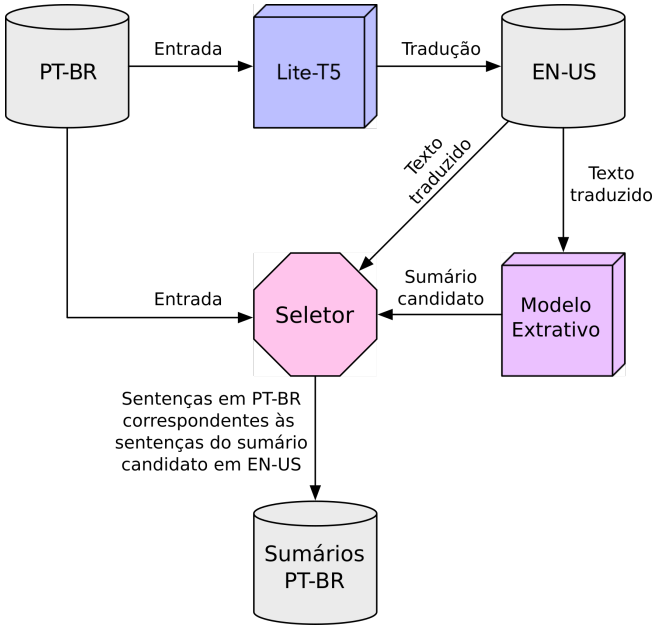


Figura 6 – Fluxograma proposto para a sumarização extrativa em português a partir de arquiteturas baseadas em inglês.

### 3.2.2 Treinamento de modelos com bases em português

A partir do levantamento bibliográfico realizado e avaliando o tempo e recursos disponíveis, foram definidas três linhas de experimentos: utilizando sumarização monolíngue, multilíngue e entre línguas. Para melhor avaliar os resultados obtidos, foi utilizada a arquitetura do modelo T5 para todos estes experimentos. O motivo desta escolha deve-se aos fatos de que o T5 já foi utilizado para experimentos em sumarização abstrativa em inglês (RAFFEL et al., 2020), obtendo bons resultados, próximos do estado-da-arte; e também por possuir modelos pré-treinados em português e também em bases multilíngue, dispensando a necessidade de uma fase de pré-treinamento nos experimentos deste trabalho, o que demandaria mais tempo e recursos.

#### 3.2.2.1 Treinamento monolíngue

Nesta primeira linha de experimentos foi realizado o treinamento monolíngue. No caso das bases WikiLingua e XL-Sum, que são multilíngues, utilizou-se apenas os textos em português.

Considerando o desempenho relatado de modelos pré-treinados em outros trabalhos de sumarização, inclusive quando aplicados a bases de dados com poucas amostras, foi escolhida esta abordagem para o treinamento dos modelos de sumarização abstrativa monolíngue. O modelo pré-treinado selecionado foi o PTT5 (CARMO et al., 2020), que consiste no modelo T5 pre-treinado com a base brWaC (FILHO et al., 2018).

Com o modelo PTT5 pré-treinado, foi realizado o ajuste-fino em cada uma das bases escolhidas, seguindo o processo descrito na Figura 7. Os modelos treinados nas bases WikiLingua e XL-Sum também sofreram um segundo ajuste-fino nas bases TeMário e CSTNews. A ideia deste segundo ajuste-fino é verificar, considerando a pequena quantidade de amostras destas duas últimas bases, se os modelos conseguem adaptar o "conhecimento" que adquiriram nas bases maiores para conseguirem melhores resultados nas bases com poucos recursos.

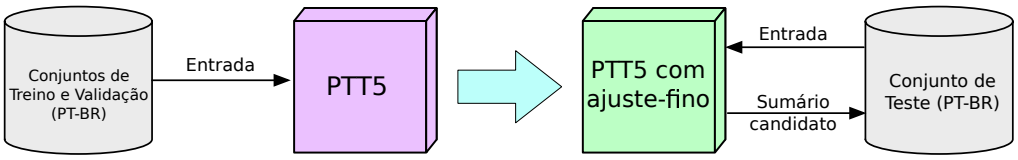


Figura 7 – Fluxograma proposto para o ajuste-fino monolíngue do modelo PTT5.

Para o treinamento dos modelos, a implementação foi realizada em Python, utilizando os módulos Transformers e PyTorch. Para a otimização foi utilizado foi o algoritmo Adam (KINGMA; BA, 2014), variando a taxa de aprendizado entre $3 \times 10^{-5}$ e $3 \times 10^{-4}$. Os modelos foram treinados em uma GPU NVidia K80, com 16GB de RAM, por $30$ épocas. Com relação ao número máximo de *tokens* de entrada e saída, foram utilizados, respectivamente, os valores $512$ e $150$ para as bases CSTNews, WikiLingua e XL-Sum, e $1024$ e $512$ para a TeMário. O

processo de divisão do texto em *tokens*, identificando unidades de texto, em especial, palavras, foi realizado utilizando o *tokenizer* do próprio modelo PTT5. É importante destacar que os parâmetros, de forma geral, foram definidos de forma empírica, com base nos parâmetros utilizados em outros trabalhos, nas características das bases de dados e nas limitações de tempo e recursos disponíveis.

Com os modelos já treinados, a geração dos sumários candidatos foi realizada utilizando o algoritmo de busca em feixe (*beam search*). Diferente do método guloso que sempre escolhe o próximo *token* mais provável para compor o texto, este algoritmo heurístico considera as $k$ palavras com maior probabilidade, e segue desta forma, mantendo as $k$ sequências mais promissoras e descartando as outras, de forma que o número de feixes seja sempre limitado a $k$. Ao final do processo, a sequência com maior probabilidade é selecionada. Neste trabalho, utilizou-se $k = 5$ como o número de feixes.

#### 3.2.2.2 Treinamento multilíngue

Além do treinamento monolíngue, aproveitando o fato das bases WikiLingua e XL-Sum possuírem textos anotados em diversas línguas, também foi realizado um treinamento multilíngue, seguindo o processo descrito na Figura 8. Para isso foi utilizado o modelo mT5 (XUE et al., 2020), que consiste em uma variante multilíngue do modelo T5, treinado com textos em 101 idiomas diferentes.

Para o treinamento multilíngue, foi utilizada a implementação fornecida por Hasan et al. (2021). Como este trabalho já disponibiliza um modelo treinado na base XL-Sum, foi realizado o treinamento apenas na base WikiLingua. Para as bases TeMário e CSTNews foi realizado um ajuste-fino dos modelos treinados na XL-Sum e WikiLingua, segundo os mesmos parâmetros da implementação monolíngue.

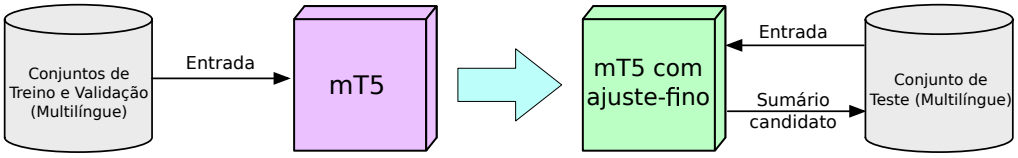


Figura 8 – Fluxograma proposto para o ajuste-fino multilíngue do modelo mT5.

Para fins de comparação, os experimentos monolíngue também foram replicados para o modelo mT5, tanto ao se realizar um ajuste-fino direto no modelo mT5 as bases TeMário e CSTNews, como também ao se realizar um primeiro ajuste-fino monolíngue na WikiLingua e XL-Sum. Sendo assim, é possível verificar se este treinamento realmente propicia algum ganho no resultado final.

#### 3.2.2.3 Treinamento entre línguas

Por fim, também foi experimentado o treinamento entre línguas, seguindo a abordagem *translate-then-summarize*, conforme o processo descrito pela Figura 9. A hipótese avaliada é que pode-se obter vantagem no processo de sumarização abstrativa ao se utilizar modelos em inglês que consistem, ou estão próximos, do estado-da-arte nesta língua, uma vez que eles foram treinados com um número maior de amostras anotadas do que há disponível em português.

Utilizando textos das bases TeMário e CSTNews traduzidos para o inglês, realizou-se o ajuste-fino do modelo T5 (Base), que já havia sido avaliado nestas bases nos experimentos com sistemas em inglês, mas sem a fase de ajuste-fino. Com o modelo treinado, foram gerados os sumários candidatos e então traduzidos para o português novamente. O treinamento e a geração dos sumários candidatos foram realizados nos mesmos moldes dos experimentos monolíngue.

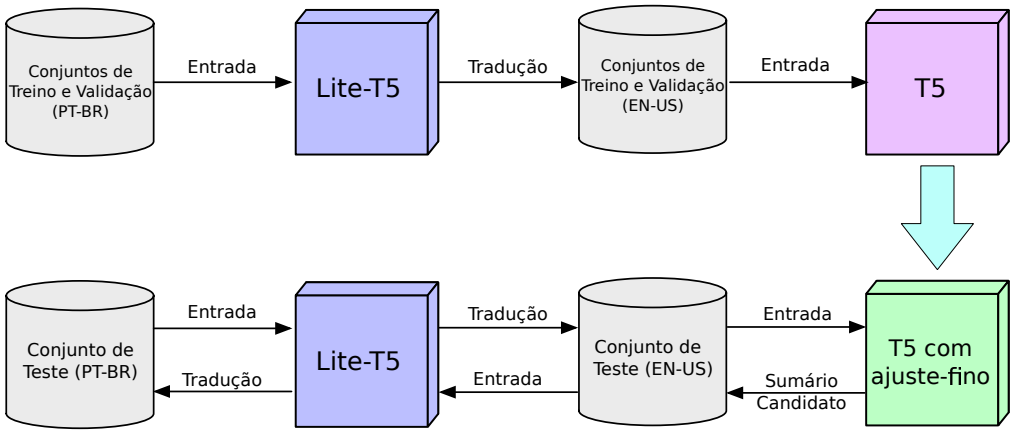


Figura 9 – Fluxograma proposto para o ajuste-fino entre línguas do modelo T5.

### 3.2.3 Avaliação dos sumários candidatos

Os sumários candidatos produzidos pelos modelos foram avaliados pelo conjunto de medidas ROUGE. Além disso, foram calculadas outras medidas de avaliação, como BERTScore e o MoverScore, apresentadas anteriormente. Além dessas medidas de avaliação, também são calculadas as médias das taxas de compressão e de abstração dos sumários produzidos por cada modelo, com o objetivo de fornecer pistas sobre as características dos comportamentos de cada um deles.

A taxa de compressão de um sumário é calculada a partir da razão entre o número de *tokens* do sumário candidato e o número de *tokens* do texto-fonte. Sendo assim, quanto mais a taxa de compressão se aproxima do valor $1$, maior o tamanho do sumário, mais próximo do tamanho do texto-fonte.

Já a taxa de abstração corresponde a porcentagem de *n-grams* que aparecem no sumário candidato, mas não no texto-fonte. Sendo assim, a taxa de abstração de um certo sumário candidato pode ser calculada pela Equação 8, em que $C_n$ representa o conjunto de *n-grams* do sumário candidato e $S_n$ o conjunto de *n-grams* do texto-fonte:

$$Abs_n(C_n, S_n) = 1 - \frac{|C_n \cap S_n|}{|C_n|}. \tag{8}$$

Por fim, foi realizada a análise dos resultados obtidos pelos modelos. Esses resultados foram confrontados entre si e também com os experimentos com sistemas em inglês com o objetivo de verificar se os resultados obtidos foram minimamente satisfatórios, ou se a utilização de modelos não treinados com estas bases ainda atingem melhores resultados.

# 4 Resultados

## 4.1 Experimentos com sistemas treinados em inglês

Nesta seção são apresentados os resultados referentes aos experimentos realizados com sistemas treinados com bases em inglês, aplicados à textos em português traduzidos para o inglês.

As Tabelas 4 e 5 apresentam a avaliação dos sumários candidatos produzidos pelos sumarizadores extrativos aplicados as bases TeMário e CSTNews, respectivamente. A aplicação dos sistemas em inglês, baseados em aprendizado em profundidade, no geral, superam o método GistSumm. Porém, não se pode deixar de considerar o impacto do tamanho do sumário candidato sobre as medidas ROUGE. Outros trabalhos de sumarização apontam que as medidas ROUGE tendem a beneficiar sumários candidatos que possuam tamanho similar aos sumários de referência, penalizando sumários com tamanhos diferentes, ainda que estes tenham um alto nível de similaridade semântica com os sumários de referência (LIU; LUO; ZHU, 2018). Isso ajuda a explicar, por exemplo, as baixas medidas encontradas pelo método de Liu e Lapata (2019) na base TeMário.

Tabela 4 – Resultados dos experimentos com sumarizadores extrativos aplicados à base TeMário, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS), MoverScore (MS) e grau de compressão (Comp)

| Modelo | R1 | R2 | RL | BS | MV | Comp |
|---|---|---|---|---|---|---|
| GistSumm | $38,96$ | $13,60$ | $21,98$ | $69,50$ | $15,96$ | $17,44$ |
| (MILLER, 2019) | $\mathbf{45,77}$ | $\mathbf{18,96}$ | $\mathbf{28,48}$ | $\mathbf{69,62}$ | $\mathbf{23,11}$ | $22,37$ |
| (LIU; LAPATA, 2019) | $22,18$ | $9,03$ | $13,91$ | $66,02$ | $7,44$ | $8,47$ |

Tabela 5 – Resultados dos experimentos com sumarizadores extrativos aplicados à base CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS), MoverScore (MS) e grau de compressão (Comp)

| Modelo | R1 | R2 | RL | BS | MV | Comp |
|---|---|---|---|---|---|---|
| GistSumm | $43,70$ | $22,93$ | $30,50$ | $73,21$ | $20,21$ | $25,4$ |
| (MILLER, 2019) | $49,70$ | $31,05$ | $38,74$ | $72,66$ | $25,55$ | $26,94$ |
| (LIU; LAPATA, 2019) | $\mathbf{53,62}$ | $\mathbf{37,24}$ | $\mathbf{43,07}$ | $\mathbf{78,56}$ | $\mathbf{27,72}$ | $27,47$ |

Já as Tabelas 6 e 8 apresentam os resultados experimentais relativos aos sumarizadores abstrativos, aplicados às bases de dados TeMário e CSTNews, respectivamente. As Tabelas 7

e 9, por suas vezes, apresentam as taxas de compressão e abstração referente aos sumários candidatos.

Tabela 6 – Resultados dos experimentos com sumarizadores abstrativos aplicados à base TeMário, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Modelo | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| Song et al. (SONG et al., 2020) | $\mathbf{41,25}$ | $9,78$ | $\mathbf{21,15}$ | $\mathbf{66,80}$ | $\mathbf{16,88}$ |
| PEGASUS (Large) | $32,56$ | $\mathbf{10,80}$ | $19,85$ | $66,01$ | $13,39$ |
| PEGASUS (Multi-News) | $34,40$ | $6,45$ | $16,40$ | $64,54$ | $10,60$ |
| PEGASUS (Newsroom) | $18,78$ | $5,32$ | $11,72$ | $58,35$ | $1,23$ |
| PEGASUS (XSum) | $10,48$ | $3,35$ | $7,30$ | $58,93$ | $-0,77$ |
| T5 (Base) | $25,94$ | $8,57$ | $15,98$ | $64,62$ | $9,94$ |
| T5 (Large) | $25,78$ | $8,20$ | $15,87$ | $64,78$ | $9,74$ |

Tabela 7 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos sumários abstrativos gerados para a base TeMário.

| Modelo | Abs (1-gram) | Abs (2-gram) | Abs (3-gram) | Comp |
|---|---|---|---|---|
| Song et al. (SONG et al., 2020) | $25,80$ | $65,33$ | $83,72$ | $32,16$ |
| PEGASUS (Large) | $13,83$ | $41,64$ | $57,67$ | $14,32$ |
| PEGASUS (Multi-News) | $29,80$ | $73,60$ | $86,75$ | $23,64$ |
| PEGASUS (Newsroom) | $39,26$ | $74,72$ | $83,61$ | $9,04$ |
| PEGASUS (XSum) | $26,50$ | $70,23$ | $83,69$ | $3,25$ |
| T5 (Base) | $14,04$ | $45,70$ | $62,79$ | $9,89$ |
| T5 (Large) | $16,25$ | $47,33$ | $64,33$ | $9,97$ |

Com relação aos sumários abstrativos, os resultados parecem ser relativamente satisfatórios, considerando as medidas obtidas. No entanto, não foi possível realizar qualquer tipo de análise mais profunda, devido a falta de bases de comparação. Por este motivo, pode-se tentar comparar estes resultados com os obtidos pelos sumarizadores extrativos, desde que não se desconsidere a limitação desta análise, pois tratam-se de abordagens consideravelmente diferentes.

Ao se comparar os resultados obtidos pelos sumarizadores abstrativos com aqueles obtidos pelos extrativos, é possível perceber que estes são, no geral, inferiores. Isso, porém, era esperado, pois os métodos abstrativos possuem maior complexidade que os extrativos, em especial pela dificuldade de gerar sentenças novas que sejam adequadas.

Tabela 8 – Resultados dos experimentos com sumarizadores abstrativos aplicados à base CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2) e ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Modelo | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| Song et al. (SONG et al., 2020) | $33,98$ | $7,49$ | $18,98$ | $65,38$ | $11,56$ |
| PEGASUS (Large) | $37,89$ | $16,32$ | $26,47$ | $67,84$ | $15,14$ |
| PEGASUS (Multi-News) | $36,80$ | $10,30$ | $20,83$ | $68,10$ | $14,59$ |
| PEGASUS (Newsroom) | $35,86$ | $17,99$ | $27,31$ | $67,82$ | $14,56$ |
| PEGASUS (XSum) | $18,82$ | $5,38$ | $13,74$ | $63,50$ | $3,06$ |
| T5 (Base) | $40,97$ | $17,90$ | $28,96$ | $68,15$ | $16,90$ |
| T5 (Large) | $\mathbf{41,45}$ | $\mathbf{18,56}$ | $\mathbf{30,01}$ | $\mathbf{69,45}$ | $\mathbf{18,65}$ |

Tabela 9 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos sumários abstrativos gerados para a base CSTNews.

| Modelo | Abs (1-gram) | Abs (2-gram) | Abs (3-gram) | Comp |
|---|---|---|---|---|
| Song et al. (SONG et al., 2020) | $37,60$ | $76,03$ | $90,16$ | $32,42$ |
| PEGASUS (Large) | $14,45$ | $43,55$ | $58,52$ | $19,45$ |
| PEGASUS (Multi-News) | $30,07$ | $70,56$ | $84,74$ | $43,31$ |
| PEGASUS (Newsroom) | $25,87$ | $54,71$ | $67,76$ | $21,06$ |
| PEGASUS (XSum) | $32,04$ | $75,15$ | $90,06$ | $6,83$ |
| T5 (Base) | $18,54$ | $48,65$ | $64,11$ | $23,39$ |
| T5 (Large) | $16,33$ | $45,64$ | $60,50$ | $23,22$ |

Além disso, as medidas ROUGE tendem a beneficiar os métodos extrativos, conforme foi abordado na Revisão Bibliográfica. Estas medidas penalizam sumários que usem um conjunto distinto de palavras daquele encontrado no sumário de referência, mesmo que preserve o mesmo significado. As medidas BERTScore e MoverScore, em tese, são mais tolerantes com sumários abstrativos, uma vez que avaliam os *embeddings*, e não as palavras em si. Porém nestes experimentos o comportamento destas medidas foi bastante semelhante as ROUGE.

O viés da tradução também pode afetar os resultados finais, devido as limitações inerentes a esta tarefa, como, por exemplo, a ocorrência de palavras desconhecidas nos textos a serem traduzidos.

Outro ponto crítico é que os sistemas utilizados nos experimentos foram treinados em bases diferentes das avaliadas (TeMário e CSTNews). Assim, é importante considerar que tais modelos adquiriram um certo viés em relação a parâmetros como nível de abstração,

comprimento do texto e do sumário e estilo de escrita, a partir das bases de dados em que foram treinadas. O modelo PEGASUS treinado com a base Multi-News, por exemplo, costuma produzir sumários apresentando trechos como "segundo relatório da NBC News" e "da New York Times", que não aparecem nos textos-fonte.

Esses vieses adquiridos pelos modelos também podem justificar a diferença nos resultados entre os sistemas. A arquitetura PEGASUS treinada com a base de dados Multi-News (FABBRI et al., 2019), por exemplo, alcançou resultados mais significativos do que a mesma arquitetura treinada com a base XSum (NARAYAN; COHEN; LAPATA, 2018). Analisando as características dessas bases, um fator que pode ter influenciado esses resultados é o comprimento dos resumos anotados. A Multi-News é anotada com resumos relativamente mais extensos, de forma semelhante as bases TeMário e a CSTNews, enquanto a XSum é anotada com resumos de uma única frase. Esta diferença influencia na taxa de compressão dos sumários gerados pelos modelos treinados, como evidenciado nas Tabelas 7 e 9.

### 4.1.1 Treinamento entre línguas

Os resultados apresentados anteriormente dizem respeito aos modelos treinados em inglês aplicados diretamente as bases em português traduzidas para o inglês, sem a realização de qualquer ajuste-fino. Porém, na metodologia deste trabalho também foi apresentada uma linha de experimento baseada em sumarização entre línguas, segundo a abordagem *translate-then-summarize*, utilizando o modelo T5 (Base). Na Tabela 10 são apresentados os resultados da avaliação deste experimento.

Tabela 10 – Resultados do ajuste-fino do modelo T5 (Base) aplicados as bases TeMário e CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Base de dados | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| TeMário | $43,38$ | $14,26$ | $24,47$ | $69,31$ | $19,81$ |
| CSTNews | $44,56$ | $19,16$ | $31,83$ | $70,89$ | $19,32$ |

As medidas obtidas superam todos os resultados dos modelos abstrativos anteriores, se aproximando até mesmo de alguns modelos extrativos. Este resultado é interessante ao se considerar que o modelo T5 (Base) tinha obtido um desempenho mediano na base TeMário, em comparação aos outros modelos utilizados, e também o fato de que as bases TeMário e CSTNews possuem poucos recursos, com apenas 175 e 94 amostras anotadas em seus conjuntos de treinamento, respectivamente.

## 4.2 Modelos treinados com bases em português

### 4.2.1 WikiLingua e XL-Sum

As bases WikiLingua e XL-Sum têm como grande vantagem em relação à TeMário e à CSTNews o fato de terem um grande número de amostras, o que é uma característica particularmente importante para o treinamento de modelos de aprendizado em profundidade. Não há uma definição precisa que permita dizer claramente se uma base de dados possui poucos recursos ou não, mas Hasan et al. (2021), por exemplo, consideram bases com poucos recursos aquelas com menos de $15.000$ amostras, e tanto a WikiLingua quanto a XL-Sum trazem, só em seus conjuntos de treinamento, mais de $50.000$ amostras.

Na Tabela 11 são apresentados os resultados da avaliação dos modelos treinados nestas bases, a partir dos experimentos descritos na metodologia, considerando tanto o treinamento monolíngue como multilíngue.

Tabela 11 – Avaliação dos modelos treinados com as bases WikiLingua e XL-Sum, conforme as medidas ROUGE-1 (R1), ROUGE-2 (R2) e ROUGE-L (RL)

| Modelo | R1 | R2 | RL |
|---|---|---|---|
| WikiLingua (multi mt5) | $30,68$ | $10,11$ | $22,25$ |
| WikiLingua (mono PTT5) | $33,00$ | $12,58$ | $25,39$ |
| XL-Sum (HASAN et al., 2021) | $37,17$ | $15,90$ | $28,56$ |
| XL-Sum (mono PTT5) | $35,42$ | $14,21$ | $26,25$ |

Para a base WikiLingua não foram encontrados resultados anteriores para a língua portuguesa. Entre os dois modelos treinados neste trabalho, o PTT5 apenas com textos em português e o mt5 com os textos de todas as línguas, o PTT5 obteve um maior desempenho. Diversos trabalhos demonstram que modelos pre-treinados em apenas um idioma apresentam resultados melhores neste mesmo idioma do que modelos pre-treinados em corpora multilíngues (CANETE et al., 2020; SOUZA; NOGUEIRA; LOTUFO, 2019; VIRTANEN et al., 2019), sendo um exemplo o próprio PTT5 (CARMO et al., 2020).

Já para a XL-Sum, os próprios autores da base apresentam os resultados do ajuste fino do modelo mT5 a este corpus, realizando o treinamento com todos os idiomas ao mesmo tempo. Por este motivo, neste trabalho utilizou-se o modelo treinado disponibilizado pelos autores, e o treinamento não foi replicado. Os resultados do modelo monolíngue treinado neste trabalho são ligeiramente inferiores ao do artigo original. A justificativa mais provável para isso deriva do fato de que o modelo multilíngue disponibilizado e utilizado neste trabalho foi treinado por mais tempo e com maiores recursos.

A seguir são apresentados exemplos de textos das bases WikiLingua e XL-Sum, respectivamente, junto com seus sumários de referência e o sumário gerado pelos modelos treinados

neste trabalho. Nestes exemplos, alguns trechos são destacados em coloridos, representando ideias correspondentes entre o texto-fonte e os sumários, para auxiliar a análise dos mesmos.

É importante ressaltar que os comentários sobre estes exemplos visam elucidar algumas características dos sumários gerados, mas não obter conclusões mais gerais, para o qual seria necessário uma análise qualitativa mais rigorosa.

#### 4.2.1.1 Exemplo de sumarização na base WikiLingua

**Texto-fonte**

Falar pessoalmente irá possibilitar uma troca livre de informações e evitar falhas na comunicação, o que pode acontecer se isso for tratado por e-mail. Exponha o problema e então ouça o que todos têm a dizer. Se está preocupado em anotar a conversa, pode gravá-la usando um telefone ou um gravador de voz. Caso não esteja no papel de liderança, talvez prefira falar com o líder antes de falar com os seus colegas. Pode fazer isso se preferir que a questão seja levantada por ele e não por você. Não acuse ou culpe outros membros de terem causado o conflito, mesmo se acreditar nisso. O foco dos seus comentários e argumentos deve ser a questão e a resolução dele pela equipe. [...] Fale e aja profissionalmente. Mesmo que ache que está certo, é importante que escute o lado de todos. [...] Na maioria dos casos, conflitos dentro de uma equipe acontecem porque as pessoas têm perspectivas e histórias diferentes, mas isso é uma coisa boa! Entenda a opinião de todos e porque eles se sentem assim para achar a melhor solução para o problema. No final das contas, isso ajudará o seu time a produzir melhores resultados. [...] Permita que o conflito se torne um gatilho de sessões de brainstorming, com o objetivo de escolher as melhores ideias. [...] Sempre que possível, colegas devem ceder de ambos os lados para que todos possam decidir o rumo que o grupo vai tomar. Esteja aberto a um compromisso que seja o melhor para todos. Se não consegue chegar a um acordo por conta de algum tipo de restrição, ofereça aos colegas que não vão conseguir o que querem, outra coisa de interesse deles. Dessa forma, eles ainda se sentirão parte da decisão. [...]

**Sumário de referência**

Converse sobre a questão pessoalmente com os seus colegas. Foque sua preocupação no problema e não nos seus colegas. Dê a todos a chance de expressar as suas opiniões, se for você o líder. Use o conflito para gerar novas ideias, se for possível. Chegue a um meio-termo para que todos se sintam incluídos.

**Sumário gerado pelo modelo multilíngue (mt5)**

Faça a conversa pessoalmente. Não culpe outros membros. Converse com o líder antes de falar com eles. Seja profissional. Fale com todos os seus colegas.

**Sumário gerado pelo modelo monolíngue (PTT5)**

Converse pessoalmente com todos os membros da equipe. Não acuse ou culpe os outros

membros do time. Escute o lado de todos. Ouça Avalie as perspectivas dos seus colegas.

**Análise dos sumários**

Os textos da WikiLingua consistem em tutoriais extraídos da página WikiHow, sendo assim, os sumários anotados consistem basicamente em um resumo dos principais passos contidos no texto. Ao observar o sumários gerados para o exemplo apresentado anteriormente, é possível perceber que os modelos treinados adquiriram essa mesma característica, trazendo textos muito mais sucintos que o original, com sentenças breves e imperativas. Também é perceptível que os sumários gerados possuem um certo nível de abstração, não se limitando a extração de sentenças do texto-fonte, sendo capazes de parafrasear ou encurtar sentenças e até de produzir sentenças novas a partir de duas ou mais.

#### 4.2.1.2 Exemplo de sumarização na base XL-Sum

**Texto-fonte**

"A tendência de queda da taxa de juros no Brasil é real, é visível", disse Meirelles, que participou na capital americana de uma série de reuniões e encontros com banqueiros e investidores que aconteceram paralelamente às reuniões do Fundo Monetário Internacional (FMI) e do Banco Mundial (Bird) no fim de semana. Para o presidente do BC, a atual política econômica do governo e a manutenção da taxa de inflação dentro da meta são fatores que garantem queda na taxa de juros a longo prazo. "Mas é importante que nós não olhemos para isso apenas no curto prazo. Temos que olhar no médio e longo prazos", disse Meirelles. [...] Neste domingo, Meirelles participou da cerimônia de entrega do prêmio "Banco Central do ano", oferecido pela revista The Banker à instituição que preside. "Este é um sinal importante de reconhecimento do nosso trabalho, de que o Brasil está indo na direção correta", disse ele. [...]

**Sumário de referência**

O presidente do Banco Central, Henrique Meirelles, disse neste domingo em Washington que há uma tendência "real" de queda nas taxas de juro em vigor hoje em dia no Brasil.

**Sumário gerado pelo modelo dos autores** (HASAN et al., 2021):

O presidente do Banco Central, Henrique Meirelles, disse neste domingo que a tendência de queda da taxa de juros no Brasil é visível.

**Sumário gerado pelo modelo monolíngue (PTT5)**

O presidente do Banco Central, Henrique Meirelles, disse neste domingo, em Washington, que a taxa de juros no Brasil é real, mas que o Brasil está indo na direção correta.

**Análise dos sumários**

No exemplo da XL-Sum, também é possível perceber a capacidade do modelo de trazer

informações do texto-fonte de uma forma muito mais sucinta, assim como de gerar novas sentenças a partir de ideias dispersas ao longo do texto. Porém, este exemplo também evidencia um dos problemas mais recorrentes em sumarizadores abstrativos, que é em relação à correção factual, ou à confiabilidade. O sumário gerado pelo modelo deste trabalho não aponta a queda da taxa de juros, conforme o texto-fonte, dizendo apenas que "a taxa de juros no Brasil é real", o que é pouco esclarecedor, mas soa uma visão pessimista, contrária ao texto original. O sumário ainda termina de forma um pouco mais otimista: "mas que o Brasil está indo na direção correta". Ainda assim, o sentido que o sumário gerado traz difere do apresentado no texto-fonte e no sumário de referência.

### 4.2.2 TeMário e CSTNews

Ao contrário das bases WikiLingua e XL-Sum, a Temário e a CSTNews apresentam poucos recursos, o que dificulta o aprendizado de modelos profundos. Neste trabalho, foram realizados o ajuste-fino de modelos pre-treinados a estas bases, buscando verificar se resultados minimamente satisfatórios são alcançados, mesmo com a limitação de recursos. A seguir são apresentados os resultados dos treinamentos monolíngue e multilíngue, respectivamente.

#### 4.2.2.1 Treinamento monolíngue

Na Tabela 12 são apresentadas as avaliações dos sumários gerados pelos modelos PTT5 treinados no corpus TeMário, segundo as medidas ROUGE, BERTScore e MoverScore. Os modelos PTT5-A e PTT5-B foram treinados diretamente na base TeMário, e se diferenciam entre si apenas pelo tamanho dos sumários gerados. Já os modelos PTT5-C e PTT5-D foram previamente treinados nas bases WikiLingua e XL-Sum, respectivamente, para depois sofrerem um segundo ajuste-fino na base TeMário.

Tabela 12 – Resultados dos modelos treinados na base TeMário, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Modelo | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| PTT5-A | $43,90$ | $19,15$ | $27,11$ | $69,73$ | $21,08$ |
| PTT5-B | $48,78$ | $20,45$ | $29,26$ | $70,39$ | $23,33$ |
| PTT5-C (WikiLingua) | $\mathbf{49,91}$ | $\mathbf{20,96}$ | $\mathbf{29,84}$ | $\mathbf{71,24}$ | $\mathbf{25,47}$ |
| PTT5-D (XL-Sum) | $47,95$ | $19,32$ | $27,81$ | $70,72$ | $23,36$ |

O modelo PTT5-C obteve os melhores resultados em todas as medidas utilizadas. Porém, as diferenças não são muito significativas, principalmente entre os modelos PTT5-B e PTT5-C. A maior diferença foi a avaliação do modelo PPT5-A pela medida ROUGE-1, que obteve um valor entre $4,05$ a $6,01$ pontos inferior aos outros modelos.

Com relação as medidas BERTScore e MoverScore, essa proximidade de valores era até esperada, pois o intervalo de valores possíveis tende a ser mais limitado, dependendo do modelo utilizado para calcular os *embeddings* e do corpus em que ele foi treinado (ZHANG et al., 2019). Existem técnicas de redimensionamento deste intervalo, como apontado pelos próprios autores do BERTScore, porém isto demandaria mais tempo para a realização dos experimentos, e os valores atuais já permitem a comparação dos modelos entre si.

Para analisar melhor os resultados apresentados anteriormente, a Tabela 13 apresenta o nível de abstração dos sumários gerados por cada modelo, assim como a taxa de compressão dos mesmos, em relação aos textos-fonte. Também é apresentado o nível de abstração e a taxa de compressão dos sumários de referência em relação aos textos-fonte. Esta análise não tem o objetivo de chegar a conclusões gerais de como estes aspectos influenciam na qualidade de um sumário, mas sim de como eles aparentam influenciar as avaliações destes modelos específicos.

Tabela 13 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos treinados na base TeMário.

| Modelo | Abs (1-gram) | Abs (2-gram) | Abs (3-gram) | Comp |
|---|---|---|---|---|
| Referência | $17,91$ | $55,23$ | $71,26$ | $34,08$ |
| PTT5-A | $2,11$ | $10,12$ | $16,13$ | $21,01$ |
| PTT5-B | $2,77$ | $11,85$ | $18,37$ | $29,42$ |
| PTT5-C (WikiLingua) | $4,04$ | $19,16$ | $29,48$ | $31,03$ |
| PTT5-D (XL-Sum) | $3,20$ | $17,28$ | $27,54$ | $25,94$ |

Em relação ao nível de abstração dos modelos, é possível perceber que os modelos PTT5-C e PTT5-D apresentam resultados consideravelmente superiores aos outros. Isto pode ser explicado pelo fato de que estes dois modelos foram primeiramente treinados com as bases WikiLingua e XL-Sum, respectivamente, para depois sofrerem um segundo ajuste-fino na base TeMário. Como as bases WikiLingua e XL-Sum apresentam muito mais amostras que a TeMário, os modelos conseguiram se adaptar muito melhor a elas, em especial ao estilo de escrita dos sumários anotadas, incluindo o nível de abstração dos mesmos. Com estes resultados apresentados na Tabela 13 pode-se inferir que houve alguma transferência do aprendizado obtido nestas bases para a sumarização na TeMário.

Buscando verificar se existe alguma correlação entre estas medidas e as pontuações obtidas pela ROUGE, BERTScore e MoverScore, foi calculado o coeficiente de correlação de Pearson entre estes valores. Os resultados estão apresentados na Figura 10.

| | R1 | R2 | RL | BS | MS | Comp | Abs1 | Abs2 | Abs3 |
|---|---|---|---|---|---|---|---|---|---|
| **R1** | 1.000000 | 0.866919 | 0.931321 | 0.892422 | 0.945377 | 0.990739 | 0.855617 | 0.701335 | 0.666462 |
| **R2** | 0.866919 | 1.000000 | 0.987843 | 0.702525 | 0.836575 | 0.921836 | 0.714043 | 0.452724 | 0.395350 |
| **RL** | 0.931321 | 0.987843 | 1.000000 | 0.762552 | 0.879995 | 0.970388 | 0.756396 | 0.515049 | 0.462526 |
| **BS** | 0.892422 | 0.702525 | 0.762552 | 1.000000 | 0.977571 | 0.844781 | 0.991780 | 0.944784 | 0.925018 |
| **MS** | 0.945377 | 0.836575 | 0.879995 | 0.977571 | 1.000000 | 0.924757 | 0.973816 | 0.860251 | 0.828140 |
| **Comp** | 0.990739 | 0.921836 | 0.970388 | 0.844781 | 0.924757 | 1.000000 | 0.814999 | 0.623512 | 0.581668 |
| **Abs1** | 0.855617 | 0.714043 | 0.756396 | 0.991780 | 0.973816 | 0.814999 | 1.000000 | 0.947341 | 0.924839 |
| **Abs2** | 0.701335 | 0.452724 | 0.515049 | 0.944784 | 0.860251 | 0.623512 | 0.947341 | 1.000000 | 0.997874 |
| **Abs3** | 0.666462 | 0.395350 | 0.462526 | 0.925018 | 0.828140 | 0.581668 | 0.924839 | 0.997874 | 1.000000 |

Figura 10 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos treinados na base TeMário.

Em primeiro lugar, pode-se notar que as medidas ROUGE, BERTScore e MoverScore apresentam correlações entre si, principalmente a ROUGE-2 com a ROUGE-L e a ROUGE-1 com a BERTScore e MoverScore (sendo ambas as três últimas avaliadas a partir de *n-grams* de tamanho $1$). As menores correlações encontradas entre estas medidas é em relação a BERTScore e MoverScore com a ROUGE-2 e ROUGE-L.

Observando os resultados, é possível perceber que as medidas ROUGE avaliaram o modelo PTT5-B como sendo superior ao PTT5-D, enquanto as medidas BERTScore e MoverScore apresentam o comportamento oposto. Uma possível explicação para esse comportamento é dada ao se observar o nível de abstração destes modelos: o modelo PTT5-D é consideravelmente mais abstrativo do que o PTT5-B, em especial ao se avaliar *n-grams* de tamanho $2$ e $3$. Como já exposto anteriormente, na Revisão Bibliográfica, as medidas ROUGE tendem a prejudicar a avaliação de sumários com alto nível de abstração, já que o cálculo realizado considera apenas a sobreposição de *n-grams*. As medidas BERTScore e MoverScore, por outro lado, ao utilizarem *embeddings* contextuais, buscam comparar os textos no âmbito da semântica, penalizando menos textos com maior nível de abstração.

Agora com relação à taxa de compressão dos sumários, os resultados apresentados na Figura 10 mostram que ela está fortemente correlacionada com as medidas ROUGE e MoverScore, e também apresenta uma correlação considerável com a BERTScore. Como citado anteriormente, sabe-se as medidas ROUGE tendem a beneficiar sumários candidatos que possuam tamanho similar aos sumários de referência.

Agora, para a base CSTNews, a Tabela 14 apresenta as avaliações dos sumários gerados pelos modelos treinados neste corpus, segundo as medidas ROUGE, BERTScore e MoverScore. Os modelos PTT5-E e PTT5-F foram treinados diretamente na base CSTNews, e se diferenciam

entre si apenas pelo tamanho dos sumários gerados. Já os modelos PTT5-G e PTT5-H foram previamente treinados nas bases WikiLingua e XL-Sum, respectivamente, para depois sofrerem um segundo ajuste fino na base CSTNews.

Tabela 14 – Resultados dos modelos treinados na base CSTNews, avaliados com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Modelo | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| PTT5-E | $52,48$ | $\mathbf{35,62}$ | $\mathbf{41,78}$ | $74,35$ | $26,59$ |
| PTT5-F | $52,64$ | $34,04$ | $40,74$ | $74,34$ | $26,70$ |
| PTT5-G (WikiLingua) | $\mathbf{54,15}$ | $34,53$ | $41,01$ | $74,86$ | $\mathbf{28,16}$ |
| PTT5-H (XL-Sum) | $52,34$ | $33,33$ | $38,42$ | $\mathbf{75,18}$ | $27,16$ |

Assim como nos modelos treinados para a TeMário, os resultados não diferem muito entre si. Porém, aqui as melhores medidas obtidas não correspondem a um único modelo. O PTT5-E foi o melhor avaliado pela ROUGE-2 e ROUGE-L, o PTT5-G pela ROUGE-1 e MoverScore, e o PTT5-H pela BertScore.

Buscando melhor analisar os resultados obtidos, levantando hipóteses para quais fatores os influenciaram, segue na Tabela 15 os níveis de abstração dos sumários gerados por cada modelo, assim como as taxas de compressão dos mesmos, em relação aos textos-fonte.

Tabela 15 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos treinados na base CSTNews

| Modelo | Abs (1-gram) | Abs (2-gram) | Abs (3-gram) | Comp |
|---|---|---|---|---|
| Referência | $9,49$ | $30,18$ | $42,29$ | $29,97$ |
| PTT5-E | $1,43$ | $5,80$ | $9,03$ | $31,45$ |
| PTT5-F | $2,08$ | $8,00$ | $12,03$ | $41,94$ |
| PTT5-G (WikiLingua) | $3,16$ | $12,95$ | $20,20$ | $35,62$ |
| PTT5-H (XL-Sum) | $2,69$ | $12,67$ | $20,99$ | $32,38$ |

Vale destacar que o nível de abstração dos sumários de referência da CSTNews, em relação ao texto-fonte, seja para *n-grams* de tamanho $1$, $2$ ou $3$, é consideravelmente menor do que o nível de abstração dos sumários de referência da TeMário. Com isso já é possível intuir que os modelos menos abstrativos devem ser ainda mais beneficiados com relação as medidas ROUGE.

Assim como para a base anterior, os modelos treinados primeiramente com a base WikiLingua ou XL-Sum apresentam um maior nível de abstração em relação aos outros.

Para verificar se existe alguma correlação entre as medidas apresentadas, foi calculado o coeficiente de correlação de Pearson entre elas. Os resultados estão apresentados na Figura 11.

| | R1 | R2 | RL | BS | MS | Comp | Abs1 | Abs2 | Abs3 |
|---|---|---|---|---|---|---|---|---|---|
| R1 | 1.000000 | 0.141372 | 0.329487 | 0.165897 | 0.891592 | 0.160907 | 0.675471 | 0.502947 | 0.427996 |
| R2 | 0.141372 | 1.000000 | 0.901499 | -0.667073 | -0.193811 | -0.286012 | -0.592696 | -0.675947 | -0.693211 |
| RL | 0.329487 | 0.901499 | 1.000000 | -0.813756 | -0.106348 | 0.141031 | -0.471638 | -0.646800 | -0.702337 |
| BS | 0.165897 | -0.667073 | -0.813756 | 1.000000 | 0.594460 | -0.408497 | 0.768390 | 0.903569 | 0.941542 |
| MS | 0.891592 | -0.193811 | -0.106348 | 0.594460 | 1.000000 | -0.052866 | 0.904882 | 0.825950 | 0.782105 |
| Comp | 0.160907 | -0.286012 | 0.141031 | -0.408497 | -0.052866 | 1.000000 | 0.095227 | -0.078309 | -0.149668 |
| Abs1 | 0.675471 | -0.592696 | -0.471638 | 0.768390 | 0.904882 | 0.095227 | 1.000000 | 0.967953 | 0.938134 |
| Abs2 | 0.502947 | -0.675947 | -0.646800 | 0.903569 | 0.825950 | -0.078309 | 0.967953 | 1.000000 | 0.995019 |
| Abs3 | 0.427996 | -0.693211 | -0.702337 | 0.941542 | 0.782105 | -0.149668 | 0.938134 | 0.995019 | 1.000000 |

Figura 11 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos treinados na base CSTNews.

Com relação as medidas de avaliação dos sumários, ao contrário do caso anterior, elas parecem menos correlacionadas entre si, com exceção da medida ROUGE-2 com a ROUGE-L e da medida ROUGE-1 com a MoverScore. Em alguns casos chega-se a apresentar uma forte correlação negativa, como entre a BERTScore e a ROUGE-L.

Também diferente do caso anterior, a taxa de compressão dos sumários não apresenta uma correlação significativa com as avaliações dos mesmos. Por vezes a correlação chega a ser negativa. Isso pode ser explicado pelo fato que, ao exato oposto do que aconteceu com a base TeMário, todos os modelos treinados na CSTNews apresentaram uma taxa de compressão maior do que a dos sumários de referência. Dessa forma, sabendo que as medidas ROUGE tendem a beneficiar modelos que produzem sumários com taxa de compressão próxima a de referência, é de se esperar que conforme a taxa de compressão dos modelos cresça, se afastando ainda mais da taxa de referência, as medidas ROUGE sejam prejudicadas.

Olhando para o nível de abstração dos modelos, pode-se levantar uma hipótese para a diferença entre as medidas de avaliação. Como já havia sido suposto anteriormente, as medidas ROUGE puniram mais severamente os modelos mais abstrativos, em comparação com os modelos da TeMário. Isso não aconteceu com a medida ROUGE-1, que ainda apresenta alguma correlação com o nível de abstração, o que pode ser justificado pelo fato que o nível de abstração de *1-gram* é relativamente baixa em todos os modelos. Porém, as medidas ROUGE-2 e ROUGE-L já apresentam uma correlação negativa mais forte com os níveis de abstração. O BERTScore e o MoverScore, por outro lado, já apresentam uma forte correlação com os níveis de abstração. E de fato, os modelos melhores avaliados por ambos são justamente os que apresentam maior nível de abstração.

De forma geral, treinar os modelos nas bases WikiLingua ou XL-Sum antes do ajuste fino na TeMário ou CSTNews não trouxe melhorias significativas na avaliação dos modelos, principalmente considerando as medidas ROUGE. Por outro lado, esses modelos se sobressaíram em todos os casos na avaliação pelo BERTScore e MoverScore, ainda que ligeiramente em algumas situações. Além disso, esse método também proporcionou maiores níveis de abstração para os sumários gerados.

Para determinar de forma mais conclusiva a qualidade de cada um destes modelos seria necessário realizar uma análise qualitativa dos sumários candidatos de cada um, o que demandaria mais tempo e esforço humano. Buscando suprir de alguma forma a falta de uma análise qualitativa adequada, a seguir são apresentados exemplos de textos da TeMário e da CSTNews com seus respectivos sumários candidatos e de referência. A análise destes casos não permite a obtenção de conclusões mais profundas sobre a qualidade dos modelos, mas pode auxiliar na compreensão do funcionamento dos diferentes modelos apresentados.

##### 4.2.2.1.1 Exemplo de sumarização na base TeMário

Para auxiliar a identificação das ideias principais do texto, os trechos que correspondem a passagens presentes no sumário de referência foram destacados com cores diferentes.

**Texto-fonte**

Dois meses depois de amargar a reprovação de 86% dos cariocas - que o consideraram omisso quando da enchente responsável pela morte de 67 pessoas -, o governador Marcello Alencar conseguiu recuperar, em parte, a credibilidade junto ao eleitor. A pesquisa JB-Vox Populi feita nos dias 30 e 31 de março com 697 moradores da capital apontou um índice de 54% de avaliação positiva do governo do Rio de Janeiro.

Sem dúvida, é uma recuperação, mas a administração do governador teve muito mais conceito péssimo (18%) do que ótimo (4%). A maior parte dos eleitores ouvidos (32%) classificou o governo de regular positivo. Outros 18% consideraram bom; 17% optaram pela avaliação regular negativa; e 10% disseram que a administração Marcello Alencar é ruim. Procurado pelo JORNAL DO BRASIL, o governador não quis comentar o resultado da pesquisa.

O diretor do Vox Populi, Marcos Coimbra, chama a atenção para o fato de os eleitores da capital serem mais normalmente mais exigentes e críticos com os governantes do que os do interior. "No Rio de Janeiro, então, o resultado não é de se estranhar, porque a cidade tem uma tradição oposicionista intensa. Marcello Alencar sofre oposição tanto à direita quanto à esquerda", analisa Coimbra.

Na avaliação do sociólogo, o carioca reage imediatamente a situações extremas. O eleitor da capital foi decisivo para eleger Marcello Alencar no segundo turno. O atual governador teve 56% de votos na cidade do Rio, contra 32% de seu adversário, o pedetista Anthony

Garotinho. Mas, diante de uma calamidade como a enchente de fevereiro deste ano, o carioca protestou: em pesquisa feita também pelo Vox Populi para o JB, 86% consideraram regular, ruim ou péssimo o desempenho do governador na época. E 78% disseram que não votariam em Marcello para prefeito. Entre os cariocas ouvidos, 68% garantiram que não votariam em um candidato apoiado por ele. Passada a indignação com a tragédia causada pelas chuvas, parte do eleitorado carioca voltou a ver pontos positivos na administração de Marcello.

O governador tucano começou o governo em alta junto ao governo federal. Logo no início do ano passado, conseguiu do presidente Fernando Henrique Cardoso a promessa de investimentos no Rio, para obras como ampliação do Porto de Sepetiba e conclusão do metrô. Além disso, conseguiu trazer para o Rio a primeira fábrica de caminhões da Volkswagen, o que lhe garantiu grande prestígio com a população. Nem tudo, porém, foi bem no governo Marcello Alencar. A onda de seqüestros, a partir de junho de 1995, levou os cariocas a cobrar uma polícia mais eficiente e menos corrupta. E as chuvas de fevereiro ajudaram a afundar a popularidade de Marcello, agora parcialmente recuperada.

Hoje, Marcello equilibra-se em uma corda bamba. Teve a aprovação de pouco mais da metade dos entrevistados. Em uma comparação com os demais governadores tucanos, ficou bem atrás (54%) de Tasso Jereissati (76%) e Eduardo Azeredo (70%), mas saiu-se melhor do que Mário Covas, de São Paulo, que teve 47% de avaliação positiva.

**Sumário de referência**

A enchente responsável pela morte de 67 pessoas levou o governador do Rio de Janeiro, Marcello Alencar, a ser reprovado por 86% dos cariocas. Passados dois meses, já obteve 54% de avaliação positiva. É uma recuperação, mas que ainda convive com 18% de péssimo contra só 4% de ótimo.

O diretor do Vox Populi , Marcos Coimbra, alerta para o fato de ser comum essa maior exigência nas capitais , com ênfase na cidade do Rio de Janeiro.

O mesmo carioca que foi decisivo para lhe dar a vitória no segundo turno (56% contra 32% de Garotinho) foi intransigente depois da tragédia das chuvas.

No início de 1995, Alencar conseguiu do presidente Fernando Henrique promessa para investimentos no Rio: ampliação do Porto de Sepetiba e conclusão do metrô. Também trouxe a primeira fábrica de caminhões da Volkswagen , o que ampliou seu prestígio com a população. Mas logo a onda de seqüestros motivou a cobrança de uma polícia eficiente e menos corrupta.

Em comparação com outros governadores tucanos, equilibra-se: 54% , contra 76% de Tasso Jereissati, 70% de Eduardo Azeredo . Mas cima de Covas , que teve 47%.

**Sumário candidato da PTT5-A**

Dois meses depois de amargar a reprovação de 86% dos cariocas - que o consideraram omisso quando da enchente responsável pela morte de 67 pessoas -, o governador Marcello Alencar conseguiu recuperar a credibilidade junto ao eleitor. A pesquisa JB-Vox Populi feita nos dias 30 e 31 de março com 697 moradores da capital apontou um índice de 54% de avaliação positiva do governo do Rio de Janeiro. Sem dúvida, é uma recuperação, mas a administração do governador teve muito mais conceito péssimo (18%) do que ótimo (4%)., A maior parte dos eleitores ouvidos (32%) classificou o governo de regular positivo. Outros 18% consideraram bom; 17% optaram pela avaliação regular negativa; e 10% disseram que a gestão de Marcelo é ruim. Procurado pelo Jornal do Brasil, ele não quis comentar o resultado da pesquisa. O sociólogo Marcos Coimbra chama a atenção para o fato de que os eleitores da

**Sumário candidato da PTT5-B**

Dois meses depois de amargar a reprovação de 86% dos cariocas - que o consideraram omisso quando da enchente responsável pela morte de 67 pessoas -, o governador Marcello Alencar conseguiu recuperar a credibilidade junto ao eleitor. A pesquisa JB-Vox Populi feita nos dias 30 e 31 de março com 697 moradores da capital apontou um índice de 54% de avaliação positiva do governo do Rio de Janeiro. Sem dúvida, é uma recuperação, mas a administração do governador teve muito mais conceito péssimo (18%) do que ótimo (4%)., A maior parte dos eleitores ouvidos (32%) classificou o governo de regular positivo. Outros 18% consideraram bom; 17% optaram pela avaliação regular negativa; e 10% disseram que a gestão de Marcelo é ruim. procurado pelo Jornal do Brasil, ele não quis comentar o resultado da pesquisa. O sociólogo Marcos Coimbra chama a atenção para o fato de que os eleitores do interior serem mais exigentes e críticos com os governantes. "No Rio de Janeiro, por exemplo, os resultados não são de se estranhar, porque a cidade tem uma tradição oposicionista intensa. Marcellos Alecar sofre oposição tanto à direita quanto à esquerda". na avaliação do sociólogo, um carioca reage imediatamente a situações extremas. no segundo turno, 56% de votos na

**Sumário candidato da PTT5-C (WikiLingua)**

A pesquisa JB-Vox Populi, feita nos dias 30 e 31 de março com 697 moradores da capital, apontou um índice de 54% de avaliação positiva do governo do Rio de Janeiro. Sem dúvida, é uma recuperação, mas a administração do governador teve um desempenho muito mais bom (8%) do que ótimo (4%). Segundo o diretor do Vox Populi, Marcos Coimbra, chama a atenção para o fato de os eleitores do interior serem mais exigentes e críticos com os governantes do estado. "No Rio de Janeiro, o resultado não é de estranhar, porque a capital tem uma tradição oposicionista intensa, Marcello Alencar sofre oposição tanto à direita quanto à esquerda.". O eleitor da cidade foi decisivo para eleger Marcellos para prefeito, contra 32% de seu adversário, Anthony Garotinho. Mas, diante da calamidade da enchente de fevereiro deste ano, ele protestou: em pesquisa feita também pelo VoxPuli para a JB, 86% consideraram

regular, ruim ou péssimo o desempenho do candidato apoiado por ele. Entre os moradores ouvidos, 68% garantiram que não votariam em nenhum candidato apoiando-o. Apesar da indignação causada pelas chuvas, parte do eleitorado carioca voltou a ver pontos positivos na administração de Marcelo. O presidente Fernando Henrique Cardoso, prometeu investimentos no Rio, para obras como ampliação do Porto de Sepetiba e conclusão da ponte de São Vicente.

**Sumário candidato da PTT5-D (XL-Sum)**

O governador Marcello Alencar conseguiu recuperar a credibilidade junto ao eleitor com uma pesquisa do Vox Populi feita nos dias 30 e 31 de março com 697 moradores da capital, apontou um índice de 54% de avaliação positiva do governo do Rio de Janeiro. A maior parte dos eleitores não quis comentar o resultado da pesquisa. O diretor do Vox Populi, Marcos Coimbra, chama a atenção para o fato de os eleitores do Rio de Janeiro serem mais exigentes e críticos com os governantes do que os do interior. Na avaliação do sociólogo, o carioca reage imediatamente a situações extremas, como a enchente de fevereiro deste ano, que provocou a indignação dos cariocas com a traugédia causada pelas chuvas que atingiram a capital.

**Análise dos sumários**

Um dos primeiros fatos que é possível perceber é que, apesar do treinamento prévio com as bases WikiLingua ou XL-Sum não ter produzido grandes diferenças na avaliação dos modelos, é notável a diferença na forma que os sumários são gerados, evidenciando ainda mais a diferença no nível de abstração, já aferida anteriormente.

Os modelos PTT5-A e PTT5-B apresentam um baixo nível de abstração, quase se limitando a copiar sentença por sentença do texto-fonte, eventualmente substituindo alguma palavra ou suprimindo algum trecho de uma sentença. A sentença "o governador Marcello Allencar conseguiu recuperar, em parte, a credibilidade junto ao eleitor" teve o trecho "em parte" retirado dos sumários candidatos dos referidos modelos. No caso, esta remoção sequer era desejada, pois acaba alterando, em parte, o sentido do texto.

Devido ao fato destes modelos quase copiarem o texto-fonte, eles também apresentam dificuldade em serem concisos, de forma que mesmo ao limitar o tamanho dos sumários na geração dos mesmos, os modelos acabam gerando textos incompletos, terminando com sentenças inacabadas.

O motivo para estes sistemas ainda conseguirem uma pontuação relativamente alta é que, de fato, eles acabam trazendo muitas das principais ideias do texto, que tendem a se concentrar no início dos textos.

Já os modelos PTT5-C e PTT5-D apresentam um maior nível de abstração. A maioria das sentenças ainda é bastante similar as do texto-fonte, mas em alguns passagens é possível perceber que foram feitas alterações significativas, por vezes até unindo informações de

sentenças diferentes. Como por exemplo, o trecho “O eleitor da capital foi decisivo para eleger Marcello Alencar no segundo turno. O atual governador teve 56% de votos na cidade do Rio, contra 32% de seu adversário, o pedetista Anthony Garotinho”, composto por duas sentenças, foi sumarizado pelo PTT5-C em uma única sentença: “O eleitor da cidade foi decisivo para eleger Marcellos para prefeito, contra 32% de seu adversário, Anthony Garotinho”.

Porém, como é possível perceber nas sentenças citadas anteriormente, os sumários produzidos por estes modelos acabam apresentando outros problemas. Há alguns erros de gramática e ortografia, como na escrita do nome “Marcellos”, e de coerência, como na primeira sentença do sumário gerado pelo modelo PTT5-D.

Há também situações em que as sentenças estão bem escritas, mas não correspondem aos fatos do texto-fonte. No sumário produzido pelo modelo PTT5-C o Marcello Alencar, governador do Rio Janeiro na data da notícia, é referido como prefeito em certo momento, e neste mesmo sumário é apontado que os eleitores do interior são mais exigentes e críticos com os governantes do estado, enquanto a informação que o texto-fonte traz é exatamente a oposta. Em outros momentos, o modelo gera alucinações, como no final do mesmo sumário, que entre as promessas citadas do então presidente Fernando Henrique Cardoso, se encontra a conclusão da ponte de São Vicente. Tal informação nunca é citada no texto. A ponte de São Vicente realmente existe, mas sequer se encontra no estado do Rio de Janeiro. Essa alucinação provavelmente foi provocada por um viés adquirido no pré-treinamento do modelo PTT5 com a brWaC, ou no primeiro ajuste-fino com a base WikiLingua.

Problemas como esses eram, em certa medidas, esperados, sendo desafios recorrentes na área de sumarização abstrativa. A indagação que fica é o quão eles comprometem a qualidade dos sumários produzidos por estes modelos. Para responder esta questão é necessária a realização de uma análise qualitativa mais cuidadosa.

#### 4.2.2.1.2 Exemplo de sumarização na base CSTNews

Para auxiliar a identificação das ideias principais do texto, os trechos que correspondem a passagens presentes no sumário de referência foram destacados com cores diferentes.

**Texto-fonte**

O Parlamento da Turquia aprovou por ampla maioria (507 votos contra 19), nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque, a fim de combater rebeldes separatistas curdos refugiados na região montanhosa. A medida fez com que o preço do petróleo disparasse no cenário internacional , mas a Petrobras garantiu que isso não interferirá no mercado de combustíveis brasileiro.

Pouco antes de o Congresso dar o sinal verde à operação militar, o primeiro-ministro da Turquia, o islamita Recep Tayyip Erdogan, disse que "a paciência do povo turco se

esgotou"com as ações dos guerrilheiros separatistas curdos do Partido dos Trabalhadores do Curdistão (conhecido como PKK) que estão refugiados no Norte do Iraque.

A possibilidade de incursão militar da Turquia, que é membro da Otan (organização militar que reúne 26 países, entre eles EUA, Grã-Bretanha, França e Alemanha), no Iraque está deixando a Casa Branca em estado de alerta. O presidente dos EUA, George W. Bush, pediu nesta quarta-feira que o Exército turco busque uma solução diplomática para a questão.

Washington teme que uma incursão turca possa desestabilizar a mais pacífica região iraquiana e, provavelmente, toda a região do Oriente Médio, levando o Irã, seu arquiinimigo, a intervir.

A Comissão Européia, por sua vez, exortou a Turquia a respeitar a integridade do território iraquiano.

- É crucial que a Turquia continue a atacar esse problema pela cooperação entre as autoridades relevantes - declarou Krisztina Nagy, porta-voz da Comissão Européia.

Erdogan deu de ombros às preocupações de aliados da Otan:

- O que importa é o que o Parlamento diz.

O premier está sob forte pressão interna por causa de uma série de ataques cometidos por guerrilheiros curdos que usam o Norte do Iraque como base. Em Istambul, maior cidade turca, professores e estudantes realizaram grande manifestação contra o terrorismo e a política externa dos EUA.

**Sumário de referência**

O Parlamento da Turquia aprovou um pedido do governo para permitir que tropas do país cruzem a fronteira do Iraque, a fim de combater rebeldes separatistas curdos. O presidente dos EUA, George W. Bush, pediu que o Exército turco busque uma solução diplomática para a questão sob o risco de uma invasão iraniana nessa região do Iraque. A Comissão Européia, por sua vez, exortou a Turquia a respeitar a integridade do território iraquiano. No entanto, o premier turco, Edorgan, que está sob forte pressão devido aos ataques dos guerrilheiros curdos na fronteira, deu de ombros às preocupações de aliados da Otan.

**Sumário candidato da PTT5-E**

O parlamento da Turquia aprovou por ampla maioria (507 votos contra 19), nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque, a fim de combater rebeldes separatistas curdos refugiados na região montanhosa. A medida fez com que o preço do petróleo disparasse no mercado internacional, mas a Petrobras garantiu que isso não interferirá no abastecimento de combustíveis brasileiro. Pouco antes de o congresso dar o sinal verde para a operação militar, o primeiro-ministro dos EUA,

Tayyip Erdogan, disse que "a paciência do

**Sumário candidato da PTT5-F**

O parlamento da Turquia aprovou por ampla maioria (507 votos contra 19), nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque, a fim de combater rebeldes separatistas curdos refugiados na região montanhosa. A medida fez com que o preço do petróleo disparasse no mercado internacional, mas a Petrobras garantiu que isso não interferirá no abastecimento de combustíveis brasileiro. Pouco antes de o congresso dar o sinal verde para a operação militar, o primeiro-ministro dos EUA, Tayyip Erdogan, disse que "a paciência do povo turco se esgotou"com as ações dos guerrilheiros do partido dos trabalhadores do curdistão (conhecido como PKK) que estão refugiados no norte. O exército turco teme que uma incursão turca possa desestabilizar a mais pacífica região do Oriente Médio, levando o Irã, seu arquiinimigo, à intervir. as autoridades

**Sumário candidato da PTT5-G (WikiLingua)**

O parlamento da Turquia aprovou por ampla maioria (507 votos contra 19), nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque, a fim de combater rebeldes separatistas curdos refugiados na região montanhosa. O presidente dos EUA, George W. Bush, pediu para que o exército turco busque uma solução diplomática para a questão. A comissão européia, por sua vez, exortou o governo a respeitar a integridade do território iraquiano.

**Sumário candidato da PTT5-H (XL-Sum)**

O parlamento da Turquia aprovou, por ampla maioria, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque, a fim de combater rebeldes separatistas curdos que estão refugiados na região montanhosa. O presidente dos EUA, George W. Bush, pediu nesta quarta-feira que o exército turco busque uma solução diplomática para a questão. A comissão européia exortou o primeiro-ministro e o premier a atacar esse problema pela cooperação entre as autoridades.

**Análise dos sumários**

De forma similar aos modelos treinados na base TeMário, os dois primeiros PTT5-E e PTT5-F, quase que se limitam a copiar o texto-fonte. A PTT5-F apenas no final do texto omite algumas sentenças (o terceiro parágrafo do texto-fonte) e realiza algumas poucas alterações, sendo elas incorretas, a primeira ao se referir ao Tayyip Erdogan como primeiro-ministro dos EUA, e não da Turquia; e depois substituindo “Washington” por “exército turco” na última sentença do sumário.

Já os dois últimos modelos acabam apresentando um desempenho melhor, gerando candidatos menores em que todas as suas sentenças correspondem as informações destacadas no sumário de referência. Ambos os modelos também demonstram sua capacidade de abstração ao realizar certas alterações nas sentenças, como o PTT5-G suprimindo a data do pedido de George Bush, e o PTT5-H removendo a data e a quantidade de votos do Parlamento da Turquia. O único problema identificado foi na última sentença do sumário da PTT5-H, em que o primeiro-ministro e o premier são referidos como se fossem pessoas diferentes.

Porém, é importante ressaltar que estes dois últimos modelos nem sempre produzem sumários com este nível de qualidade, e em outros textos eles também apresentam problemas semelhantes aos encontrados nos sumários candidatos da TeMário.

#### 4.2.2.2 Treinamento multilíngue

Na Tabela 16 são apresentadas as avaliações dos sumários gerados pelos modelos resultantes do ajuste-fino do modelo multilíngue mT5 à base TeMário, segundo as medidas ROUGE, BERTScore e MoverScore. O modelo mT5-A foi gerado a partir do ajuste-fino direto do mT5 à base TeMário. Os modelos mT5-B e mT5-C, por outro lado, sofreram primeiro um ajuste-fino na base WikiLingua, seguido por um segundo na base TeMário. A diferença entre estes dois modelos é que o ajuste-fino na WikiLingua foi monolíngue para o mT5-B, utilizando apenas textos em português, e multilíngue para a mT5-C, utilizando todos os textos disponíveis nesta base. De forma análoga aos dois anteriores, o modelo mT5-D sofreu primeiro um ajuste-fino nos textos em português da base XL-Sum e o mT5-E é resultado do ajuste-fino na TeMário do modelo de sumarização multilíngue disponibilizado pelos próprios autores da base XL-Sum.

Tabela 16 – Resultados do modelo mT5 treinado na base TeMário, avaliado com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Modelo | Modo de treinamento | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|---|
| mT5-A | Monolíngue | $33,96$ | $13,07$ | $20,09$ | $67,83$ | $17,47$ |
| mT5-B (WikiLingua) | Monolíngue | $\mathbf{37,81}$ | $12,96$ | $20,66$ | $\mathbf{68,50}$ | $18,46$ |
| mT5-C (WikiLingua) | Multilíngue | $37,06$ | $\mathbf{14,19}$ | $\mathbf{22,55}$ | $68,14$ | $\mathbf{19,34}$ |
| mT5-D (XL-Sum) | Monolíngue | $32,70$ | $11,95$ | $19,23$ | $67,59$ | $16,00$ |
| mT5-E (XL-Sum) | Multilíngue | $34,48$ | $13,22$ | $20,57$ | $68,22$ | $16,91$ |

Os melhores resultados obtidos foram pelos modelos que sofreram o primeiro ajuste-fino na base WikiLingua. Os modelos mT5-B e mT5-C obtiveram resultados próximos, de forma que o primeiro ajuste-fino multilíngue não parece ter trazido grande vantagem, em uma primeira análise. No caso dos modelos mT5-D e mT5-E, o segundo se sobressaiu em todas

as avaliações sobre o primeiro, apontando uma melhor performance do modelo multilíngue. Porém, é importante considerar que o ajuste-fino do modelo mT5-E na base XL-Sum não foi realizado neste trabalho, e sim pelos autores da XL-Sum, que dispunham de maiores recursos computacionais e realizaram o treinamento por mais tempo.

Da mesma forma que foi realizado para os experimentos monolíngue, segue a Tabela 17 com o nível de abstração e taxa de compressão dos sumários gerados por cada modelo; e a Figura 12, com os coeficientes de correlação de Pearson entre cada uma das medidas de avaliação anteriores.

Tabela 17 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos multilíngues treinados na base TeMário

| Modelo | Abs (1-gram) | Abs (2-gram) | Abs (3-gram) | Comp |
|---|---|---|---|---|
| Referência | 17,91 | 55,23 | 71,26 | 34,08 |
| mT5-A | 6,98 | 27,95 | 41,32 | 14,45 |
| mT5-B | 11,17 | 38,10 | 52,61 | 20,05 |
| mT5-C | 8,41 | 31,43 | 45,21 | 18,46 |
| mT5-D | 7,85 | 31,31 | 45,97 | 14,06 |
| mT5-E | 8,80 | 33,66 | 47,49 | 14,85 |

| | R1 | R2 | RL | BS | MS | Comp | Abs1 | Abs2 | Abs3 |
|---|---|---|---|---|---|---|---|---|---|
| R1 | 1.000000 | 0.652949 | 0.748330 | 0.867535 | 0.901139 | 0.975723 | 0.745823 | 0.632887 | 0.583028 |
| R2 | 0.652949 | 1.000000 | 0.953493 | 0.525141 | 0.850718 | 0.498051 | 0.074456 | -0.025750 | -0.105315 |
| RL | 0.748330 | 0.953493 | 1.000000 | 0.545258 | 0.898954 | 0.643714 | 0.212597 | 0.118801 | 0.059467 |
| BS | 0.867535 | 0.525141 | 0.545258 | 1.000000 | 0.645059 | 0.803559 | 0.844414 | 0.788738 | 0.722991 |
| MS | 0.901139 | 0.850718 | 0.898954 | 0.645059 | 1.000000 | 0.835594 | 0.383672 | 0.238172 | 0.179825 |
| Comp | 0.975723 | 0.498051 | 0.643714 | 0.803559 | 0.835594 | 1.000000 | 0.800869 | 0.694419 | 0.664968 |
| Abs1 | 0.745823 | 0.074456 | 0.212597 | 0.844414 | 0.383672 | 0.800869 | 1.000000 | 0.983078 | 0.969382 |
| Abs2 | 0.632887 | -0.025750 | 0.118801 | 0.788738 | 0.238172 | 0.694419 | 0.983078 | 1.000000 | 0.994598 |
| Abs3 | 0.583028 | -0.105315 | 0.059467 | 0.722991 | 0.179825 | 0.664968 | 0.969382 | 0.994598 | 1.000000 |

Figura 12 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos mT5 treinados na base TeMário.

As medidas de avaliação da qualidade dos sumários, de forma geral, são correlacionadas entre si. Porém, ao se comparar com os resultados dos experimentos com o modelo PTT5, essa correlação se enfraqueceu em alguns casos, em especial as medidas ROUGE-2 e ROUGE-L

em relação as outras. O correlação do grau de compressão dos sumários com as medidas de avaliação também diminui, caindo mais vertiginosamente em relação a ROUGE-2 e ROUGE-L.

Com relação aos graus de abstração, eles apresentam uma baixa correlação com as medidas ROUGE-2 e ROUGE-L, e uma correlação moderada com as medidas ROUGE-1 e BertScore. Nos experimentos com o PTT5, quase todas as medidas de avaliação estavam pelo menos moderadamente correlacionadas com o grau de abstração. Este resultado não era esperado, visto que as medidas ROUGE tendem a punir sumários com maior grau de abstração. Uma possível explicação para isso é que todos os modelos PTT5 apresentavam graus de abstração relativamente baixos, bem menores do que a abstração dos sumários de referência, de forma que este não foi um fator muito prejudicial para a avaliação dos sumários. Já os experimentos com o modelo mT5 geraram sumários com maiores graus de abstração, e por este motivo a sua correlação com as medidas de avaliação enfraqueceu, em especial a ROUGE-2 e ROUGE-L, chegando a se tornar negativa em alguns casos.

A Tabela 18 apresenta as avaliações dos sumários gerados pelos modelos resultantes do ajuste-fino do modelo multilíngue mT5 à base CSTNews, de forma análoga aos resultados apresentados para a TeMário.

Tabela 18 – Resultados do modelo mT5 treinado na base CSTNews, avaliado com as medidas ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BertScore (BS) e MoverScore (MS)

| Modelo | Modo de treinamento | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|---|
| mT5-F | Monolíngue | $44,90$ | $24,69$ | $29,88$ | $72,12$ | $22,38$ |
| mT5-G (WikiLingua) | Monolíngue | $44,99$ | $25,21$ | $31,38$ | $72,10$ | $22,23$ |
| mT5-H (WikiLingua) | Multilíngue | $43,70$ | $26,27$ | $32,54$ | $71,87$ | $22,18$ |
| mT5-I (XL-Sum) | Monolíngue | $45,94$ | $26,39$ | $30,07$ | $73,33$ | $23,81$ |
| mT5-J (XL-Sum) | Multilíngue | $\mathbf{47,57}$ | $\mathbf{29,31}$ | $\mathbf{36,33}$ | $\mathbf{74,26}$ | $\mathbf{25,31}$ |

Neste caso, os modelos que tiveram um primeiro ajuste-fino na base XL-Sum foram os que obtiveram os melhores resultados, na maioria dos casos, sendo que o modelo multilíngue mT5-J se sobressaiu em todos os resultados. Não se pode concluir que o treinamento multilíngue trouxe melhorias significativas para a qualidade dos sumários. Da mesma forma que ocorreu para a base TeMário, os modelos mT5-G e mT5-H, com primeiro ajuste-fino na WikiLingua, apresentam resultados próximos, com o primeiro se sobressaindo nas medidas ROUGE-1, BertScore e MoverScore, e o segundo na ROUGE-2 e ROUGE-L.

Segue a Tabela 19 com o nível de abstração e taxa de compressão dos sumários gerados por cada modelo; e a Figura 13, com os coeficientes de correlação de Pearson entre cada uma das medidas de avaliação anteriores.

Tabela 19 – Nível de abstração (Abs), em relação aos sumários de referência, e grau de compressão (Comp), em relação aos textos-fonte, dos modelos multilíngues treinados na base CSTNews

| Modelo | Abs (1-gram) | Abs (2-gram) | Abs (3-gram) | Comp |
|---|---|---|---|---|
| Referência | $9,49$ | $30,18$ | $42,29$ | $29,97$ |
| mT5-F | $8,23$ | $28,61$ | $42,07$ | $29,69$ |
| mT5-G | $7,84$ | $25,62$ | $37,41$ | $23,42$ |
| mT5-H | $6,62$ | $24,52$ | $35,81$ | $23,00$ |
| mT5-I | $6,08$ | $21,79$ | $33,44$ | $23,84$ |
| mT5-J | $5,30$ | $20,02$ | $29,78$ | $24,58$ |

| | R1 | R2 | RL | BS | MS | Comp | Abs1 | Abs2 | Abs3 |
|---|---|---|---|---|---|---|---|---|---|
| R1 | 1.000000 | 0.763144 | 0.577250 | 0.958680 | 0.945573 | 0.014398 | -0.638605 | -0.709823 | -0.688099 |
| R2 | 0.763144 | 1.000000 | 0.898717 | 0.864582 | 0.900921 | -0.363483 | -0.919518 | -0.893790 | -0.920406 |
| RL | 0.577250 | 0.898717 | 1.000000 | 0.619816 | 0.678413 | -0.341732 | -0.694016 | -0.667441 | -0.744033 |
| BS | 0.958680 | 0.864582 | 0.619816 | 1.000000 | 0.995280 | -0.127029 | -0.820915 | -0.851822 | -0.824408 |
| MS | 0.945573 | 0.900921 | 0.678413 | 0.995280 | 1.000000 | -0.121512 | -0.841415 | -0.853290 | -0.834211 |
| Comp | 0.014398 | -0.363483 | -0.341732 | -0.127029 | -0.121512 | 1.000000 | 0.513639 | 0.604511 | 0.637956 |
| Abs1 | -0.638605 | -0.919518 | -0.694016 | -0.820915 | -0.841415 | 0.513639 | 1.000000 | 0.965705 | 0.959912 |
| Abs2 | -0.709823 | -0.893790 | -0.667441 | -0.851822 | -0.853290 | 0.604511 | 0.965705 | 1.000000 | 0.992144 |
| Abs3 | -0.688099 | -0.920406 | -0.744033 | -0.824408 | -0.834211 | 0.637956 | 0.959912 | 0.992144 | 1.000000 |

Figura 13 – Coeficiente de correlação de Pearson entre as medidas ROUGE, BERTScore, MoverScore, nível de abstração e taxa de compressão dos modelos mT5 treinados na base CSTNews.

Diferente dos experimentos realizados com o PTT5 na base CSTNews, as medidas de avaliação apresentam uma correlação de moderada a forte entre si. Já com relação ao grau de compressão, os resultados se mantém semelhantes no sentido de que apresentam uma correlação fraca com as medidas de avaliação.

Por outro lado, os graus de abstração apresentam uma correlação negativa de moderada a forte com todas as medidas de avaliação. No caso dos experimentos com o PTT5, as medidas ROUGE-2 e ROUGE-L já puniam moderadamente os modelos que geravam sumários com maior grau de abstração, porém as outras medidas ainda os beneficiavam, chegando a apresentar uma correlação forte com a BertScore e a MoverScore. Com o modelo mT5, porém, em que os sumários apresentam graus de abstrações consideravelmente maiores, esta característica parece ser ainda mais punida, até mesmo pela BertScore e MoverScore, que deveriam ser mais tolerantes com este aspecto, visto que utilizam *embeddings*, ao invés de $n-grams$, para o

cálculo de sua pontuação. Uma hipótese para este comportamento é que os modelos com maior nível de abstração também são aqueles com menor confiabilidade, com maior divergência do sentido original do texto-fonte, e por este motivo eles também são penalizados pela BertScore e MoverScore.

#### 4.2.2.2.1 Exemplo de sumarização na base TeMário

A seguir são apresentados os sumários candidatos gerados pelos modelos mT5 para o mesmo texto da TeMário utilizado de exemplo nos experimentos com o PTT5. Por esse motivo, omite-se o texto-fonte e o sumário de referência.

**Sumário candidato da mT5-A**

A pesquisa JB-Vox Populo feita nos dias 30 e 31 de março com 697 moradores da capital apontou um índice de 54% de avaliação positiva do governo Rio de Janeiro. Sem dúvida, é uma recuperação da credibilidade junto ào eleitor. O governador Marcello Alencar teve 56% de votos na cidade, contra 32% de seu adversário, pedetista Anthony Garotinho. Mas os eleitores do interior são mais exigentes, mas não se estranha quando da enchente responsável pela morte de 67 pessoas.

**Sumário candidato da mT5-B (WikiLingua)**

A pesquisa JB-Vox Populi feita nos dias 30 e 31 de março com 697 moradores da capital apontou um índice de 56% de avaliação positiva do governo do Rio de Janeiro. Segundo o governador Marcello Alencar, conseguiu recuperar, em parte, credibilidade junto à eleitor. Os 18% classificaram bom; outros optaram por ótimo (3%). O maior parte dos eleitores ouvidos (62%) avaliou os resultados de regular positivo. As outras avaliações são positivas. No Rio de Janeiro, ele reage imediatamente. A enchente responsável pela morte de 67 pessoas - diante de uma catastrófa de protestos.

**Sumário candidato da mT5-C (WikiLingua)**

A pesquisa JB-Vox Populi feita nos dias 30 e 31 de março com 6974 moradores da capital, o governador Marcello Alencar conseguiu recuperar sua credibilidade junto ào eleitor. Suas 56% de avaliação positiva do governo Rio de Janeiro. Sem dúvida, não é de se estranhar porque os eleitores são mais exigentes. O diretor da jornada, Marcos Coimbra, alerta para isso, por meio de um índice péssimo do que ótimo (4%). Ele vê imediatamente as situações extremas: diante de uma calamidade cometida pelo enchente de fevereiro deste ano, protestou contra dois partidos do pedetista Anthony Garotinho. mas, apesar do seu adversário, ele reagiu de maneira extrema.

**Sumário candidato da mT5-D (XL-Sum)**

O governador Marcello Alencar conseguiu, em parte, a credibilidade junto com o eleitor. Segundo pesquisa JB-Vox Populi feita nos dias 30 e 31 de março de 697 moradores da capital, um índice de 56% de avaliação negativa do governo do Rio de janeiro. Sem dúvida, não é uma recuperação, mas os outros 18% classificaram negativo; 17% optaram pela classificação positiva; 10% disseram que são ruim. Procurado pelo Jornal do Brasil, ele não quis comentar. O sociólogo Marcos Coimbra, atenção à capacidade dos eleitores serem mais exigentes. Analisa também o comportamento do carioca na cidade, que reage imediatamente contra as situações extremas. No Rio de Janeiro, então, tem uma tradição ofensiva intensa. Mas, mesmo diante de uma enchente que ameaçou 67 pessoas.

**Sumário candidato da mT5-E (XL-Sum)**

O governador do Rio de Janeiro, Marcello Alencar, conseguiu recuperar a credibilidade junto ào eleitor. A pesquisa j-vox populi feita nos dias 30 e 31 de março com 697 moradores da capital mostrou um índice de 54% de avaliação positiva, mas o governo teve muito mais conceito péssimo do que ótimo (4%. maior parte dos eleitores ouvidos) classificou-o de regular positivo; outros 18% optaram por negativa. procurado pelo jornal do Brasil, ele não quis comentar os resultados. O diretor da entidade, Marcos Coimbra, acha que as eleições na capital não são estranhas.

**Análise dos sumários**

Em comparação com os sumários gerados pelos modelos obtidos a partir do PTT5, é possível perceber facilmente uma grande queda na qualidade dos mesmos. Por um lado, os sumários são mais concisos e com maior nível de abstração. Porém, há muito mais erros de gramática e ortografia, concordância, coerência e de confiabilidade modo geral, trazendo algumas informações incorretas, distorcidas ou incoerentes.

#### 4.2.2.2.2 Exemplo de sumarização na base CSTNews

A seguir são apresentados os sumários candidatos gerados pelos modelos mT5 para o mesmo texto da CSTNews utilizado de exemplo nos experimentos com o PTT5. Por esse motivo, omite-se o texto-fonte e o sumário de referência.

**Sumário de referência**

O Parlamento da Turquia aprovou um pedido do governo para permitir que tropas do país cruzem a fronteira do Iraque, a fim de combater rebeldes separatistas curdos. O presidente dos EUA, George W. Bush, pediu que o Exército turco busque uma solução diplomática para

a questão sob o risco de uma invasão iraniana nessa região do Iraque. A Comissão Européia, por sua vez, exortou a Turquia a respeitar a integridade do território iraquiano. No entanto, o premier turco, Edorgan, que está sob forte pressão devido aos ataques dos guerrilheiros curdos na fronteira, deu de ombros às preocupações de aliados da Otan.

**Sumário candidato da mT5-F**

O parlamento da Turquia aprovou por ampla maioria (507 votos contra 19), um pedido do governo para permitir que tropas do país cruzem a fronteira para combater rebeldes separatistas curdos refugiados na região montanhosa. Washington teme que uma incursão turca se esgotou com as ações dos guerrilheiros da partido dos trabalhadores do Curdistão se posicione em estado alerta e, por sua vez, todas as regiões do oriente médio sejam prejudicadas pela segurança do povo turco. Uma medida fez que os empregos do petróleo disparassem no cenário internacional, levando à casa branca em alertas. Em primeiro-ministro da Turquia pediu

**Sumário candidato da mT5-G (WikiLingua)**

O parlamento da Turquia aprovou por ampla maioria (507 votos contra 19), nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do seu Iraque, à fim de combater rebeldes separatistas curdos refugiados na região montanhosa. O exército turco busque uma solução diplomática para essa situação. A Comissão Européia, por sua vez, exortou que uma incursão turca possa desestabilizar os Estados Unidos e outros países, pois isso não interferirá no mercado de combustível brasileiro.

**Sumário candidato da mT5-H (WikiLingua)**

O parlamento da Turquia aprovou por ampla maioria, nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte da região montanhosa e combater rebeldes separatistas curdos refugiados no Iraque. Por sua vez, Erdogan disse que "a paciência do povo turco se esgotou" com as ações dos guerrilheiros separados do Curdistão, que estão no estado de alerta. Eles foram muito atentados.

**Sumário candidato da mT5-I (XL-Sum)**

O Parlamento da Turquia aprovou por ampla maioria (5507 votos contra 19, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque. A medida fez com que os preços do petróleo disparasse no cenário internacional e, provavelmente, todas as regiões do Oriente Médio estão sendo afetadas pelos Estados Unidos. O primeiro-ministro da Turquia pediu nesta quarta-feira que uma incursão turca possa desestabilizar ainda mais pacífica região.

**Sumário candidato da mT5-J (XL-Sum)**

O parlamento da Turquia aprovou por ampla maioria (501 votos contra 19), nesta quarta-feira, um pedido do governo para permitir que tropas do país cruzem a fronteira para o norte do Iraque e combater rebeldes separatistas curdos refugiados na região montanhosa. Isso fez com que os preços do petróleo disparasse no cenário internacional, mas não interferirá no mercado de combustíveis brasileiro. Em meio à operação militar, que reúne 26 países, entre eles EUA, Grã-bretanha, França, Alemanha.

**Análise dos sumários**

Os resultados são semelhantes aos apresentados para a base TeMário. Os sumários candidatos destes modelos, em comparação com os obtidos a partir do PTT5, possuem uma quantidade maior de erros e com baixa confiabilidade nas informações apresentadas.

## 4.3 Comparação dos experimentos

Considerando todos os experimentos realizados neste trabalho, as Tabelas 20 e 21 apresentam novamente as avaliações de cada modelo nas bases TeMário e CSTNews, respectivamente. No caso dos experimentos com modelos treinados em inglês foram mantidos apenas os melhores avaliados e o T5 (Base).

Os resultados apontam que o ajuste-fino do modelo PTT5, pré-treinado em português, obteve melhores resultados que todos os outros modelos, tanto na base TeMário como na base CSTNews. E de fato, ao se analisar um caso de sumarização por este modelo e por aqueles obtidos a partir do mT5, como feito anteriormente, é possível perceber uma diferença bastante considerável na qualidade dos sumários gerados.

Também é possível perceber que os modelos que sofreram o ajuste-fino nas bases avaliadas quase sempre se sobressaíram aos sumários pelos modelos treinados em inglês sem ajuste-fino, com poucas exceções. De certo modo este resultado era esperado, já que os sistemas em inglês sofrem com os problemas trazidos pela tradução e pelo fato de que, já que eles não foram treinados na base TeMário, eles estão reproduzindo o comportamento (incluindo tipo de escrita, nível de abstração e taxa de compressão) adquirido com suas bases originais. Por outro lado, a escassez de amostras das bases TeMário e CSTNews poderia dificultar muito o ajuste-fino dos modelos, ou conduzi-los a um sobreajuste. Mas a partir da análise dos resultados apresentados anteriormente, embora estes modelos apresentem determinados problemas, eles ainda demonstram resultados superiores aos sistemas em inglês atuando com os textos traduzidos.

Já com relação ao modelo T5 utilizado para a sumarização entre línguas é difícil obter maiores conclusões. De fato, ele performou melhor que os modelos em inglês que não passaram

Tabela 20 – Avaliações dos sumários produzidos para a base TeMário pelos diferentes modelos experimentados neste trabalho.

| Modelo | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| Song et al. (2020) | 41,25 | 9,78 | 21,15 | 66,80 | 16,88 |
| Pegasus (Large) | 32,56 | 10,80 | 19,85 | 66,01 | 13,39 |
| T5 (Base) | 25,94 | 8,57 | 15,98 | 64,62 | 9,94 |
| Fine-tuned T5 (Base) | 43,38 | 14,26 | 24,47 | 69,31 | 19,81 |
| PTT5-A | 43,90 | 19,15 | 27,11 | 69,73 | 21,08 |
| PTT5-B | 48,78 | 20,45 | 29,26 | 70,39 | 23,33 |
| PTT5-C (WikiLingua) | **49,91** | **20,96** | **29,84** | **71,24** | **25,47** |
| PTT5-D (XL-Sum) | 47,95 | 19,32 | 27,81 | 70,72 | 23,36 |
| mT5-A | 33,96 | 13,07 | 20,09 | 67,83 | 17,47 |
| mT5-B (WikiLingua) | 37,81 | 12,96 | 20,66 | 68,50 | 18,46 |
| mT5-C (WikiLingua) | 37,06 | 14,19 | 22,55 | 68,14 | 19,34 |
| mT5-D (XL-Sum) | 32,70 | 11,95 | 19,23 | 67,59 | 16,00 |
| mT5-E (XL-Sum) | 34,48 | 13,22 | 20,57 | 68,22 | 16,91 |

Tabela 21 – Comparação das avaliações dos sumários produzidos pelos sistemas em inglês e os sistemas treinados na base CSTNews.

| Modelo | R1 | R2 | RL | BS | MV |
|---|---|---|---|---|---|
| T5 (Base) | 40,97 | 17,90 | 28,96 | 68,15 | 16,90 |
| T5 (Large) | 41,45 | 18,56 | 30,01 | 69,15 | 18,65 |
| Fine-tuned T5 (Base) | 44,56 | 19,16 | 31,83 | 70,89 | 19,32 |
| PTT5-E | 52,48 | **35,62** | **41,78** | 74,35 | 26,59 |
| PTT5-F | 52,64 | 34,04 | 40,74 | 74,34 | 26,70 |
| PTT5-G (WikiLingua) | **54,15** | 34,53 | 41,01 | 74,86 | **28,16** |
| PTT5-H (XL-Sum) | 52,34 | 33,33 | 38,42 | **75,18** | 27,16 |
| mT5-F | 44,90 | 24,69 | 29,88 | 72,12 | 22,38 |
| mT5-G (WikiLingua) | 44,99 | 25,21 | 31,38 | 72,10 | 22,23 |
| mT5-H (WikiLingua) | 43,70 | 26,27 | 32,54 | 71,87 | 22,18 |
| mT5-I (XL-Sum) | 45,94 | 26,39 | 30,07 | 73,33 | 23,81 |
| mT5-J (XL-Sum) | 47,57 | 29,31 | 36,33 | 74,26 | 25,31 |

pelo processo de ajuste-fino, mas obtiveram resultados inferiores a todos os experimentos com o PTT5. Porém, com relação ao modelo mT5, as diferenças variam. No caso na base TeMário, o modelo T5 obteve resultados superiores à todos os obtidos pelo mT5. Por outro lado, na base CSTNews, houve um comportamento inverso, com o mT5 obtendo resultados superiores em quase todas as avaliações, com diferenças significativas especialmente na medida ROUGE-2.

Nos Apêndices A e B são apresentados um exemplo de texto de cada base de dados, juntamente com seu sumário de referência e o sumário candidato gerado por cada um dos modelos utilizados nos experimentos deste trabalho.

# 5 Conclusões e Trabalhos Futuros

A sumarização automática é uma tarefa muito útil e cada vez mais necessária no atual contexto, com o crescimento exponencial na quantidade de dados no ambiente *Web* junto ao crescente interesse na extração de valores destes dados. Porém, esta não é uma tarefa simples, especialmente no caso da sumarização abstrativa, que de alguma forma deve ser capaz de abstrair o conteúdo, a semântica, de documentos.

Para textos em inglês, já existem diversos trabalhos, tanto de sumarização extrativa quanto abstrativa, que alcançam bons resultados, com o estado da arte avançando cada vez mais rapidamente. Já para textos em português, os trabalhos sobre sumarização abstrativa ainda são escassos, principalmente pela falta de bases anotadas suficientemente grandes para o treinamento de modelos de aprendizagem em profundidade. O foco deste trabalho consiste justamente nos métodos de sumarização abstrativa voltada para a língua portuguesa.

Inicialmente foram realizados experimentos utilizando sistemas treinados com bases de dados em inglês e um sistema de tradução. Não pode-se chegar a muitas conclusões apenas a partir dos resultados destes experimentos, devido a falta de uma base de comparação. Porém, eles serviram como um ponto de partida para a avaliação dos próximos modelos, de forma há se avaliar se treinar um modelo com as bases TeMário e CSTNews, mesmo com seus poucos recursos, leva a resultados melhores que estes primeiros experimentos ou não.

Por fim, foram treinados modelos de aprendizado em profundidade em bases de dados com textos em português, anotados com sumários de referência. O treinamento nas bases maiores, WikiLingua e XL-Sum, obteve resultados satisfatórios. No caso da XL-Sum, em especial, há uma base de comparação a partir do trabalho dos próprios autores do corpus. O modelo treinado neste trabalho não superou os resultados originais, mas se mantiveram próximos.

Já os treinamentos com as bases TeMário e CSTNews, mais tradicionais nos estudos de sumarização automática em português, impunham mais desafios, principalmente devido ao pequeno número de amostras anotadas. Ainda assim, os resultados foram superiores aos obtidos com os experimentos utilizando sistemas treinados em inglês, segundo todas as medidas de avaliação utilizadas. Em especial, o ajuste-fino do modelo PTT5, pré-treinado em português, obteve os melhores resultados, com destaque para os modelos treinados inicialmente com a WikiLingua e XL-Sum, que apresentaram diversas características interessantes para um sumarizador, seja em relação a taxa de abstração ou a capacidade de extrair as principais informações dos textos. Já a sumarização multilíngue e entre línguas, de forma geral, obtiveram resultados inferiores, porém melhores dos que ao se utilizar sistemas em inglês sem a fase de ajuste-fino.

Considerando estes resultados, e analisando-os sob à óptica das perguntas e da hipótese de pesquisas apresentadas na introdução, as principais conclusões obtidas neste trabalho foram:

- Modelos de aprendizado em profundidade são capazes de conduzir a resultados minimamente satisfatórios de sumarização abstrativa, mesmo em bases de dados com poucas amostras anotadas;
- Para uma base de dados com poucos recursos, como a TeMário e a CSTNews, a realização de um primeiro ajuste-fino em outro corpus de sumarização com maior quantidade de dados, seguido pelo ajuste-fino na base em si, tende a trazer melhorias nos resultados qualitativos obtidos;
- A utilização de um modelo pré-treinado apenas em português conduziu a resultados superiores à utilização de modelos multilíngue ou mesmo à sumarização entre línguas.

Os modelos treinados e apresentados neste trabalho ainda apresentam uma série de problemas, principalmente em relação a coerência e a correspondência com os fatos trazidos pelo texto-fonte. Porém, estes são desafios inerentes a sumarização abstrativa, e considerando que estes são os primeiros resultados de sumarização abstrativa nas bases TeMário e CSTNews, utilizando aprendizado em profundidade, segundo o conhecimento dos autores, este trabalho pode vir a ser um ponto de partida para novos estudos. Com isso em mente, a seguir são sugeridas algumas possibilidades para serem exploradas em trabalhos futuros.

Em primeiro lugar, considerando os modelos já apresentados, a realização de uma análise qualitativa mais profunda dos sumários gerados forneceria uma visão mais clara sobre a qualidade dos mesmos, possibilitando apontamentos mais precisos sobre o que ainda precisa ser melhorado.

Considerando a aplicação de aprendizado em profundidade para a tarefa de sumarização, sugere-se também o treinamento de outros modelos além do T5, como aqueles baseados no BERT ou no BART. Em especial, uma grande oportunidade de ultrapassar os presentes resultados é através de aplicações de arquiteturas e técnicas mais específicas para a sumarização, se espelhando nos estudos que se encontram no estado da arte em sumarização abstrativa em inglês. Para a base TeMário e CSTNews, particularmente, poderia-se aplicar técnicas de aprendizado mais voltadas para cenários com pouco recursos, além de se considerar algoritmos de treinamento não-supervisionado ou semi-supervisionado.

Ainda em relação ao tamanho das bases TeMário e CSTNews, levanta-se a possibilidade de aplicações de abordagens mais tradicionais de sumarização abstrativa. Ainda que tais métodos normalmente demandem mais esforço, considerando a limitação destas bases, é possível que eles obtenham um desempenho melhor que os modelos profundos.

# Referências


ALSHAINA, S.; JOHN, A.; NATH, A. G. Multi-document abstractive summarization based on predicate argument structure. In: IEEE. *2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*. Kollam, Índia, 2017. p. 1–6.

ANTONITISCH, A.; FIGUEIRA, A.; AMARAL, D.; FONSECA, E. B.; ABREU, S. C. de; VIEIRA, R. Summ-it++: an enriched version of the summ-it corpus. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION (ELRA). *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)*. Eslovênia, 2016. p. 2047–2051.

BANARESCU, L.; BONIAL, C.; CAI, S.; GEORGESCU, M.; GRIFFITT, K.; HERMJAKOB, U.; KNIGHT, K.; KOEHN, P.; PALMER, M.; SCHNEIDER, N. Abstract meaning representation for sembanking. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Bulgaria, 2013. p. 178–186.

BANERJEE, S.; LAVIE, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Michigan, 2005. p. 65–72.

BANERJEE, S.; MITRA, P.; SUGIYAMA, K. Multi-document abstractive summarization using ilp based multi-sentence compression. In: AAAI. *Twenty-Fourth International Joint Conference on Artificial Intelligence*. Buenos Aires, Argentina, 2015. p. 1208–1214.

BARZILAY, R.; ELHADAD, M. Using lexical chains for text summarization. *Advances in Automatic Text Summarization*, p. 111–121, 1999.

BLACK, W. J.; JOHNSON, F. C. A practical evaluation of two rule-based automatic abstraction techniques. *Expert Systems for Information Management*, v. 1, n. 3, p. 159–177, 1988.

CAI, D.; LAM, W. AMR parsing via graph-sequence iterative inference. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 1290–1301.

CANETE, J.; CHAPERON, G.; FUENTES, R.; HO, J.-H.; KANG, H.; PÉREZ, J. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR 2020*, OpenReview.net, Etiópia, 2020.

CARDOSO, P. C.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. M.; FELIPPO, A. D.; RINO, L. H. M.; NUNES, M. d. G. V.; PARDO, T. A. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: NILC. *Proceedings of the 3rd RST Brazilian Meeting at the 8th Brazilian Symposium in Information and Human Language Technology*. Cuiabá, Mato Grosso, 2011. p. 88–105.

CARDOSO, P. C.; PARDO, T. A. Multi-document summarization using semantic discourse models. *Procesamiento del Lenguaje Natural*, Sociedad Española para el Procesamiento del Lenguaje Natural, n. 56, p. 57–64, 2016.

CARMO, D.; PIAU, M.; CAMPIOTTI, I.; NOGUEIRA, R.; LOTUFO, R. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

CHEN, Y. C.; BANSAL, M. Fast abstractive summarization with reinforce-selected sentence rewriting. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Austrália: Association for Computational Linguistics (ACL), 2018. p. 675–686.

COLLOVINI, S.; CARBONEL, T. I.; FUCHS, J. T.; COELHO, J. C.; RINO, L.; VIEIRA, R. Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In: NILC. *Proceedings of 5th Workshop in Information and Human Language Technology*. Rio de Janeiro, RJ, 2007.

CONDORI, R. E. L.; PARDO, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, v. 78, p. 124–134, 2017. ISSN 0957-4174.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DOHARE, S.; GUPTA, V.; KARNICK, H. Unsupervised semantic abstractive summarization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of ACL 2018, Student Research Workshop*. Melbourne, Australia, 2018. p. 74–83.

DOU, Z.-Y.; LIU, P.; HAYASHI, H.; JIANG, Z.; NEUBIG, G. GSum: A general framework for guided neural abstractive summarization. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021. p. 4830–4842.

EDMUNDSON, H. P. New methods in automatic extracting. *Journal of the ACM*, Association for Computing Machinery, New York, NY, USA, v. 16, n. 2, p. 264–285, 1969. ISSN 0004-5411.

EYAL, M.; BAUMEL, T.; ELHADAD, M. Question answering as an automatic evaluation metric for news article summarization. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 3938–3948.

FABBRI, A.; LI, I.; SHE, T.; LI, S.; RADEV, D. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florença, Itália: Association for Computational Linguistics, 2019. p. 1074–1084.

FABBRI, A. R.; KRYśCIńSKI, W.; MCCANN, B.; XIONG, C.; SOCHER, R.; RADEV, D. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, v. 9, p. 391–409, 04 2021. ISSN 2307-387X.

FEIJÓ, D. de V.; MOREIRA, V. P. RulingBR: A summarization dataset for legal texts. In: *Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language (PROPOR)*. Canela, Rio Grande do Sul: Springer, 2018. p. 255–264.

FELIPPO, A. D. CM2News: Towards a corpus for multilingual multi-document summarization. In: *CORPORA AND TOOLS FOR PROCESSING CORPORA WORKSHOP (CTPC)(PROPOR)*. Tomar, Portugal: Portuguese Language Department of the Directorate-General of Translation of the European Commission, 2016. v. 12.

FILHO, J. A. W.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. The brWaC corpus: A new open resource for Brazilian Portuguese. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

GANESAN, K.; ZHAI, C.; HAN, J. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In: *23rd International Conference on Computational Linguistics*. Pequim, China: Coling, 2010. p. 340–348.

GEHRMANN, S.; DENG, Y.; RUSH, A. Bottom-up abstractive summarization. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Bruxelas, Bélgica: Association for Computational Linguistics, 2018. p. 4098–4109.

GUPTA, S.; GUPTA, S. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, Elsevier, v. 121, p. 49–65, 2019.

HASAN, T.; BHATTACHARJEE, A.; ISLAM, M. S.; MUBASSHIR, K.; LI, Y.-F.; KANG, Y.-B.; RAHMAN, M. S.; SHAHRIYAR, R. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021. p. 4693–4703.

INÁCIO, M. L. *Sumarização de Opinião com base em Abstract Meaning Representation*. Tese (Doutorado) — Universidade de São Paulo, 2021.

JIANG, S.; TU, D.; CHEN, X.; TANG, R.; WANG, W.; WANG, H. *ClueGraphSum: Let Key Clues Guide the Cross-Lingual Abstractive Summarization*. Online: arXiv, 2022.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, MCB UP Limited, 1972.

JONES, K. S. IDF term weighting and ir research lessons. *Journal of Documentation*, Emerald Group Publishing Limited, 2004.

KÅGEBÄCK, M.; MOGREN, O.; TAHMASEBI, N.; DUBHASHI, D. Extractive summarization using continuous vector space models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*. Gotemburgo, Suécia, 2014. p. 31–39.

KINGMA, D. P.; BA, J. *Adam: A Method for Stochastic Optimization*. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1412.6980>.

KRYŚCIŃSKI, W.; PAULUS, R.; XIONG, C.; SOCHER, R. Improving abstraction in text summarization. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Bruxelas, Bélgica: Association for Computational Linguistics, 2018. p. 1808–1817.

KUPIEC, J.; PEDERSEN, J.; CHEN, F. A trainable document summarizer. In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. Seattle, Washington: Association for Computing Machinery, 1995. p. 68–73.

KURISINKEL, L. J.; ZHANG, Y.; VARMA, V. Abstractive multi-document summarization by partial tree extraction, recombination and linearization. In: ASIAN FEDERATION OF NATURAL LANGUAGE PROCESSING. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan, 2017. p. 812–821.

KUSNER, M.; SUN, Y.; KOLKIN, N.; WEINBERGER, K. From word embeddings to document distances. In: BACH, F.; BLEI, D. (Ed.). *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France: PMLR, 2015. (Proceedings of Machine Learning Research, v. 37), p. 957–966.

LADHAK, F.; DURMUS, E.; CARDIE, C.; MCKEOWN, K. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 4034–4048.

LEIXO, P.; PARDO, T. A. S. et al. CSTNews: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory). São Carlos, SP, Brasil., 2008.

LIANG, Y.; MENG, F.; ZHOU, C.; XU, J.; CHEN, Y.; SU, J.; ZHOU, J. *A Variational Hierarchical Model for Neural Cross-Lingual Summarization*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2203.03820>.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. Barcelona, Espanha: Association for Computational Linguistics, 2004. p. 74–81.

LIN, H.; NG, V. Abstractive summarization: A survey of the state of the art. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 33, n. 01, p. 9815–9822, Jul. 2019.

LIU, Y. Fine-tune bert for extractive summarization. *CoRR*, abs/1903.10318, 2019.

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)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3730–3740.

LIU, Y.; LUO, Z.; ZHU, K. Controlling length in abstractive summarization using a convolutional neural network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Bruxelas, Bélgica, 2018. p. 4110–4119.

LOPES, A.; NOGUEIRA, R.; LOTUFO, R.; PEDRINI, H. Lite training strategies for Portuguese-English and English-Portuguese translation. In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, 2020. p. 833–840.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165, 1958.

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Berlin, v. 8, n. 3, p. 243–281, 1988.

MARTSCHAT, S.; MARKERT, K. Improving rouge for timeline summarization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Espanha, 2017. p. 285–290.

MAZIERO, E. G.; UZÊDA, V.; PARDO, T. A. S.; NUNES, M. d. G. V. *TeMário 2006: Estendendo o Córpus TeMário*. São Carlos, São Paulo: Série de Relatórios do NILC. NILC-TR-07-06, 2007.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: BURGES, C.; BOTTOU, L.; WELLING, M.; GHAHRAMANI, Z.; WEINBERGER, K. (Ed.). *Advances in Neural Information Processing Systems*. Lake Tahoe, Nevada: Curran Associates, Inc., 2013. v. 26.

MILLER, D. *Leveraging BERT for Extractive Text Summarization on Lectures*. Online: arXiv, 2019.

MOHAN, M. J.; SUNITHA, C.; GANESH, A.; JAYA, A. A study on ontology based abstractive summarization. *Procedia Computer Science*, Elsevier, v. 87, p. 32–37, 2016.

MOIRANGTHEM, D. S.; LEE, M. Abstractive summarization of long texts by representing multiple compositionalities with temporal hierarchical pointer generator network. *Neural Networks*, Elsevier, v. 124, p. 1–11, 2020.

MORATANCH, N.; CHITRAKALA, S. A survey on extractive text summarization. In: *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. Chennai, Índia: IEEE, 2017. p. 1–6.

MORENO, J. M. T. *Automatic text summarization*. Nova Jersey: John Wiley & Sons, 2014.

NARAYAN, S.; COHEN, S. B.; LAPATA, M. 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*. Bruxelas, Bélgica: Association for Computational Linguistics, 2018. p. 1797–1807.

NENKOVA, A.; MCKEOWN, K. *Automatic summarization*. Boston: Now Publishers Inc, 2011.

NENKOVA, A.; PASSONNEAU, R.; MCKEOWN, K. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, Association for Computing Machinery, Nova Iorque, v. 4, n. 2, 2007. ISSN 1550-4875.

NETO, J. L.; SANTOS, A. D.; KAESTNER, C. A.; FREITAS, A. A. Generating text summaries through the relative importance of topics. In: *Advances in Artificial Intelligence*. Berlim, Heidelberg: Springer, 2000. p. 300–309.

NGUYEN, T. T.; LUU, A. T. *Improving Neural Cross-Lingual Abstractive Summarization via Employing Optimal Transport Distance for Knowledge Distillation*. 2022. 11103-11111 p.

NÓBREGA, F. A. A.; PARDO, T. A. S. Investigating machine learning approaches for sentence compression in different application contexts for portuguese. In: SILVA, J.; RIBEIRO, R.; QUARESMA, P.; ADAMI, A.; BRANCO, A. (Ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2016. p. 245–250. ISBN 978-3-319-41552-9.

OUYANG, J.; SONG, B.; MCKEOWN, K. A robust abstractive system for cross-lingual summarization. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 2025–2031.

PARDO, T. A. Sentence compression for portuguese. In: SPRINGER NATURE. *Computational Processing of the Portuguese Language: 14th International Conference (PROPOR 2020)*. Evora, Portugal, 2020.

PARDO, T. A. S. *GistSumm-GIST SUMMarizer: Extensões e novas funcionalidades*. São Carlos, São Paulo: Série de Relatórios do NILC. NILC-TR-05-05, 2005.

PARDO, T. A. S.; RINO, L. H. M. *TeMário: Um corpus para sumarização automática de textos*. São Carlos, São Paulo: Série de Relatórios do NILC. NILC-TR-03-09, 2003.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. d. G. V. Gistsumm: A summarization tool based on a new extractive method. In: SPRINGER. *International Workshop on Computational Processing of the Portuguese Language (PROPOR 2003)*. Faro, Portugal, 2003. p. 210–218.

PASSONNEAU, R. J.; CHEN, E.; GUO, W.; PERIN, D. Automated pyramid scoring of summaries using distributional semantics. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgária, 2013. p. 143–147.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Catar, 2014. p. 1532–1543.

PONTES, E. L.; HUET, S.; TORRES-MORENO, J.-M.; LINHARES, A. C. Cross-language text summarization using sentence and multi-sentence compression. In: SILBERZTEIN, M.; ATIGUI, F.; KORNYSHOVA, E.; MÉTAIS, E.; MEZIANE, F. (Ed.). *Natural Language Processing and Information Systems*. Cham: Springer International Publishing, 2018. p. 467–479. ISBN 978-3-319-91947-8.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2020.

RINO, L. H. M.; MÓDOLO, M. Supor: An environment for as of texts in brazilian portuguese. In: SPRINGER. *International Conference on Natural Language Processing*. Espanha, 2004. p. 419–430.

ROCHA, V. J. C. *PragmaSUM: novos métodos na utilização de palavras-chave na sumarização automática*. Dissertação (Mestrado) — Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, Minas Gerais, 2017.

RUSH, A. M.; CHOPRA, S.; WESTON, J. A neural attention model for abstractive sentence summarization. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisboa, Portugal: Association for Computational Linguistics, 2015. p. 379–389.

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic text structuring and summarization. *Information Processing & Management*, Elsevier, v. 33, n. 2, p. 193–207, 1997.

SCIALOM, T.; LAMPRIER, S.; PIWOWARSKI, B.; STAIANO, J. Answers unite! unsupervised metrics for reinforced summarization models. 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)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3246–3256. Disponível em: <https://www.aclweb.org/anthology/D19-1320>.

SEE, A.; LIU, P. J.; MANNING, C. D. Get to the point: Summarization with pointer-generator networks. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 1073–1083.

SENO, E. R. M.; NUNES, M. d. G. V. Fusão automática de sentenças similares em português. *Anais do VII Simpósio Brasileiro em Tecnologia da Informação e da Linguagem Humana–STIL*, 2009.

SODRÉ, L. C.; OLIVEIRA, H. T. A. Evaluating regression algorithms for automatic text summarization in brazilian portuguese. In: SBC. *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. São Paulo, 2018. p. 634–645.

SONG, K.; WANG, B.; FENG, Z.; LIU, R.; LIU, F. Controlling the amount of verbatim copying in abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, n. 05, p. 8902–8909, Apr. 2020.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese Named Entity Recognition using BERT-CRF. *arXiv e-prints*, p. arXiv:1909.10649, 2019.

STEINBERGER, J.; JEŽEK, K. Evaluation measures for text summarization. *Computing and Informatics*, v. 28, n. 2, p. 251–275, 2012.

VADAPALLI, R.; KURISINKEL, L. J.; GUPTA, M.; VARMA, V. Ssas: semantic similarity for abstractive summarization. In: ASIAN FEDERATION OF NATURAL LANGUAGE PROCESSING. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan, 2017. p. 198–203.

VEDANTAM, R.; ZITNICK, C. L.; PARIKH, D. Cider: Consensus-based image description evaluation. In: IEEE. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. San Juan, Porto Rico, 2015. p. 4566–4575.

VIRTANEN, A.; KANERVA, J.; ILO, R.; LUOMA, J.; LUOTOLAHTI, J.; Salakoski, T.; Ginter, F.; Pyysalo, S. *Multilingual is not enough: BERT for Finnish*. 2019. arXiv:1912.07076 p.

WAN, X.; LI, H.; XIAO, J. Cross-language document summarization based on machine translation quality prediction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Suécia, 2010. p. 917–926.

WAN, X.; LUO, F.; SUN, X.; HUANG, S.; YAO, J.-g. Cross-language document summarization via extraction and ranking of multiple summaries. In: *Knowledge and Information Systems*. Berlim: Springer, 2019. p. 481–499.

WANG, J.; MENG, F.; LU, Z.; ZHENG, D.; LI, Z.; QU, J.; ZHOU, J. *ClidSum: A Benchmark Dataset for Cross-Lingual Dialogue Summarization*. Online: arXiv, 2022.

WANG, J.; MENG, F.; ZHENG, D.; LIANG, Y.; LI, Z.; QU, J.; ZHOU, J. *A Survey on Cross-Lingual Summarization*. [S.l.]: arXiv, 2022.

WEBER, N.; SHEKHAR, L.; BALASUBRAMANIAN, N.; CHO, K. Controlling Decoding for More Abstractive Summaries with Copy-Based Networks. *arXiv e-prints*, p. arXiv:1803.07038, 2018.

XUE, L.; CONSTANT, N.; ROBERTS, A.; KALE, M.; AL-RFOU, R.; SIDDHANT, A.; BARUA, A.; RAFFEL, C. *mT5: A massively multilingual pre-trained text-to-text transformer*. Online: arXiv, 2020.

YANG, Q.; PASSONNEAU, R.; MELO, G. D. Peak: Pyramid evaluation via automated knowledge extraction. In: AAAI. *Proceedings of the AAAI Conference on Artificial Intelligence*. Phoenix, Arizona, 2016. v. 30, n. 1.

YANG, Z.; ZHU, C.; GMYR, R.; ZENG, M.; HUANG, X.; DARVE, E. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 1865–1874.

ZHANG, J.; ZHAO, Y.; SALEH, M.; LIU, P. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In: III, H. D.; SINGH, A. (Ed.). *Proceedings of the 37th International Conference on Machine Learning*. Online: PMLR, 2020. (Proceedings of Machine Learning Research, v. 119), p. 11328–11339.

ZHANG, J.; ZHOU, Y.; ZONG, C. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 24, n. 10, p. 1842–1853, 2016.

ZHANG, T.; KISHORE, V.; WU, F.; WEINBERGER, K. Q.; ARTZI, Y. Bertscore: Evaluating text generation with bert. In: *International Conference on Learning Representations*. Nova Orleans: OpenReview.net, 2019.

ZHAO, W.; PEYRARD, M.; LIU, F.; GAO, Y.; MEYER, C. M.; EGER, S. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 563–578.

ZHONG, M.; LIU, P.; CHEN, Y.; WANG, D.; QIU, X.; HUANG, X. Extractive summarization as text matching. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 6197–6208.

ZHOU, J.; RUSH, A. Simple unsupervised summarization by contextual matching. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florença, Itália: Association for Computational Linguistics, 2019. p. 5101–5106.

# Apêndices

# APÊNDICE A – Sumários gerados a partir de um texto do corpus TeMário

## A.1 Texto fonte

PARIS - Está em ponto de ebulição na França a polêmica provocada pelo apoio do Abade Pierre, considerado o homem mais popular do país, a seu velho amigo o filósofo Roger Garaudy - ex-comunista e ex-católico convertido ao islamismo - em torno das teses de revisão do Holocausto que este sustenta em seu livro Os Mitos fundadores da política israelense. A simpatia hipotecada pelo sacerdote ao polemista - e não necessariamente a todo o conteúdo de suas teses - repercutiu mal no momento em que Garaudy é processado com base na lei que pune a negação de crimes contra a humanidade, e semeou a confusão na comunidade judaica e na opinião pública.

Ânimos - O tema não podia ser mais delicado, e os ânimos, mais facilmente exaltáveis. Militante islâmico, Garaudy considera no livro que genocídio e holocausto são palavras exageradas para os "pogroms" nazistas; propõe "uma história crítica dos crimes hitleristas"; e assume posição combatente contra "o dogma dos seis milhões de judeus exterminados", que segundo ele é usado para justificar os excessos da política de Israel na Palestina e para deixar o Estado judeu "acima das leis internacionais". Garaudy apresenta os "historiadores críticos" - ou negacionistas - como pesquisadores perseguidos cujos trabalhos não foram contestados cientificamente. E se aventura numa tentativa de relativizar o horror: segundo ele, gente de outros povos foi morta também pelos nazistas e nem todos os judeus morreram em câmaras de gás, mas também de fome, em marcha forçada ou a bala.

A questão poderia ser: aonde semelhante linha de pensamento levaria? Mas o fato é que, relativização, revisão ou negação, o que escreve Garaudy é passível, na França, das penas da lei Gayssot, que no seu caso foi invocada pelo Movimento Contra o Racismo e pela Amizade entre os Povos. O pensador foi indiciado e poderá passar um ano na cadeia.

Onde a porca torceu o rabo, no entanto, foi na intervenção do Abade Pierre, o homem que espalhou pelo mundo as Comunidades de Emaús, que ensinou os franceses a pensarem mais nos pobres, que renunciou à fortuna pessoal, ajudou judeus a escaparem para a Suíça durante a Segunda Guerra e é membro da Liga Contra o Racismo e o Anti-semitismo (Licra).

Calúnia - Também simpatizante da causa palestina, o sacerdote de 83 anos foi solicitado por Garaudy a sair em sua defesa, como outros amigos. Começou por considerar "uma calúnia confundir teu livro com as teses revisionistas". Mais adiante, reconheceu não ter lido o livro, mas um resumo, insistindo porém em argumentos como o do número de mortos no campo

de concentração de Auschwitz, onde se afirmou inicialmente que houve 4 milhões de vítimas, número corrigido posteriormente para 1 milhão; mesmo considerando que "a abominação é a mesma", o abade sustenta haver aí uma demonstração de que o tema deve ser objeto de investigação imparcial.

O Abade Pierre foi chamado a explicações na Liga contra o Racismo. Recuou, disse que não entra no mérito do livro nem apóia suas teses, mas repisou o argumento da integridade intelectual de Garaudy e da necessidade de debater livremente este tema.

Os jornais estão cheios de contestações, respostas e desafios ao abade. E ontem, enquanto a hierarquia católica mantém cauteloso silêncio, ele voltou à carga em entrevista ao Libération, dizendo-se satisfeito com a polêmica: "Muita gente me tem dito obrigado pela coragem de questionar um tabu. É preciso parar de chamar de anti-semita quem questiona a história do Holocausto. Não nos deixaremos mais chamar de antijudeus ou anti-semitas por dizer que um judeu canta mal."

## A.2 Sumário de referência

Está quente na França a polêmica provocada pelo apoio do Abade Pierre ao filósofo Roger Garaudy, convertido ao islamismo, que , no seu livro Os mitos fundadores da política israelense, propõe uma revisão do holocausto judeu , considerado por ele uma forma de justificar os excessos israelenses contra os palestinos.

O filósofo vê exagero na atribuição a Hitler dos crimes contra os judeus , quer na quantidade quer na qualidade. Mesmo não inocentando o ditador, ele é passível de infringir a lei Gayssot, invocada pelo Movimento Contra o Racismo e pela Amizade entre os Povos para indiciá-lo.

A questão se acende mais, porque o Abade, de 83 anos, é conhecido pelo seu desapego aos bens, por ajudar judeus a escapar, por ser membro da Liga Contra o Racismo e o Anti-Semitismo . Mas também é simpatizante da causa palestina , e, por isso, Garaudy pediu que o defendesse. E na defesa, mesmo dizendo só ter lido um resumo do livro , insiste em argumentos que põem em cheque a quantidade de judeus mortos.

Chamado a dar explicações na Liga contra o Racismo, o abade recuou dizendo que não entrava no mérito do livro e nem apoiava as suas teses, mas reforçou sua convicção na integridade intelectual de Garaudy e a necessidade de um debate despreconcebido sobre o tema.

## A.3 Sumários candidatos

### A.3.1 GistSumm

Mais adiante, reconheceu não ter lido o livro, mas um resumo, insistindo porém em argumentos como o do número de mortos no campo de concentração de Auschwitz, onde se afirmou inicialmente que houve 4 milhões de vítimas, número corrigido posteriormente para 1 milhão; mesmo considerando que "a abominação é a mesma", o abade sustenta haver aí uma demonstração de que o tema deve ser objeto de investigação imparcial.

### A.3.2 (MILLER, 2019)

PARIS - Está em ponto de ebulição na França a polêmica provocada pelo apoio do Abade Pierre, considerado o homem mais popular do país, a seu velho amigo o filósofo Roger Garaudy - ex-comunista e ex-católico convertido ao islamismo - em torno das teses de revisão do Holocausto que este sustenta em seu livro Os Mitos fundadores da política israelense. Ânimos - o tema não podia ser mais delicado, e os ânimos, mais facilmente exaltáveis. Militante islâmico, Garaudy considera no livro que genocídio e holocausto são palavras exageradas para os "pogroms" nazistas; propõe "uma história crítica dos crimes hitleristas"; e assume posição combatente contra "o dogma dos seis milhões de judeus exterminados", que segundo ele é usado para justificar os excessos da política de Israel na Palestina e para deixar o Estado judeu "acima das leis internacionais". a questão poderia ser: aonde semelhante linha de pensamento levaria?

### A.3.3 (LIU; LAPATA, 2019)

PARIS - Está em ponto de ebulição na França a polêmica provocada pelo apoio do Abade Pierre, considerado o homem mais popular do país, a seu velho amigo o filósofo Roger Garaudy - ex-comunista e ex-católico convertido ao islamismo - em torno das teses de revisão do Holocausto que este sustenta em seu livro Os Mitos fundadores da política israelense. A simpatia hipotecada pelo sacerdote ao polemista - e não necessariamente a todo o conteúdo de suas teses - repercutiu mal no momento em que Garaudy é processado com base na lei que pune a negação de crimes contra a humanidade, e semeou a confusão na comunidade judaica e na opinião pública. Ânimos - O tema não podia ser mais delicado, e os ânimos, mais facilmente exaltáveis.

### A.3.4 (SONG et al., 2020)

O apoio do filósofo francês no momento da ebulição na França. O padre anti-sestimista. A arte da agitação. Um livro sobre o holocausto. Garaudy apresenta "historiadores críticos". Os sobreviventes do Holocausto contam o horror. O que é a linha de pensamento. Garaudy's

‘ Gayssot. O pensador é indiciado. O abade Pierre, que ensinou francês a pensar mais sobre os pobres. Priest acusava matar soldados israeli. O revisor. As observações do papa sobre auschwitz. Liga africana contra o racismo. O livro de Garaudy sobre a integridade intelectual. O abade. O texto de Pope comenta o tabu. Pare de chamar antissêmicos que questionam o Holocausto. Antissêmicos já não são chamados anti- judeus.

### A.3.5 PEGASUS (Large)

Paris - a controvérsia provocada pelo apoio de abade Pierre, considerado o homem mais popular na França, está no ponto de ebulição na França, ao seu antigo amigo o filósofo Roger Garaudy - ex-comunista e ex-católico convertido ao islamismo - em torno da tese de revisão do holocausto que ele sustenta em seu livro os mitos que fundam a política israeli.

### A.3.6 PEGASUS (MultiNews)

– Um padre francês popular foi obrigado a retirar seu apoio para um filósofo que argumenta em seu livro que o Holocausto era um hoax perpetrado pelos nazistas. A controvérsia começou quando Roger Garaudy, um ex-comunista que se converteu ao islamismo, publicou um livro chamado mitos foundingisraeli policy. Argumenta que o Holocausto era um tiro-furo perpetrado pelos nazistas para justificar suas políticas na palestina, e que nem todos os 6 milhões de judeus que foram exterminados morreram em câmaras de gás, fome ou em mortes por balastro, relata Haaretz. A garçonaria de 84 anos foi responsável pela negação de crimes contra a humanidade por lei que torna a negação de crimes contra a humanidade um crime puntável até um ano na prisão. O padre em questão, 83 anos de calunaia, foi orientado por Garaudy a defender seu amigo. Ele inicialmente considerou “uma calúnia para confundir seu livro com as teses revisionistas”, mas acabou concordando em fazê-lo. Agora, a Igreja Católica em França pediu a calunaia que retirasse o apoio para a garaudia. “É necessário deixar de chamar antissêmicos que questionam a história do holocaus. não mais iremos chamar de anti-jewish ou anti-semitics

### A.3.7 PEGASUS (Newsroom)

Sobre a Fé é uma conversação inovadora e provocativa sobre todos os aspectos da religião com o melhor autor vendedor Jon Meacham da semana de semana e Sally Quinn do Posto de Washington. Manter actualizado sobre os desenvolvimentos religiosos globais com a fé.

### A.3.8 PEGASUS (XSum)

O caso de Roger Garaudy, um ex-comunista convertido em islamismo, provocou um tumulto na comunidade judaica em França.

### A.3.9 T5 (Base)

Roger Garaudy é um ex-comunista e ex-católico convertido em islamismo. Garaudy sustenta a tese de revisão do holocausto em seu livro os mitos que fundam a política de israeli. Garaudy: genocídio e holocausto são palavras exageradas para os "pogroms" nazistas, ele propõe "uma história crítica dos crimes hitleristas" o livro de Garaudy está sujeito às frases dos gays.

### A.3.10 T5 (Large)

Roger Garaudy é processado com base no direito que puni a negação dos crimes contra a humanidade. Garaudy, um ativo islâmico, considera em seu livro que o genocídio e o holocausto são palavras exageradas para os "pogroms" nazistas o pensador foi indiciado e pode passar um ano na cadeia. Mas, quando a porca torta a espada, ela foi na intervenção do abade Pierre, que disseminou a espada,

### A.3.11 Fine-tuned T5 (Base)

Paris - a controvérsia provocada pelo apoio de abade Pierre, considerado o homem mais popular na frança, está no ponto de ebulição, para seu antigo amigo o filósofo Roger Garaudy, um ex-comunista e católico convertido no islamismo, em torno da tese de revisão do holocausto que ele sustenta em seu livro os mitos que fundam a política de israeli. A simpatia que o padre tinha em relação ao polemista procurou a confusão na comunidade judaica e na opinião pública. Agitação – o tema não poderia ser mais delicado, e é de fácil expiração. O muslim active considera que o genocídio é palavras exageradas para os "pogromas nazistas"; propõe "uma história crítica dos crimes hitleristas"; e toma posição de combate ao "dogma de seis milhões de exterminados, que foi usado para justificar os excessos da política de Israel na palestina e deixar o Estado acima das leis internacionais.

### A.3.12 PTT5-A

Paris - está em ponto de ebulição na França a polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, a seu velho amigo, o filósofo Roger Garaudy. A simpatia de seu amigo pelo polemista, que sustenta em seu livro os mitos fundadores da política israelense. O apoio a sua tese de revisão do holocausto, em torno das teses de relativização do genocídio e da negação de crimes contra a humanidade, gerou uma grande confusão na comunidade judaica e na opinião pública. Ânimos - o tema não podia ser mais delicado, e os descontentamentos, mais facilmente exaltáveis. Militante islâmico, ele considera que os "pogroms" nazistas são palavras exageradas para justificar a política de Israel na Palestina e para deixar o Estado judeu "acima das leis internacionais". ele apresenta os "historiadores críticos" - ou negacionista

### A.3.13 PTT5-B

Paris - está em ponto de ebulição na França a polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, a seu velho amigo, o filósofo Roger Garaudy. A simpatia de seu amigo pelo polemista, que sustenta em seu livro os mitos fundadores da política israelense. O apoio a sua tese de revisão do holocausto, em torno das teses de relativização do genocídio e da negação de crimes contra a humanidade, gerou uma grande confusão na comunidade judaica e na opinião pública. Ânimos - o tema não podia ser mais delicado, e os descontentamentos, mais facilmente exaltáveis. Militante islâmico, ele considera que os "pogroms" nazistas são palavras exageradas para justificar a política de Israel na Palestina e para deixar o Estado judeu "acima das leis internacionais". ele apresenta os "historiadores críticos" - ou negacionista- como pesquisadores perseguidos pela lei Gayssot. e se aventura numa tentativa de rever o horror: segundo ele, gente de outros povos foi morta pelos nazistas e não só por fome, mas também de fome e em marcha forçada ou a bala. A questão poderia ser: aonde semelhante linha de pensamento levaria? Mas o fato é que, releva, revisão ou negação do que escreve gar

### A.3.14 PTT5-C (WikiLingua)

A polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, a seu velho amigo o filósofo Roger Garaudy. A simpatia hipotecada pelo sacerdote ao polemista – não necessariamente a todo o conteúdo de suas teses - repercutiu mal no momento em que ele é processado com base na lei que pune a negação de crimes contra a humanidade - e semeou a confusão na comunidade judaica e na opinião pública.. Militante islâmico, ele considera no livro que genocídio e holocausto são palavras exageradas para os nazistas; propõe uma história crítica dos crimes hitleristas"; e assume posição combatente contra o dogma dos seis milhões de judeus exterminados, que segundo ele, é justificado pelos excessos da política de Israel na Palestina e para deixar o Estado judeu acima das leis internacionais. E que se aventura numa tentativa de relativizar o horror: como gente de outros povos foi morta também pelos nazistas e os judeus morreram em câmaras de gás, mas também de fome, em marcha forçada ou a bala. Mas a questão não é passível de ebulição na frança: a lei Gayssot, invocada pela luta contra racismo e pela amizade entre os povos. O pensador foi indiciado e poderá passar um ano na cadeia, onde a porca torceu o rabo, no entanto, foi na intervenção o abade Pierro, um homem que espalhou pelo mundo as comunidades de Emaús e ajudou judeus a pensarem mais nos pobres.

### A.3.15 PTT5-D (XL-Sum)

A polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, está em ponto de ebulição na França: a simpatia hipotecada pelo sacerdote ao polemista - e não necessariamente a todo o conteúdo de suas teses de revisão do holocausto -

repercutiu mal no momento em que Garaudy é processado com base na lei que pune a negação de crimes contra a humanidade, e semeou a confusão na comunidade judaica e na opinião pública. Ânimos: o genocídio é palavras exageradas para os nazistas; propõe uma história crítica dos crimes hitleristas; e assume posição de combatente contra o dogma dos seis milhões de judeus exterminados. E propõe relativizar o horror: segundo ele, gente de outros povos foi morta também pelos nazistas e nem todos morreram em câmara de gás, em marcha forçada ou a bala. Mas o fato é que a revisão, revisão ou negação, é passível, no seu caso, da lei Gayssot. O pensador foi indiciado e poderá passar um ano na cadeia, onde a porca torceu o rabo.

### A.3.16 mT5-A

A polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, ao filósofo Roger Garaudy - ex-comunista e islâmico, em torno das teses de revisão do holocausto, que é usado para justificar os excessos da política de Israel na Palestina. A simpatia hipotecada pelos sacerdotes ào polemista com base na lei que punha negação de crimes contra humanidade, não parece ser mais delicado. Ainda, se aventura numa tentativa de relativizar isso: segundo ele, gente de outros povos foi morta também pelo nazistas, mas não todos morreram em câmaras de fome, na marcha forçada ou no carro forçado. O fato não significa que, no momento atual, existem outras palavras exagadas para alguns pogroms. ele assume posição combatente contra o dogma dos seis milhões de judeus exterminados.

### A.3.17 mT5-B (WikiLingua)

O militante islâmico, Garaudy considera no livro que Genocídio e holocausto são palavras exageradas para os pogroms dos judeus exterminados; propõe "uma história crítica dos crimes hitleristas"; assume posição combatente contra "o dogma dos seis milhões de Judeus Exterminadois", que segundo ele é justificar a confusão na comunidade judaica. A simpatia hipotecada pelo sacerdote do polêmica provocada por apoio do abade-Pierre, em torno das teses de revisão do HHolocausto que sustenta sua política israelense. Não pode ser mais delicado o tema, mas não podia ser exaltável. E seus ânimos não podem estar exaltaráveis. Ele também considera que, com base na lei que puniu as negações contra humanidade, ele temeu mal no momento em que ele processado. Apesar de não ter feito nenhum esforço para relativizar ou negação do horror: ele, gente de outros povos foi morta pelos nazistas; não de fome; mas às vezes, não foram contestados cientificamente. Se aventura numa tentativa de relativização, referência feita por históricos ou negacionistas - como pesquisadores perseguidos. O que aconteceria? Até agora?

### A.3.18 mT5-C (WikiLingua)

A polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, e do filósofo Roger Garaudy, em torno das teses de revisão do holocausto que este

sustenta em seu livro os mitos fundadores da política israelense. A simpatia hipotecada pelos sacerdotes ào polemista repercutiu mal no momento em que ele processa com base na lei que punha negação de crimes contra humanidade, sem que foram contestados cientificamente. Militante islâmico, se aventura numa tentativa de relativização do horror: segundo ele, gente de outros povos foi morta por "pogroms" nazistas; propõe uma história combatente contra "o dogma dos seis milhões de judeus exterminatos"; assume posição contra seus excessos de política de Israel na Palestina, para justificar isso. Ele afirma que genocídio é palavras exageradas para "polocaustos hitleristas", um historiador crítico - ou negacionista- como pesquisadores perseguidos científicomente.

### A.3.19 mT5-D (XL-Sum)

A polêmica provocada pelo apoio do abade Pierre, em torno das teses de revisão do holocausto, o filósofo Roger Garaudy - ex-comunista e convertido ào islamismo- é um amigo muito popular do país. O livro  Genocídio, Holocausto são palavras exageradas para os pogroms nazistas; assume posição combatente contra "o dogma dos seis milhões de judeus exterminados"; propõe "uma história de crítica dos crimes hitleristas". Segundo ele, não podia ser mais delicado, mas ainda mais exaltáveis. Militante islâmico, Garoudy, que também insiste em justificar esses excessos da política de Israel na Palestina, para que não sejam contestados científicomente.

### A.3.20 mT5-E (XL-Sum)

A polêmica provocada pelo apoio do abade Pierre, considerado o homem mais popular do país, ao filósofo Roger Garaudy, em torno das teses de revisão do holocausto que este sustenta em seu livro os mitos fundadores da política israelense, repercutiu mal no momento em que ele é processado por base na lei que permite negação de crimes contra as humanidades. Não se sabe quanto à sua simpatia pelos seus assessores e não apenas no conteúdo de suas Teses. O ódio da comunidade islâmica na opinião pública vem provocando confusão na população judaica. A crítica feita pelo velho amigo do líder norte-coreano, ele argumenta que as palavras exageradas são usadas para justificar as excessos na política de Israel na Palestina; propõe uma história dos seis milhões de judeus exterminados.

# APÊNDICE B – Sumários gerados a partir de um texto do corpus CSTNews

## B.1 Texto fonte

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). Braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES e comandou a Secretaria de Previdência Complementar no governo Fernando Henrique Cardoso.

- A Solange vai ser a nova presidente da Anac - disse Jobim, em jantar que celebrou os 50 anos da Rede RBS em Brasília.

Alvo de críticas incisivas da oposição desde o acidente com o Airbus da TAM, o atual presidente da Anac, Milton Zuanazzi, já teria concordado em renunciar e deve entregar o cargo nos próximos dias. Porém, procurado pela GloboNews TV na noite desta terça-feira, ele disse que não pretende deixar a função.

Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa. Os dois ministros que davam sustentação a Zuanazzi no cargo, Dilma Rousseff e Walfrido dos Mares Guia, avaliam nesta quarta-feira que ele se tornou uma figura muito vinculada à crise aérea e passaram a defender sua substituição.

O último remanescente da diretoria da Anac à época do acidente, Josef Barat, diretor de Relações Internacionais, Pesquisas e Capacitação da Anac, também deve sair. Outros três diretores já entregaram os cargos.

Em 2001, por cerca de oito meses, Solange comandou com mão-de-ferro os fundos de pensão do país, como titular da Secretaria de Previdência Complementar (SPC) do Ministério da Previdência. Após vários desentendimentos com o então ministro da pasta, Roberto Brant, foi demitida. Durante aquele período, comandou a maior reforma dos últimos anos na legislação que regulamenta os fundos de pensão, aumentando a transparência do setor. Mas também conquistou muito inimigos, principalmente entre os dirigentes de fundos de pensão, que a acusavam de ser muito vaidosa e de não ter cuidado ao divulgar informações sobre o sistema.

O relatório final da CPI do Apagão da Câmara, que começou a ser lido nesta terça-feira, recomenda o ingresso da iniciativa privada na administração da infra-estrutura dos aeroportos, hoje sob o comando de uma estatal, a Infraero. O relator, deputado Marco Maia (PT-RS), só

deve apontar conclusões nesta quinta-feira.

Cobrado por familiares de vítimas do acidente da TAM e pela oposição, dificilmente o relator irá sugerir o indiciamento de Zuanazzi. Ele disse que não está convencido da "participação objetiva" de Zuanazzi nas denúncias contra a agência:

- Não podemos indiciar para agradar à oposição, ao governo ou a quem quer que seja - disse Maia.

## B.2 Sumário de referência

O ministro da Defesa, Nelson Jobim, informou que a economista Solange Vieira será a nova presidente da Agência Nacional de Aviação Civil. Ela substitui o atual presidente, Milton Zuanazzi, porque ele está renunciando devido a críticas incisivas da oposição desde o acidente com o Airbus da TAM. Os dois ministros que davam sustentação a Zuanazzi no cargo, Dilma Rousseff e Walfrido dos Mares Guia, avaliam que ele se tornou uma figura muito vinculada à crise aérea e passaram a defender sua substituição. Solange foi escolhida devido à sua experiência quando comandou os fundos de pensão do país, promovendo uma expressiva reforma em sua legislação e aumentando a transparência do setor. O acidente da TAM ainda levou a uma análise CPI do Apagão da Câmara sobre a administração da infra-estrutura dos aeroportos.

## B.3 Sumários candidatos

### B.3.1 GistSumm

Os dois ministros que davam sustentação a Zuanazzi no cargo, Dilma Rousseff e Walfrido dos Mares Guia, avaliam nesta quarta-feira que ele se tornou uma figura muito vinculada à crise aérea e passaram a defender sua substituição. O último remanescente da diretoria da Anac à época do acidente, Josef Barat, diretor de Relações Internacionais, Pesquisas e Capacitação da Anac, também deve sair. Durante aquele período, comandou a maior reforma dos últimos anos na legislação que regulamenta os fundos de pensão, aumentando a transparência do setor. - Não podemos indiciar para agradar à oposição, ao governo ou a quem quer que seja - disse Maia.

### B.3.2 (MILLER, 2019)

O ministro da defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). em 2001, por cerca de oito meses, Solange comandou com mão-de-ferro os fundos de pensão do país, como titular da Secretaria de Previdência Complementar (SPC)

do Ministério da Previdência. Após vários desentendimentos com o então ministro da pasta, Roberto Brant, foi demitida. Mas também conquistou muito inimigos, principalmente entre os dirigentes de fundos de pensão, que a acusavam de ser muito vaidosa e de não ter cuidado ao divulgar informações sobre o sistema.

### B.3.3 (LIU; LAPATA, 2019)

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). Braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES e comandou a Secretaria de Previdência Complementar no governo Fernando Henrique Cardoso. Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa.

### B.3.4 (SONG et al., 2020)

Economist Solange vieira to head angola's Civil Aviation Agency. O chefe da Previdência Social assume o cargo de. Novo presidente do brasil - Jobim. Anac presidente de mandato. Para a GloboNews, ele não está pronto para ir. Os juízes não podem sair da corte. Os dois ministérios do Brasil defendem Zuanazzi. A última diretoria da ANAC. Três outros diretores já assumiram suas funções. Solange lidera os fundos de pensão da china. Ministro de massa italiano despediu. Jiang zemin sobre reforma previdenciária. O chefe do fundo de pensão enfrenta críticas. Os aeroportos do Brasil suscitaram a melhoria da infraestrutura. Decisão da equipa de rugby francesa na quinta-feira. Famílias de vítimas de acidentes de IAM. U. s. oficial diz não estar convencido da "participação objetiva"de Zuanazzi. Por que queremos agradar à oposição.

### B.3.5 PEGASUS (Large)

O ministro da Defesa, Nelson Jobim, informou no final da terça-feira que o economista Solange Vieira, 38 anos, será o novo presidente da Agência Nacional de Aviação Civil (Anac). Alvo de críticas incisivas pela oposição desde o acidente com o avião de TAM, o atual presidente da Anac, Milton Zuanazzi, teria concordado em resignar e devia entregar o cargo nos próximos dias.

### B.3.6 PEGASUS (MultiNews)

– O governo brasileiro nomeou o novo presidente da Agência de Aviação Civil do país depois de o chefe da agência ter descer na sequência do acidente de avião que matou 154 pessoas. O ministro da Defesa Nelson Jobim nomeou economista Solange Vieira de Souza como o novo presidente da Agência de Aviação Civil, ou Anac, Agência Nacional de Aviação Civil do país, relatórios da BBC. Vieira de Souza, 38 anos, esteve junto ao governo desde

1999 e era anteriormente o chefe dos fundos de pensão do país. A demissão do presidente da agência, Milton Zuanazzi, era esperada após o acidente do voo 214, mas mudou de cabeça na noite passada, dizendo que não queria sair da posição, relatórios de celeuma.

### B.3.7 PEGASUS (Newsroom)

Classe P "MsoNormal"estilo "MARGIN: 0 em 0 em 0pt; DIRECÇÃO: ltr; unicode-bidi: embed; TEXT-ALIGN: esquerda; mso-layout-grid-align: nenhum»Ministério da Defesa, nelson

### B.3.8 PEGASUS (XSum)

Ela é o braço direito da peemedebista desde seu escritório, é uma funcionária da carreira do bndes e comandou a secretaria de assistência complementar na administração Fernando Henrique Cardoso. Ela foi o braço direito da peemedebista desde seu escritório, ela é uma

### B.3.9 T5 (Base)

Solange Vieira, 38 anos, será o novo presidente da Agência Nacional de Aviação Civil (Anac), que tem sido o braço direito da peemedebista desde seu escritório. É uma funcionária da BNDES e comandou a secretaria de assistência complementar na administração Fernando Henrique Cardoso. Milton Zuanazzi, atual presidente da Anac, teria concordado em renunciar e devia entregar o cargo nos próximos dias.

### B.3.10 T5 (Large)

Solange Bieira, 38 anos, foi o braço direito da peemedebista desde seu escritório. Ela comandou a secretaria de assistência complementar na administração Fernando Henrique Cardoso. O atual presidente da Anac, Milton Zuanazzi, teria concordado em resignificar. Mas procurou a TV GloboNews à noite de terça-feira, ele disse não querer deixar a função.

### B.3.11 Fine-tuned T5 (Base)

O ministro da Defesa, Nelson Jobim, informou que o economista Solange Vieira, 38 anos, será o novo presidente da Agência Nacional de Aviação Civil (Anac). Como os diretores das agências têm um mandato de cinco anos, podem sair do tribunal ou da decisão judicial ou ser acusados de improbidade administrativa. Os dois ministérios que sustentaram o Zuanazzi na posição avaliaram na quarta-feira que ele se tornou uma figura importante para a crise da área e passou a defender sua substituição. Em 2001, há cerca de oito meses, Solange liderou os fundos de pensão

### B.3.12 PTT5-E

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira será a nova presidente da Agência Nacional de Aviação Civil (Anac). Braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES e comandou a secretaria de previdência complementar no governo Fernando Henrique Cardoso. Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa. Os dois ministros que davam sustentação a Zuanazzi no cargo

### B.3.13 PTT5-F

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira será a nova presidente da Agência Nacional de Aviação Civil (Anac). Braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES e comandou a secretaria de previdência complementar no governo Fernando Henrique Cardoso. Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa. Os dois ministros que davam sustentação a Zuanazzi no cargo, Dilma Rousseff e Walfrido dos Mares Guia, avaliam que ele se tornou uma figura muito vinculada à crise aérea e passaram a defender sua substituição. O último remanescente da diretoria da Anac à época do acidente, Josef Barat, diretor de relações internacionais, pesquisas e capacitação, também deve sair. Outros

### B.3.14 PTT5-G (WikiLingua)

A economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil no governo Fernando Henrique Cardoso. O ministro da defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira será o novo presidente na Agência Nacional de Aviação civil. Ele já renunciou e deve entregar o cargo nos próximos dias. Os dois diretores de agências têm mandato de cinco anos e só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa. Quando o último remanescente da diretoria da Anac à época do acidente, Josef Barat, diretor de relações internacionais, pesquisas e capacitação da Anac, também deve sair.

### B.3.15 PTT5-H (XL-Sum)

O ministro da Defesa, Nelson Jobim, informou na noite desta terça-feira que Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). Ela é funcionária de carreira do BNDES e comandou a secretaria de previdência complementar no governo Fernando Henrique Cardoso. Ela será vice-presidente da Anac desde o acidente com

o Airbus da TAM em Brasília e já teria concordado em renunciar e deve deixar o cargo nos próximos dias.

### B.3.16 mT5-F

O economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). O governo Fernando Henrique Cardoso se tornou uma figura muito vinculada à crise e passaram para defender sua substituição. Os diretores de agências têm mandato de cinco anos. Em 2001, por cerca de oito meses, ele comandou com mão-de-ferro outros fundos de pensão do país, como titular da secretaria de previdência complementar do ministério da prevença, desde sua posse, ela é funcionária de carreira do BNDES, que foi morto em dois aeroportos da TAM, na cidade de Brasília, em seu cargo, já teria concordado em renunciar ou acusação de improbidade administrativa.

### B.3.17 mT5-G (WikiLingua)

A economista Solange Vieira, de 38 anos, será a presidente da Agência Nacional de Aviação Civil (Anac). O ministro da defesa, Milton Zuanazzi, já teria concordado em renunciar e deve entregar os cargos nos próximos dias. Os diretores de agências têm mandato de cinco anos. O governo também pode sair por renúncia, decisão judicial ou acusação de improbidade administrativa. Porém, outros dois diretores de relações internacionais, avaliam que ele se tornou uma figura muito vinculada à crise

### B.3.18 mT5-H (WikiLingua)

o ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será nova presidente da Agência Nacional de Aviação Civil (Anac). Porém, os diretores de agências têm mandato de cinco anos. Em 2003, por cerca de oito meses, ele comandou à secretaria de previdência complementar no governo Fernando Henrique Cardoso.

### B.3.19 mT5-I (XL-Sum)

A Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES, com mão-de-ferro os fundos de pensão do país, como titular da secretaria de previdência. O atual presidente, Milton Zuanazzi, já teria concordado em renunciar e deve entregar seus cargos.

### B.3.20 mT5-J (XL-Sum)

A economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil, braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES no governo Fernando Henrique Cardoso. Ela comandou com mão-de-ferro os fundos de pensão do país, como titular da secretaria de previdência complementar do Ministério da Previdência. O atual presidente, Milton Zuanazzi, já teria concordado em renunciar e deve entregar o cargo nos próximos dias, porém, procurado pela GloboNews TV, ele foi renunciodo.