Sumarização Automática de Artigos de Notícias em Português: Da Extração à Abstração com Abordagens Clássicas e Modelos Neurais

Marcio Alves Sarmento[1], **Hilário Tomaz Alves de Oliveira**[1]

[1]Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES) – Campus Serra

malves.sarmento@gmail.com, hilario.oliveira@ifes.edu.br

Abstract. *Automatic text summarization aims to generate a summary that captures the most relevant information from one or more textual documents. Although there have been significant advances in this field, research on documents written in Brazilian Portuguese remains limited. This article provides an analysis of various summarization approaches, ranging from classical baselines to extractive methods, including the fine-tuning of different architectures of the $PPT5$ and $FLAN$-$T5$ and the use of large language models for abstractive summarization. Experiments were conducted using three datasets of news articles written in Portuguese. The results showed that the models fine-tuned for the abstractive summarization achieved competitive performance based on ROUGE-L and BERTScore when compared to larger models like GPT-4o.*

Resumo. *A sumarização automática de texto tem como objetivo a criação de um resumo com as informações mais relevantes extraídas de um ou mais documentos textuais. Apesar dos avanços obtidos na área, pesquisas envolvendo documentos escritos em português do Brasil ainda são escassas. Este artigo apresenta uma análise envolvendo diferentes abordagens de sumarização, desde baselines clássicas, passando por sistemas extrativos, o ajuste fino de diferentes arquiteturas dos modelos $PPT5$ e $FLAN$-$T5$, até o uso de modelos de linguagem de larga escala para sumarização abstrativa. Experimentos foram realizados considerando três bases de dados de artigos de notícias escritos em português. Os resultados demonstraram que os modelos ajustados para a tarefa de sumarização abstrativa obtiveram resultados competitivos com base nas medidas do ROUGE-L e do BERTScore com modelos maiores, como o GPT-4o.*

1. Introdução

A crescente demanda por informações impulsiona o desenvolvimento de tecnologias capazes de processar e sintetizar grandes volumes de dados de forma rápida e eficiente. Em um cenário em que a produção de conteúdo digital online é cada vez mais abundante, torna-se cada vez mais desafiador para os leitores acompanhar todas as notícias relevantes [Levitin 2014, Zhang et al. 2024]. Nesse contexto, sistemas de Sumarização Automática de Texto (SAT) podem ser ferramentas úteis para auxiliar os usuários, oferecendo a capacidade de gerar resumos concisos que capturam as informações mais relevantes de um texto ou de múltiplos documentos relacionados, permitindo uma assimilação mais rápida do conteúdo [Lin and Ng 2019, Zhang et al. 2022].

A SAT é uma área de pesquisa em Processamento de Linguagem Natural (PLN) que busca gerar resumos de documentos textuais de forma automática. Existem duas abordagens principais para a tarefa de SAT: a *Extrativa* e a *Abstrativa* [Nenkova and McKeown 2012]. A *sumarização extrativa* seleciona as frases mais relevantes diretamente do texto original para compor o resumo, enquanto a *sumarização abstrativa* envolve a reescrita do conteúdo de forma mais condensada e frequentemente utiliza técnicas de geração de linguagem natural, sendo capaz de criar novas frases que não necessariamente aparecem no texto original [Zhang et al. 2022]. A sumarização pode ser aplicada tanto a um único documento (*monodocumento*) quanto a um conjunto de documentos (*multidocumento*).

Nos últimos anos, houve uma mudança no foco das pesquisas na área de SAT das abordagens extrativas para as abstrativas [Lin and Ng 2019]. Essa mudança foi impulsionada pelo desenvolvimento de algoritmos baseados em redes neurais profundas, especialmente na arquitetura *Transformer* [Vaswani 2017], capazes de geração de linguagem natural. Assim, diversas abordagens surgiram usando modelos neurais pré-treinados e, mais recentemente, modelos de linguagem de larga escala, do inglês *Large Language Models* (LLMs) [Zhang et al. 2022]. Contudo, apesar dos resultados promissores usando essas abordagens neurais, elas impõem diversos desafios, como a necessidade de grandes bases de dados para treinamento e demandam muitos recursos computacionais. Apesar dos avanços na área, a maioria das pesquisas tem como foco a língua inglesa, e poucos estudos têm sido dedicados ao português, especialmente para a sumarização abstrativa [Zhang et al. 2022]. Essa lacuna limita a aplicabilidade de sistemas de sumarização automáticos projetados especificamente para o português, que enfrenta a ausência de modelos e bases de dados para suportar essas pesquisas [Paiola et al. 2024].

Este artigo busca contribuir para o desenvolvimento da tarefa de SAT em português do Brasil por meio de uma investigação de algoritmos de sumarização aplicados a artigos de notícias. O estudo abrange desde o uso de sistemas extrativos, comumente usados como *baselines* de comparação, o ajuste fino dos modelos $PTT5$ e $FLAN\text{-}T5$ para sumarização abstrativa, até o uso de LLMs de código aberto e proprietários, como o GPT4-o[1], Llama 3 [Touvron et al. 2023] e o Gemma [Team et al. 2024]. Para isso, foram realizados experimentos em três bases de dados, o Temário e Recognnasum para a tarefa de sumarização monodocumento e o CSTNews para sumarização multidocumento. O desempenho dos modelos foi avaliado usando as medidas de avaliação automáticas do ROUGE-L e BERTScore, que são usualmente adotadas na literatura.

Os códigos desenvolvidos e os resumos gerados neste trabalho estão públicos em `https://github.com/laicsiifes/benchmark_ptbr_summ`.

2. Trabalhos Relacionados

A literatura da área de SAT é vasta e existem diversos *surveys* que fornecem uma visão ampla do desenvolvimento da área desde a sua origem [Nenkova and McKeown 2012, Lin and Ng 2019, Zhang et al. 2022]. Por limitações de espaço, esta seção foca apenas em trabalhos que envolveram documentos escritos em português ou que usaram técnicas adotadas nos experimentos realizados neste estudo.

[1] `https://openai.com/index/hello-gpt-4o/`

Diversos indicadores de relevância vêm sendo explorados para a execução da tarefa de sumarização extrativa [Leite and Rino 2008, Oliveira et al. 2016a]. Em sua maioria, esses indicadores baseiam-se em técnicas estatísticas, como frequência e centralidade, ou em heurísticas, como a posição das sentenças nos documentos. O estudo conduzido por Oliveira et al. [Oliveira et al. 2016a] avaliou diferentes técnicas para mensurar a relevância de sentenças em tarefas de SAT de artigos jornalísticos em inglês. Os autores analisaram os métodos individualmente e em combinação, utilizando-os como atributos em algoritmos de classificação. Leite e Rino [Leite and Rino 2008] investigaram uma abordagem combinando múltiplas *features* e algoritmos de aprendizado de máquina para a sumarização extrativa de documentos em português.

No trabalho de Sodré e Oliveira [Sodré and de Oliveira 2019], os autores investigaram a estratégia de combinar alguns dos indicadores analisados por Oliveira et al. [Oliveira et al. 2016a] e aplicaram algoritmos de regressão para estimar um escore de relevância das sentenças na tarefa de sumarização de artigos jornalísticos em português. Gomes e Oliveira [Gomes and de Oliveira 2019] propuseram um sistema usando Programação Linear Inteira (PLI) para sumarização extrativa multidocumento. O sistema desenvolvido usa bigramas como conceitos e aplica métodos estatísticos tradicionais para identificar as informações mais relevantes para a construção do resumo.

Diferentemente dos trabalhos anteriores, Paiola et al. [Paiola et al. 2022] investigaram a tarefa de sumarização abstrativa. Os autores usaram diversas bases de dados em português (TeMário, CSTNews, WikiLingua e XL-Sum) e um sistema de tradução para aplicar modelos treinados em inglês. Em [Paiola et al. 2024], os autores apresentam a base de dados do RecognaSumm, um conjunto de dados contendo mais de 135 mil artigos de notícias para a tarefa de SAT. Os autores realizaram diferentes análises da base de dados proposta e avaliaram o desempenho do modelo base do $PTT5$ para estabelecer um desempenho de referência para comparações futuras.

Este trabalho busca expandir os anteriores ao realizar uma análise mais ampla considerando três bases de dados (Temário, RecognaSumm e CSTNews) para sumarização monodocumento e multidocumento, além de envolver desde técnicas de sumarização extrativas tradicionais de ponderação das frases assim como os usados nos trabalhos em [Oliveira et al. 2016a, Sodré and de Oliveira 2019], adaptação do sistema de PLI proposto em [Gomes and de Oliveira 2019], ajuste fino e avaliação de diferentes tamanhos de arquiteturas ($small$, $base$ e $large$) dos modelos $PTT5$ e $FLAN$-$T5$ e o uso de LLMs de código aberto ($Llama3$ e $Gemma2$) e proprietários (GPT-3.5 e GPT-4o).

3. Materiais e Métodos

3.1. Bases de Dados

Neste trabalho, foram utilizadas três bases de dados comumente usadas na literatura para a tarefa de SAT no domínio de artigos de notícias escritas em português.

TeMário. Esse conjunto de dados é formado por 100 textos jornalísticos, provenientes da Folha de S.Paulo e do Jornal do Brasil. Os artigos, que abordam uma variedade de temas, foram selecionados por sua linguagem clara e objetiva. Todos os textos possuem resumos elaborados por um especialista, o que garante a qualidade dos resumos de referência [Pardo and Rino 2003].

CSTNews. Essa base de dados é formada por 50 conjuntos de notícias, cada um com aproximadamente quatro artigos sobre o mesmo tema, coletados manualmente em sites de notícias como Folha de São Paulo e Estadão. Essa abordagem permitiu a seleção de notícias com linguagem clara e acessível, provenientes de diferentes fontes sobre um mesmo assunto [Cardoso et al. 2011].

RecognaSumm. Com o objetivo de construir um conjunto de dados robusto para estudos de sumarização de textos, Paiola et al. [Paiola et al. 2024] coletaram 135.272 artigos de notícias usando sistemas de *web crawlers* personalizados. A diversidade temática dos artigos foi garantida pela coleta de dados em diferentes portais de notícias e categorias. A base de dados é dividida em três subconjuntos: treinamento, validação e teste. Por conta de limitações de *hardware*, foi feita uma filtragem no conjunto de treinamento, sendo removidos os artigos com resumos contendo menos do que 25 palavras.

Na Tabela 1 são apresentadas algumas estatísticas das bases de dados usadas nos experimentos. Para cada base, foi computado o total de documentos ou grupos, média e Desvio Padrão (DP) de frases e palavras no texto dos artigos.

Tabela 1. Estatísticas das bases de dados usadas nos experimentos.

Base de Dados	Conjunto	Docs / Grupos	Média (DP) Frases	Média (DP) Palavras
TeMário	Único	100	32,4 (10,38)	618,67 (163,93)
CSTNews	Único	50	47,06 (19,47)	939,56 (331,42)
RecognaSumm	Treino	64.347	27,07 (24,82)	527,33 (468,38)
	Validação	21.538	26,73 (24,15)	519,91 (458,68)
	Teste	21.493	27,05 (24,88)	526,41 (470,02)

3.2. Modelos de Sumarização

Os seguintes modelos de sumarização foram investigados:

Baselines. Foram utilizadas como baselines oito técnicas de ponderação de frases [Sodré and de Oliveira 2019]. As técnicas utilizadas foram: *Bushy Path*, Centralidade das Frases, Frequência de Palavras, Frequência de Entidades Nomeadas, Frequência do Termo - Frequência Inversa das Sentenças (TF-ISF), Posição das Frases, Similaridade Agregada, *TextRank*. Essas técnicas foram usadas em conjunto com uma abordagem clássica de sumarização extrativa composta por três etapas [Nenkova and McKeown 2012]:

- **Pré-processamento:** O documento ou grupo de documentos de entrada é pré-processado usando várias técnicas tradicionais de PLN, como divisão do texto em frases, palavras, lematização, identificação das classes gramaticais e reconhecimento de entidades nomeadas.
- **Ponderação das frases:** Nesta etapa, cada uma das oito técnicas de ponderação de frases é aplicada para analisar cada frase do(s) documento(s) de entrada e gerar um valor que deve refletir sua relevância para ser incluído no resumo. Todos os valores gerados são normalizados no intervalo de 0 a 1.
- **Geração de resumo:** As frases com os maiores valores de relevância geradas na etapa anterior são inseridas iterativamente no resumo até que o tamanho máximo desejado seja atingido. Uma nova frase é inserida no resumo somente

se sua similaridade de cosseno com as frases já inseridas for menor que $0,5$ [Nenkova and McKeown 2012].

Sistema de PLI. Foi utilizado o sistema de PLI proposto por Gomes e Oliveira [Gomes and de Oliveira 2019] para sumarização multidocumento e uma adaptação de um sistema similar apresentado em [Oliveira et al. 2016b] para a tarefa de sumarização monodocumento.

Modelos Pré-treinados. Foram utilizados os modelos $PTT5$ [Carmo et al. 2020] e $FLAN\text{-}T5$ em suas arquiteturas *small*, *base* e *large*, que se diferenciam pelo tamanho da arquitetura. O $PTT5$ é uma versão em português do modelo de linguagem $T5$, pré-treinado no BrWac, um grande corpus de páginas da web em português brasileiro. O $FLAN\text{-}T5$ é um modelo multilíngue desenvolvido pela google que foi treinado para múltiplas tarefas de PLN [Chung et al. 2024].

LLMs. Os recentes avanços no progresso de LLMs têm impulsionado o desenvolvimento de diversas aplicações. Neste trabalho, foram utilizados os modelos: Gemma 2 9B [Team et al. 2024], o Llama 3.1 8B [Touvron et al. 2023] e os modelos Text-davinci-003, GPT-3.5 Turbo, GPT-4o e GPT-4o mini desenvolvidos pela empresa OpenAI [OpenAI 2024].

3.3. Desenho Experimental

A análise de desempenho dos modelos de sumarização foi dividida em dois experimentos. No primeiro experimento, foi utilizada somente a base de dados do RecognaSumm, sendo considerados os sistemas extrativos (*baselines* e o sistema de PLI) e foram treinados seis modelos de sumarização abstrativos baseados no $PTT5$ e $FLAN\text{-}T5$. O segundo experimento foi realizado usando as bases de dados do Temário e CSTNews. Para esse experimento, foram usados os sistemas extrativos (*baselines* e o sistema de PLI), os modelos abstrativos baseados no $PTT5$ e $FLAN\text{-}T5$ treinados no primeiro experimento e os modelos de LLMs. Em todas as abordagens avaliadas, foi configurado o tamanho máximo de resumo para 150 palavras.

Para os métodos de *baselines* e o sistema de PLI foram usadas implementações próprias. Os modelos da OpenAI foram acessados usando a API oficial disponibilizada pela empresa. A implementação dos modelos $PTT5$, $FLAN\text{-}T5$ e dos LLMs do *Gemma 2 9B* e *Llama 3.1 8B* foi baseada na biblioteca *Transformers*[2] e foram usados os modelos pré-treinados disponibilizados publicamente pelos autores e empresas na plataforma do Hugging Face[3]. Para o ajuste fino das três arquiteturas dos modelos $PTT5$ e $FLAN\text{-}T5$, o tamanho máximo de entrada foi definido para 512 *tokens* e o tamanho máximo do resumo a ser gerado foi configurado para 150 *tokens*. Os modelos foram ajustados por no máximo 20 épocas, sendo utilizada a estratégia de parada antecipada com uma paciência de 5 épocas. Para evitar sobreajuste dos modelos, foi feito um monitoramento do treinamento, no qual, ao final de cada época, o modelo resultante é aplicado no conjunto de validação e é computada a medida do ROUGE-L, sendo armazenado somente o modelo com maior valor. Para a geração dos resumos, foi usado o algoritmo de decodificação do *Beam Search* com tamanho 5 de largura.

[2] `https://huggingface.co/docs/transformers/index`
[3] `https://huggingface.co/`

Baseado no trabalho de [Zhang et al. 2024], o seguinte *prompt* foi usado nos LLMs para geração dos resumos: "*Escreva um resumo em PORTUGUÊS DO BRASIL para o artigo de notícias a seguir com no MÁXIMO 150 palavras. ARTIGO:* $\{TEXTO\}$.", onde $\{TEXTO\}$ foi substituído pelo conteúdo completo do(s) artigo(s) de notícias.

O desempenho dos modelos foi avaliado utilizando as medidas de avaliação do *Recall-Oriented Understudy for Gisting Evaluation Longest Common Subsequence* (LCS) (ROUGE-L) [Lin 2004] e a do BERTScore [Zhang et al. 2019]. O ROUGE-L computa a maior cadeia em comum entre um resumo candidato e o resumo de referência, enquanto o BERTScore calcula a similaridade do cosseno entre dois textos usando representações de *embeddings* extraídas do modelo *Bidirecional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019]. Por questões de espaço, são reportados somente a métrica do *f1-score*, que combina as métricas de precisão e revocação. Apesar de terem diversas limitações, essas medidas são alternativas válidas à realização de avaliações manuais e, conforme análise feita em Zhang et al. [Zhang et al. 2024], elas apresentaram correlação moderada com avaliações humanas na tarefa de sumarização.

4. Resultados

4.1. Experimento na base de dados do RecognaSumm

Na Tabela 2 são apresentados os resultados dos experimentos na base de dados do RecognaSumm. Analisando o desempenho dos *baselines*, pode-se observar que a técnica da *Posição das Frases* foi a que obteve os melhores resultados. Essa técnica consiste em selecionar as n primeiras frases do documento para compor o resumo até que o tamanho máximo do resumo desejado seja alcançado. Essa técnica tem sido um dos *baselines* mais competitivos para sumarização de artigos de notícias [Oliveira et al. 2016a]. O sistema baseado em PLI obteve melhor desempenho do que quase todos os *baselines*, com exceção da *Posição das Frases*. Os modelos $PTT5$ e $FLAN\text{-}T5$ demonstraram melhor desempenho geral do que as demais abordagens analisadas. Em especial, os melhores desempenhos neste experimento foram obtidos pelos modelos $FLAN\text{-}T5_{Large}$ e $PTT5_{Large}$ em ambas as medidas de avaliação. Os resultados obtidos usando a arquitetura $base$ foram muito próximos às arquiteturas da *large*, sendo que eles são menores e consomem menos recursos computacionais.

Com base nos resultados, fica evidente que os modelos ajustados para sumarização abstrativos geraram resumos melhores do que as técnicas de *baselines* e que o sistema extrativo de PLI nas medidas do ROUGE-L e BERTScore. Essa superioridade demonstra a eficácia dos modelos $PTT5$ e $FLAN\text{-}T5$ para a tarefa de geração de resumos abstrativos. Entretanto, ao considerar o uso desses modelos, é importante levar em conta o custo computacional associado a cada um, tanto para o treinamento quanto para a geração dos resumos. Portanto, a relação custo-benefício deve ser ponderada na escolha da abordagem, especialmente em cenários com recursos computacionais limitados.

4.2. Experimento nas bases de dados do Temário e CSTNews

A Tabela 3 apresenta os resultados do experimento nas bases de dados do TeMário e CSTNews. As abordagens avaliadas incluem os métodos de *baselines*, o sistema extrativo usando PLI, os modelos do $PTT5$ e $FLAN\text{-}T5$ treinados no RecognaSumm e os LLMs analisados. Os resultados obtidos neste experimento foram bastante diversificados.

Tabela 2. Resultados do experimento usando o corpus RecognaSumm.

Abordagem		ROUGE-L	BERTScore
Baselines	Bushy Path	0,249 (0,086)	0,691 (0,037)
	Centralidade das Frases	0,249 (0,087)	0,690 (0,038)
	Frequência de Palavras	0,240 (0,085)	0,686 (0,038)
	Freq. Entidades Nomeadas	0,242 (0,088)	0,681 (0,038)
	Posição das Frases	0,279 (0,099)	0,701 (0,040)
	Similaridade Agregada	0,249 (0,085)	0,689 (0,039)
	TextRank	0,206 (0,072)	0,674 (0,034)
	TF-ISF	0,235 (0,084)	0,684 (0,037)
Extrativo	Sistema PLI	0,270 (0,095)	0,694 (0,038)
Abstrativos	$PTT5_{Small}$	0,315 (0,125)	0,713 (0,045)
	$PTT5_{Base}$	0,337 (0,132)	0,722 (0,045)
	$PTT5_{Large}$	0,346 (0,134)	0,726 (0,046)
	$FLAN\text{-}T5_{Small}$	0,314 (0,130)	0,714 (0,045)
	$FLAN\text{-}T5_{Base}$	0,338 (0,140)	0,724 (0,048)
	$FLAN\text{-}T5_{Large}$	**0,349 (0,143)**	**0,729 (0,048)**

Na base de dados do Temário, os modelos Text-davinci-003 e GPT-4o obtiveram o melhor desempenho nas medidas do ROUGE-L e BERTScore, respectivamente. Na base do CSTNews, o *baseline* de *Posição das Frases* apresentou o melhor resultado no ROUGE-L e o GPT-3.5 Turbo no BERTScore.

Tabela 3. Resultados do experimento usando o Temário e o CSTNews.

Abordagens	Sistema	Temário		CSTNews	
		ROUGE-L	BERTScore	ROUGE-L	BERTScore
Baselines	Bushy Path	0,396 (0,069)	0,694 (0,024)	0,447 (0,067)	0,720 (0,029)
	Cent. das Frases	0,384 (0,063)	0,690 (0,025)	0,454 (0,066)	0,723 (0,031)
	Freq. de Palavras	0,375 (0,069)	0,686 (0,023)	0,452 (0,063)	0,721 (0,029)
	Freq. Ent. Nom.	0,389 (0,076)	0,683 (0,024)	0,434 (0,068)	0,705 (0,032)
	Posição das Frases	0,402 (0,070)	0,686 (0,022)	**0,482 (0,047)**	0,733 (0,024)
	Sim, Agregada	0,390 (0,070)	0,696 (0,025)	0,419 (0,050)	0,712 (0,024)
	TextRank	0,350 (0,059)	0,685 (0,021)	0,415 (0,060)	0,709 (0,027)
	TF-ISF	0,379 (0,072)	0,685 (0,024)	0,451 (0,061)	0,718 (0,032)
Extrativo	Sistema PLI	0,396 (0,065)	0,687 (0,023)	0,477 (0,049)	0,736 (0,033)
Abstrativos	$PTT5_{Small}$	0,348 (0,064)	0,679 (0,024)	0,393 (0,065)	0,713 (0,031)
	$PTT5_{Base}$	0,346 (0,062)	0,681 (0,023)	0,384 (0,055)	0,712 (0,028)
	$PTT5_{Large}$	0,339 (0,062)	0,678 (0,025)	0,385 (0,055)	0,715 (0,026)
	$FLAN\text{-}T5_{Small}$	0,241 (0,053)	0,654 (0,021)	0,304 (0,070)	0,700 (0,027)
	$FLAN\text{-}T5_{Base}$	0,242 (0,048)	0,658 (0,022)	0,290 (0,064)	0,700 (0,033)
	$FLAN\text{-}T5_{Large}$	0,225 (0,049)	0,654 (0,021)	0,294 (0,070)	0,696 (0,033)
LLMs	Gemma 2 9B	0,354 (0,046)	0,690 (0,020)	0,383 (0,034)	0,717 (0,023)
	Llama 3.1 8B	0,320 (0,037)	0,671 (0,019)	0,338 (0,037)	0,699 (0,026)
	Text-davinci-003	**0,424 (0,075)**	0,705 (0,027)	0,472 (0,048)	0,738 (0,029)
	GPT-3.5 Turbo	0,402 (0,074)	0,705 (0,025)	0,455 (0,061)	**0,740 (0,025)**
	GPT-4o	0,417 (0,062)	**0,713 (0,025)**	0,452 (0,043)	0,731 (0,023)
	GPT-4o Mini	0,402 (0,059)	0,705 (0,021)	0,444 (0,039)	0,730 (0,023)

Os métodos de *baselines*, o sistema extrativo baseado em PLI e os LLMs apresentaram resultados próximos com base nas medidas de avaliação. Por outro lado, os modelos ajustados do $PTT5$ e $FLAN\text{-}T5$ demonstraram desempenho inferior aos demais, especialmente os modelos do $FLAN\text{-}T5$. Esse baixo desempenho pode ser atribuído ao

fato desses modelos consistentemente gerarem resumos com tamanhos bem inferiores aos demais, mesmo sendo definido um tamanho máximo de 150 palavras. Essa característica aconteceu por conta do treinamento desses modelos no Recognasumm, que possui resumos de referência bem menores do que os do Temário e do CSTNews.

Apesar dos resultados quantitativos serem próximos, ao analisar os resumos gerados pelas abordagens extrativas e abstrativas, fica evidente que os resumos extrativos, em geral, possuem muitas informações contidas nos resumos de referências, mas os resumos possuem diversos problemas de coerência e coesão textual. Por outro lado, os resumos abstrativos são mais sucintos e, em sua maioria, apresentam uma boa qualidade textual em termos de coerência, coesão e estrutura ortográfica e gramatical. Os LLMs do Gemma 2 9B e do Llama 3.1 8B apresentaram uma tendência de terminar de forma brusca os resumos, por exemplo, no meio de uma frase. Cabe ressaltar que nenhum LLM foi ajustado para a tarefa de sumarização.

Por fim, é importante enfatizar que os LLMs, como Gemma, Llama e especialmente os modelos da OpenAI, possuem um custo consideravelmente maior do que os demais modelos avaliados neste trabalho. Essa característica deve ser considerada em aplicações práticas, na qual a relação custo-benefício é determinante. Nesse contexto, abordagens extrativas, como o sistema baseado PLI ou mesmo os *baselines*, podem oferecer uma alternativa que equilibra desempenho com menor custo computacional. Em cenários com recursos computacionais moderados, os modelos ajustados do $PTT5$ e $FLAN\text{-}T5$ podem ser as melhores opções.

5. Conclusões

Este trabalho apresentou uma análise comparativa de várias abordagens para sumarização automática de texto, considerando desde tradicionais métodos de ponderação de frases até modelos de linguagem de grande escala, para sumarização abstrativa e extrativa de artigos de notícias escritas em português do Brasil. Essa avaliação fez uso de três bases de dados e de duas medidas de avaliação automática comumente usadas na literatura. Os resultados obtidos demonstram que os modelos de LLMs são promissores para a tarefa de criação automática de resumos, mas são sistemas com uma alta complexidade que requerem muitos recursos computacionais. Portanto, modelos especializados para a tarefa de sumarização ou sistemas extrativos ainda podem ser opções viáveis, especialmente em cenários de poucos recursos.

Em trabalhos futuros, pretendemos expandir este trabalho visando: **(i)** analisar o desempenho de modelos de LLM de código aberto, considerando diferentes cenários de utilização, como *zero shot-learning*, *few-shot learning* e fazendo o ajuste fino desses modelos para a tarefa de sumarização; e **(ii)** realizar uma avaliação manual de um subconjunto dos resumos gerados para complementar as análises automáticas.

Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (processo 2021-2S6CD, nº FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Referências

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., and Pardo, T. A. (2011). Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gomes, L. and de Oliveira, H. (2019). A multi-document summarization system for news articles in portuguese using integer linear programming. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 622–633. SBC.

Leite, D. S. and Rino, L. H. M. (2008). Combining multiple features for automatic text summarization through machine learning. In *International Conference on Computational Processing of the Portuguese Language*, pages 122–132. Springer.

Levitin, D. J. (2014). *Organized Mind: Thinking Straight in the Age of Information Overload (9780698157224)*. Barnes & Noble.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, H. and Ng, V. (2019). Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.

Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016a). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications*, 65:68–86.

Oliveira, H., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016b). A concept-based integer linear programming approach for single-document summarization. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–408. IEEE.

OpenAI (2024). Openai models. https://openai.com/api/.

Paiola, P. H., de Rosa, G. H., and Papa, J. P. (2022). Deep learning-based abstractive summarization for brazilian portuguese texts. In Xavier-Junior, J. C. and Rios, R. A., editors, *Intelligent Systems*, pages 479–493, Cham. Springer International Publishing.

Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024). Recognasumm: A novel brazilian summarization dataset. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 575–579.

Pardo, T. A. S. and Rino, L. H. M. (2003). Temário: Um corpus para sumarização automática de textos. *São Carlos: Universidade de São Carlos, Relatório Técnico*.

Sodré, L. and de Oliveira, H. (2019). Avaliando algoritmos de regressão para sumarização automática de textos em português do brasil. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 634–645. SBC.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Zhang, M., Zhou, G., Yu, W., Huang, N., and Liu, W. (2022). A comprehensive survey of abstractive text summarization based on deep learning. *Computational intelligence and neuroscience*, 2022(1):7132226.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.