Integração de Conhecimento Linguístico Estruturado em LLMs para a Tarefa de Sumarização Extrativa

Carlos Vitor Cardoso da Silva [1], **Paula Christina Figueira Cardoso** [1]

[1] Instituto de Ciências Exatas e Naturais – Faculdade de Computação (FACOMP)
Universidade Federal do Pará (UFPA) – Belém – PA – Brasil

cvcardos5@gmail.com, pcardoso@ufpa.br

Abstract. *This work investigates the integration of Rhetorical Structure Theory (RST) into LLMs (Gemma 3, Llama 3.2, Ministral) for extractive summarization. We evaluated twelve configurations (zero/one-shot, plain/RST formats) against the baseline from Ono et al. and human summaries from the CSTNews corpus. The Ono baseline remained competitive (ROUGE-L = 0.428), not being surpassed by the best LLM configuration (Ministral/zero-shot/RST, ROUGE-L = 0.422). The RST format produced divergent and model-dependent effects on the ROUGE and BERTScore metrics, with a variable impact on extraction quality.*

Resumo. *Este trabalho investiga a integração da Rhetorical Structure Theory (RST) em LLMs (Gemma 3, Llama 3.2, Ministral) para sumarização extrativa. Avaliamos doze configurações (zero/one-shot, formatos plain/RST) contra o baseline de Ono et al. e sumários humanos do corpus CSTNews. O baseline de Ono manteve-se competitivo (ROUGE-L = 0,428), não sendo superado pela melhor configuração LLM (Ministral/zero-shot/RST, ROUGE-L = 0,422). O formato RST produziu efeitos divergentes e dependentes do modelo nas métricas ROUGE e BERTScore, com impacto variável na qualidade da extração.*

1. Introdução

A sumarização automática (SA) consiste em produzir um texto condensado que mantenha as informações relevantes de um documento original [Martins et al. 2001]. Os sumários podem ser extrativos, quando partes do texto-fonte são selecionadas sem modificações, ou abstrativos, quando há reescrita do conteúdo [Nenkova and McKeown 2012]. Este trabalho foca na SA extrativa, cuja propriedade central é reproduzir trechos do documento original sem alterá-los, dependendo da capacidade do sistema em identificar os segmentos textuais mais informativos. Nesse contexto, a Teoria da Estrutura Retórica (*Rhetorical Structure Theory* - RST) [Mann and Thompson 1987] oferece um arcabouço para modelar a organização discursiva do texto via relações retóricas hierárquicas, distinguindo segmentos nucleares (essenciais) de satélites (suporte) e fornecendo critérios estruturados para seleção de conteúdo [Ono et al. 1994, Uzêda et al. 2010, Cardoso et al. 2014].

Nos últimos anos, os *Large Language Models* (LLMs) demonstraram forte capacidade de compreensão e geração de texto [Zhang et al. 2025]. No entanto, os estudos recentes que empregam LLMs para sumarização em português [de Camargo et al. 2025, Sarmento and de Oliveira 2024] fornecem ao modelo apenas o texto-fonte, sem informações sobre sua estrutura discursiva. Permanece em aberto, portanto, a questão de se e como metadados discursivos da RST, quando incorporados explicitamente ao *prompt*, podem auxiliar LLMs na tarefa de SA extrativa.

Este trabalho de iniciação científica contribui para preencher essa lacuna por meio de um experimento que combina três LLMs (Gemma 3, Llama 3.2 e Ministral, cada um com aproximadamente 3 a 4 bilhões de parâmetros), duas estratégias de *prompting* (*zero-shot* e *one-shot*) e dois formatos de entrada (texto-fonte puro e texto com estrutura RST simplificada), totalizando doze configurações experimentais. A avaliação é realizada por comparação com os sumários *gold standard* do CSTNews [Cardoso et al. 2011], utilizando as métricas ROUGE [Lin 2004] e BERTScore [Zhang et al. 2019]. Como linha de base adicional, os resultados dos LLMs são comparados ao algoritmo de Ono et al. [Ono et al. 1994], o método baseado em RST com melhor desempenho reportado na literatura [Uzêda et al. 2010].

2. Fundamentação Teórica

Rhetorical Structure Theory (RST). A RST [Mann and Thompson 1987] analisa a organização do discurso a partir da conexão entre unidades mínimas de discurso (EDU), formando uma estrutura hierárquica em árvore. A teoria incorpora o conceito de **nuclearidade**: uma unidade atua como núcleo quando é essencial para a compreensão do trecho, ou como satélite quando fornece informação dependente do núcleo. As relações classificam-se em mononucleares (um núcleo modificado por satélites) ou multinucleares (múltiplos núcleos de igual importância).

RST aplicada à SA. A aplicação da RST em SA tem sido extensivamente estudada [Ono et al. 1994, Uzêda et al. 2010, Taboada and Mann 2006]. O trabalho de [Uzêda et al. 2010] comparou abordagens extrativas superficiais, métodos baseados em RST e híbridos, concluindo que os algoritmos guiados pela RST superam estatisticamente os métodos superficiais. Como *baseline*, este trabalho utiliza o algoritmo de [Ono et al. 1994]. Nele, a raiz da árvore RST recebe pontuação igual à profundidade da árvore; cada nó herda a pontuação do pai se for núcleo ou recebe a pontuação decrementada em 1 se for satélite.A escolha desse *baseline* é motivada por este ter sido o método baseado em RST de melhor desempenho (F-Score ROUGE-1 = 0,418)[Uzêda et al. 2010].

LLMs em SA e métricas. Para [Brown et al. 2020], incluir um (*one-shot*) ou mais (*few-shot*) exemplos no *prompt* pode reduzir a ambiguidade da instrução em modelos de menor capacidade. [de Camargo et al. 2025] avaliaram estratégias de *prompting* para sumarização abstrativa no RecognaSumm, evidenciando alta variabilidade dependente do modelo e da estratégia. Já [Sarmento and de Oliveira 2024] compararam extratos heurísticos e abstratos por LLMs. No entanto, nenhum desses trabalhos investiga o impacto de informações discursivas estruturadas no *prompt*.

3. Metodologia

Corpus. Usamos o CSTNews [Cardoso et al. 2011] (formado por 50 grupos de textos que totalizam 140 árvores RST anotadas), com sumários *gold standard*. Selecionamos aleatoriamente 57 dos 140 documentos anotados, por limitações de hardware. Os textos-fonte têm média de 328 palavras com desvio padrão igual a 134 palavras, e os sumários humanos têm $\approx$25% do tamanho original, taxa adotada para todos os métodos.

Ambiente experimental e modelos. Os experimentos foram executados em máquina com AMD Ryzen 5 4600G, 16 GB de RAM e GPU Nvidia GTX 1650 (4 GB VRAM),

via *Ollama*[1] com temperatura igual a 0. O *pipeline* foi implementado em Python e a avaliação utilizou as bibliotecas *rouge-score*[2] e *bert-score*[3]. Selecionamos três LLMs de 3–4B de parâmetros: *Gemma 3* (4B, Google), *Llama 3.2* (3B, Meta) e *Ministral* (3B, Mistral AI). Essa faixa representa o maior porte viável para inferência local com 4 GB de VRAM e corresponde à implantação acessível em *hardware* de baixo custo, cenário em que o auxílio de conhecimento linguístico estruturado pode ser mais valioso.

Prompts e formatos. O *baseline* foi o algoritmo de [Ono et al. 1994] com compressão de 25% sobre árvores RST do CSTNews. Para os LLMs, variamos dois eixos: (i) **formato de entrada**: *plain* (texto-fonte puro) e *RST* (segmentos rotulados como `[<relação> - <nuclearidade>] "<texto>"`); (ii) **estratégia**: *zero-shot* (sem exemplo) e *one-shot* (contém o exemplo do Pan-2007, excluído do conjunto de teste). Todos os prompts instruem o modelo a atuar como especialista em sumarização extrativa, proibindo paráfrase, preâmbulos, conectivos adicionados e marcadores de lista. O prompt RST inclui contexto sobre nuclearidade e regras de extração. Os modelos de prompts completos estão disponíveis em `https://github.com/carloscardoso05/tilic2026`.

Avaliação. Os sumários foram avaliados **em relação aos sumários de referência** do CSTNews, calculando F-Score da ROUGE-1/-2-/L (sobreposição lexical), F-Score do BERTScore (similaridade semântica) e cobertura extrativa (percentual de sumários compostos por sequências contíguas do texto-fonte). A significância estatística das diferenças entre pares de configurações foi avaliada pelo teste bilateral de Wilcoxon *signed rank*, emparelhado por documento ($N = 57$).

4. Resultados e Discussões

A Tabela 1 apresenta o desempenho consolidado das 12 configurações LLM e do *baseline* Ono, ordenadas por ROUGE-L decrescente.

Tabela 1. F-Score médio das 12 configurações LLM e do *baseline* Ono. Ext.: cobertura extrativa média (%).

#	Método	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Ext. %
—	Ono et al. (*baseline*)	0,538	0,354	0,428	0,774	100
1	Ministral / zero-shot / RST	**0,543**	**0,365**	**0,422**	0,784	0
2	Gemma 3 / one-shot / RST	0,524	0,349	0,413	0,777	29
3	Ministral / one-shot / RST	0,522	0,339	0,406	**0,785**	2
4	Llama 3.2 / zero-shot / RST	0,514	0,342	0,398	0,758	44
5	Ministral / one-shot / *plain*	0,523	0,343	0,396	0,782	4
6	Llama 3.2 / one-shot / *plain*	0,514	0,341	0,396	0,779	47
7	Llama 3.2 / zero-shot / *plain*	0,484	0,339	0,383	0,779	44
8	Llama 3.2 / one-shot / RST	0,470	0,323	0,371	0,733	0
9	Gemma 3 / zero-shot / *plain*	0,462	0,335	0,369	0,774	54
10	Ministral / zero-shot / *plain*	0,514	0,316	0,366	0,775	2
11	Gemma 3 / one-shot / *plain*	0,461	0,313	0,358	0,765	**67**
12	Gemma 3 / zero-shot / RST	0,454	0,315	0,356	0,762	12

***Baseline* competitivo.** O algoritmo de Ono (ROUGE-L = 0,428) superou marginalmente a melhor configuração LLM (Ministral/*zero-shot*/RST, ROUGE-L = 0,422). Essa diferença não é estatisticamente significativa (Wilcoxon, $p = 0{,}58$), tampouco a diferença em

[1] `https://ollama.com/`
[2] `https://pypi.org/project/rouge-score/`
[3] `https://pypi.org/project/bert-score/`

BERTScore-F ($p = 0{,}051$). Os LLMs avaliados não extraíram informações consistentemente melhor que este método clássico fundamentado em RST.

Efeito do formato RST. O formato RST elevou o ROUGE-L médio (+0,017), mas reduziu o BERTScore-F (-0,009), sugerindo maior sobreposição lexical, porém menor precisão semântica. Os efeitos variaram significativamente por modelo: o ganho de ROUGE-L com RST foi estatisticamente significativo para Gemma 3/*one-shot* e Ministral/*zero-shot* (Wilcoxon, $p < 0{,}01$ e $p < 0{,}001$, respectivamente), mas não para as demais combinações. A redução do BERTScore-F no Llama 3.2 foi significativa em ambas as estratégias ($p < 0{,}01$ no *zero-shot*; $p < 0{,}001$ no *one-shot*).

Dificuldades em seguir restrições. O maior desafio do trabalho concentrou-se na incapacidade dos modelos selecionados em seguir restrições negativas, como a instrução "não gere preâmbulos". Os modelos frequentemente incluíram elementos indevidos no resultado final como preâmbulos, aspas e metadados da RST, apesar da instrução para omiti-los. Embora a remoção desses elementos adicionais tornasse os sumários extrativos, a presença deles comprometeu gravemente as métricas de extratividade.

***One-shot* e extratividade.** O uso de *one-shot* teve efeitos heterogêneos: beneficiou significativamente o Gemma 3 no formato RST (ROUGE-L $p < 0{,}001$) e o Ministral no formato *plain* (ROUGE-L $p < 0{,}01$), mas prejudicou significativamente o Llama 3.2 no formato RST, tanto em ROUGE-L ($p < 0{,}05$) quanto em BERTScore-F ($p < 0{,}01$). As demais combinações não apresentaram diferença estatisticamente significativa.

Síntese por modelo. O **Ministral** obteve as melhores métricas de qualidade (ROUGE/BERTScore), com alta estabilidade, porém extratividade quase nula (2–4%). O **Llama 3.2** teve desempenho intermediário, mostrando-se sensível à combinação de *one-shot* com RST. O **Gemma 3** obteve o menor ROUGE-L, mas destacou-se por gerar a maior proporção de sumários estritamente extrativos no formato *plain* (até 67%).

5. Conclusão

Investigamos o impacto da RST em *prompts* de LLMs para SA extrativa em português. A análise revelou que (i) o *baseline* de Ono permanece competitivo, com a melhor configuração LLM (Ministral/*zero-shot*/RST, ROUGE-L = 0,422) alcançando desempenho equivalente; (ii) o formato RST melhora o ROUGE-L médio (+0,017) mas reduz o BERTScore-F (-0,009), com efeitos estatisticamente significativos apenas para configurações específicas de modelo e estratégia; (iii) o *one-shot* produz ganhos modestos e específicos por modelo; e (iv) o Ministral foi o LLM mais consistente em ROUGE e BERTScore, enquanto o Gemma 3 destacou-se em extratividade. Como limitações, destacam-se a ausência de avaliação humana e a representação RST linear, na qual se perde a hierarquia profunda da árvore. Trabalhos futuros devem explorar serializações hierárquicas da RST, modelos de maior escala via API, *few-shot* e avaliação humana.

6. Agradecimentos

Por fim, agradecemos ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da UFPA pela concessão da bolsa de pesquisa, que viabilizou a investigação teórica e o desenvolvimento deste trabalho.

Referências

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. 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. sn.

Cardoso, P. C. F. et al. (2014). Exploração de métodos de sumarização automática multi-documento com base em conhecimento semântico-discursivo. *São Carlos, SP*.

de Camargo, H. A., Paiola, P. H., Garcia, G. L., and Papa, J. P. (2025). Abstractive Summarization with LLMs for texts in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):1030–1048.

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

Mann, W. C. and Thompson, S. A. (1987). Rhetorical Structure Theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

Martins, C. B., Pardo, T. A. S., Espina, A. P., and Rino, L. H. M. (2001). Introdução à sumarização automática. *Relatório Técnico RT-DC*, 2(1):35.

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

Ono, K., Sumita, K., and Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Sarmento, M. A. and de Oliveira, H. T. (2024). Sumarização automática de artigos de notícias em português: Da extração à abstração com abordagens clássicas e modelos de neurais. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 105–114.

Taboada, M. and Mann, W. C. (2006). Applications of Rhetorical Structure Theory. *Discourse studies*, 8(4):567–588.

Uzêda, V. R., Pardo, T. A. S., and Nunes, M. D. G. V. (2010). A comprehensive comparative evaluation of RST-based summarization methods. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(4):1–20.

Zhang, H., Yu, P. S., and Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

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