Proposta recebida em Novembro 2024 e aceite para publicação em Janeiro 2025.

Verbos *Dicendi* como Indicadores de Coerência em Resumos: uma análise humana e automatizada

Dicendi Verbs as Indicators of Coherence in Abstracts: a human and automated analysis

Osmar de Oliveira Braz Junior
Universidade do Estado de Santa Catarina

Ana Julia Araujo Sanchuki
Universidade Tecnológica Federal do Paraná

Roberlei Alves Bertucci
Universidade Tecnológica Federal do Paraná

Renato Fileto
Universidade Federal de Santa Catarina

Resumo

O resumo escolar desempenha um papel fundamental no desenvolvimento da capacidade de síntese e compreensão de textos, promovendo a interpretação e expressão organizada dos conteúdos do texto-base. Este estudo investiga como a escolha de verbos *dicendi* (verbos de dizer) em resumos pode refletir a coerência e o objetivo do texto-base. Dada a importância da coerência para a qualidade dos resumos, a pesquisa explora o potencial de modelos de linguagem de larga escala (LLMs) em identificar ou sugerir verbos *dicendi* coerentes com o propósito do texto, contribuindo para uma avaliação automática mais precisa. Os experimentos relatados neste artigo utilizaram resumos propostos para um texto dado em um exame vestibular, sobre o qual especialistas e LLMs sugeriram verbos apropriados. As sugestões foram comparadas e avaliadas quanto à coerência com o texto de origem. Os resultados indicam que, embora os LLMs sejam capazes de identificar verbos adequados em alguns contextos, eles apresentam limitações quando comparados à interpretação humana. Concluímos que LLMs podem servir como ferramentas auxiliares na educação, apesar dos desafios relacionados à precisão e ao contexto na avaliação linguística automatizada.

Palavras chave

coerência semântica; semântica de palavras; verbos dicendi; modelo de linguagem de grande escala

Abstract

The school summary plays a fundamental role in developing the ability to synthesize and understand texts, promoting the interpretation and the organized expression of the contents of the source text. This study investigates how the choice of *dicendi* verbs (verbs of saying) in summaries can reflect the coherence and intention of the source text. Given the importance of coherence for the quality of summaries, the research explores the potential of large-scale language models (LLMs) in identifying or suggesting *dicendi* verbs that are coherent with the purpose of the text, contributing to a more accurate automatic evaluation. The experiments reported in this article used summaries proposed for a given text in a college entrance exam, on which experts and LLMs suggested appropriate verbs. The suggestions were compared and evaluated for coherence with the source text. The results indicate that, although LLMs are capable of identifying appropriate verbs in some contexts, they have limitations when compared to human interpretation. We conclude that LLMs can serve as auxiliary tools in education, despite the challenges related to accuracy and context in automated linguistic assessment.

Keywords

semantic coherence; word semantics; dicendi verbs; large language model

1. Introdução

O resumo escolar é um gênero textual fundamental para aprimorar as habilidades de leitura e escrita, pois, ao sintetizar informações de forma organizada, estimula tanto a interpretação do texto original quanto a capacidade de expressar essa compreensão de maneira concisa. E, para que essa compreensão seja checada, é essencial que o resumo seja totalmente coerente com o conteúdo expresso no texto-base (Machado et al., 2005; Costa & Silva, 2013; Bragagnollo, 2017) entre outros.[1] Nessa direção, assume-se a coerência como um dos principais critérios de textualidade, sendo essencial para a produção de sentidos em um texto (Koch & Travaglia, 2021; Van Dijk & Kintsch, 1983). Do ponto de vista da avaliação de uma produção textual, é comum que a coerência seja colocada em primeiro lugar, uma vez que garante que o estudante possa apresentar seu projeto de texto de forma consistente (Costa & Silva, 2013).

[1] Neste trabalho, utilizaremos "texto original", "texto lido", "texto-fonte" e "texto-base" indefinidamente para nos referirmos ao texto que dá origem ao resumo.

DOI: 10.21814/lm.17.1.461

Este trabalho foca na coerência em resumos, mais especificamente no uso adequado de verbos *dicendi* para a expressão do objetivo do texto lido. Na produção do resumo, espera-se que o escritor expresse os pontos principais do texto, o que inclui o objetivo do texto lido, tal como se vê no trabalho de Machado et al. (2005). As autoras ainda afirmam que é importante separar, de um lado, a sumarização, processo necessário ao resumo, e, de outro, o próprio resumo, uma vez que este é um gênero completo, enquanto aquela é uma estratégia necessária para a produção de diferentes textos.

Nesse aspecto, o que vamos focar neste trabalho é o uso de verbos de dizer (i.e., do latim verbos *dicendi*) na formulação do gênero resumo. Partiremos do princípio de que tais verbos têm a função de caracterizar "um discurso reportado" (Clark & Gerrig, 1990, p. 764-805), a partir do modo como o falante compreende esse discurso. Entendemos que os verbos são as peças-chave que dão vida e movimento a um texto/discurso, sobretudo no caso dos resumos. Logo, os verbos *dicendi* são modos de enunciar o que se interpreta do autor do texto-fonte, devendo ser coerentes com aquilo que ali se apresenta. Por isso, nesta pesquisa, tomamos a noção de "coerência" como central para a expressão da enunciação por meio desses verbos. Com isso, esperamos analisar as avaliações desse aspecto por humanos e por modelos de linguagem de grande escala (do inglês, *Large Language Model* - LLM) em resumos.

Diferentes trabalhos têm analisado a avaliação automática de textos como ferramenta auxiliar no ensino (Meira et al., 2023; Rassi & Lopes, 2023). Nesse contexto, o emprego de ferramentas computacionais pode gerar resultados que auxiliam tanto o estudante a refletir sobre seu desempenho de escrita, quanto professores, no processo de ensino e de correção de textos. Nesta pesquisa, assumimos que a avaliação automática de textos auxilia na compreensão mais profunda de textos, ao identificar os elementos-chave que compõem a sua estrutura. No caso dos resumos, pode oferecer indícios de sua qualidade, ao verificar se eles capturam as ideias principais do texto original. Nesse sentido, pode-se analisar se os resultados a que chegamos contribuem para as tarefas de avaliação de texto, como de nota global e *feedback* (Rassi & Lopes, 2023).

Neste momento da história humana, as inteligências artificiais (IAs), materializadas em LLMs, estão em evidência por suas diferentes capacidades de geração e transformação de conteúdo semiótico. Alguns trabalhos mostram suas capacidades e fragilidades na tarefa de lidar com geração ou análise de textos (Paes & Freitas, 2023). Outros, trazem a discussão de como elas podem impactar ou não as produções humanas, sobretudo na escola (Nunes, 2024). Nesse contexto, uma das questões é justamente a capacidade de modelos automáticos produzirem ou avaliarem conteúdos linguísticos, inclusive em ambiente de ensino (Rassi & Lopes, 2023). Tal fato nos levou a questionar: considerando um enunciado específico para resumo de um texto, um LLM poderia sugerir um verbo *dicendi* coerente com o objetivo do texto? Isso se justifica porque os verbos *dicendi*, não são de livre-escolha na composição do resumo: precisam estar diretamente ligados à ideia central emitida pelo texto.

Para chegarmos à resposta da questão, decidimos escolher um texto que fora sugerido para resumo no vestibular da Universidade Federal do Paraná (UFPR) de 2024. Pedimos a especialistas que indicassem os verbos mais coerentes com a interpretação do texto. Depois, solicitamos que os LLMs indicassem também o objetivo do texto por meio de um verbo e comparamos essas indicações com a lista humana. Finalmente, ainda testamos se os modelos poderiam contribuir para tarefas de correção automática, tomando um conjunto de textos para avaliar a escolha dos estudantes. Para realizar os experimentos, levantamos as seguintes hipóteses: *i*) LLMs que identificarem adequadamente o verbo de objetivo do texto-base atribuirão notas também mais adequadas aos resumos; e *ii*) se o LLM realizar sua análise a partir dos verbos sugeridos pelos humanos ele apresentará resultados melhores. Nos dois casos, poderemos entender que o emprego dos verbos *dicendi* pode ser um critério para análise automática de resumos.

Com esse recorte de pesquisa, temos o intuito de contribuir para a discussão das possibilidades de uso de IAs em contextos de produção e avaliação de textos, sobretudo no ensino. Nesse viés, o artigo contribui essencialmente com:

- o enfoque aos verbos *dicendi* na coerência de resumos;
- a análise propriamente dita dos textos, comparando versões humanas com versões de LLMs;
- a reflexão sobre as capacidades e limitações do emprego de IAs na produção e avaliação de textos; e
- o impacto que as IAs podem apresentar no campo da educação.

O restante do artigo está organizado assim: a Seção 2 apresenta os fundamentos da pesquisa; a Seção 3 descreve a metodologia utilizada para os

experimentos; a Seção 4 apresenta os resultados e as discussões; finalmente na Seção 5, concluímos o trabalho e apresentamos pontos a serem ainda investigados.

2. Fundamentos

2.1. Coerência semântica

Do ponto de vista linguístico, a computação contribui significativamente para as possibilidades de análise de texto/discurso, particularmente em sua forma escrita. Nesse sentido, entendemos que o texto escrito é um terreno fértil para análise, especialmente por corresponder ao discurso, considerado um evento comunicativo (Wang & Guo, 2014). Baseando-se nas principais ideias de De Beaugrande & Dressler (1981) sobre este tópico, Wang & Guo (2014) consideram a coerência — uma das sete características desejadas de um discurso — como a qualidade que distingue discursos consistentes que fazem sentido daqueles que não fazem, devido a contradições ou falhas na composição das ideias.

Um discurso semanticamente coerente consiste em agrupamentos ou cadeias de componentes textuais (palavras, frases, etc.) cujos significados compatíveis permitem uma interpretação consistente. Textos semanticamente incoerentes, por outro lado, têm componentes localizados próximos uns dos outros no texto, mas que possuem significados incompatíveis, ou seja, estão semanticamente distantes uns dos outros ou do contexto em que aparecem (e.g., frase, parágrafo, documento como um todo) (Braz Jr & Fileto, 2021). Koch & Travaglia (2021) definem coerência semântica como a compatibilidade de significado entre elementos adjacentes (i.e., local) ou entre todos os elementos em um documento de texto (i.e., global).

Costa & Silva (2013) consideram que a análise de coerência textual requer também a avaliação da situação comunicativa em que os textos aparecem. Na interação que ocorre entre autor e leitor, é preciso que se leve em conta os objetivos de leitura e escrita para que se avalie a adequação. As autoras consideram que o resumo exige uma coerência essencialmente global e situacional. No primeiro caso, por que o conteúdo do resumo precisa ser compatível com as informações do texto-base; no segundo, porque é preciso se considerar a situação de produção dele.

No caso específico que estamos analisando, análogo ao trabalho de Costa & Silva (2013), é preciso considerar que o autor do resumo está em uma situação de avaliação. Ele sabe que o resumo a ser produzido responde a uma situação bem peculiar, já que o leitor desse texto tem papel de "juiz", por avaliar o resumo com base nas informações do texto original. Nesse caso, portanto, o avaliador já conhece o texto-base e, por isso, é capaz de julgar se o resumo está coerente com as informações ali presentes. Nesse sentido, portanto, como as próprias autoras exemplificam com um caso real, vale menos a coerência interna do resumo e mais a coerência deste com o texto-fonte. Em nosso recorte, essa coerência será analisada em termos de maior ou menor relação dos verbos *dicendi* com aquilo que o texto original deixa entender.

2.2. O resumo escolar

A prática social de resumir é comum em diferentes atividades, como relatar uma experiência, contar sobre um filme ou um livro. Nesses casos, o que se espera é que o resumidor dê indicações do que originou o resumo (i.e., falado ou escrito), focando no eixo principal e excluindo detalhes desnecessários. Por isso, a habilidade principal para a produção de um resumo é a capacidade de sumarizar, entendida como a seleção dos tópicos principais daquilo que se pretende resumir. Como gênero, há diferentes funções para o resumo e, por isso, diferentes tipos, tais como escolar, acadêmico, cultural, expandido, crítico, esquemático etc. Ainda que apresentem especificidades, o eixo que os une é a capacidade de condensar, pela escrita, a interpretação global e os principais pontos do original.

No contexto educacional, o resumo escolar, foco deste trabalho, pode ser considerado um instrumento de avaliação de duas habilidades: a leitura, por verificar o grau de compreensão de um texto; e a escrita, por exigir do produtor que apresente essa compreensão em um texto autônomo, coeso e coerente. Além disso, pode servir como um instrumento de estudo, por contribuir para a fixação de ideias e conceitos enunciados nos textos lidos. Bragagnollo (2017) ressalta que o resumo tem relevância para o ensino por estar diretamente relacionado a práticas sociais diversas, incluindo o próprio contexto de aprendizagem. Para a autora, o resumo escolar tem função singular por ser usado como instrumento de avaliação de compreensão e produção textual. Bragagnollo (2017) afirma ainda que a elaboração do resumo necessariamente deve trazer informações fiéis ao texto original, focando naquilo que é essencial no texto-base.

Destacamos que o resumo se caracteriza pela paráfrase, uma vez que o que se espera é um texto novo por parte do produtor do resumo, mas que seja originário deste. Essa paráfrase não deve ser uma cópia ou alteração simples do texto original por sinônimos. Ela deve se configurar, sobretudo, como uma exposição do sentido construído na leitura do texto-base, ao, por exemplo, explicar uma passagem difícil daquilo que seu leu (Bragagnollo, 2017). Se é assim, o que menos se espera é que o resumo seja uma sumarização pura e simples, um "copia e cola" das informações lidas.

Nessa perspectiva, vamos assumir que o resumo decorre do próprio texto original, uma vez que é organizado a partir deste (Machado et al., 2005). Por isso, ele é necessariamente acarretado pelo texto-base e suas asserções são inferidas dali. Tais condições permitem que se trate de um resumo "coerente" ou não em relação ao original, sobretudo no caso do resumo escolar, foco deste trabalho. Desta forma é importante que o autor do resumo entenda que, ao

> resumir o que o autor do texto de base diz, estamos, de fato, interpretando o seu agir, atribuindo-lhe a efetivação de determinados atos. Ou seja, quando dizemos que o autor nega algo, afirma algo, questiona algo, estamos inferindo, através de nossa compreensão do conteúdo do texto, que o autor está realizando esses atos (Machado et al., 2005, p. 98-99).

Como se vê, a interpretação crucial do texto, indicada no resumo, passa necessariamente pela escolha adequada do verbo *dicendi* para aquela situação. Afinal, nesse movimento de resumir fica marcado o que o autor do texto-base teria dito, por meio da compreensão do autor do resumo.

2.3. A importância dos verbos *dicendi*

Na gramática, os verbos *dicendi* são aqueles que introduzem o discurso direto ou indireto de outrem, geralmente usados em conjunto com uma oração subordinada. Eles representam ações de comunicação, como argumentar, defender, criticar, opinar, sugerir, falar, entre tantas outras. Esses verbos podem ser encontrados em textos que visam relatar informações e opiniões de terceiros de forma objetiva e precisa, como, por exemplo, nos textos dos gêneros resumo e resenha. Isso acontece porque eles facilitam a referenciação a ideias e conteúdos apresentados por outra pessoa, servindo como ponte entre o autor do texto original e o leitor do resumo ou da resenha.

Tratando especificamente do resumo escolar, Campos & Ribeiro (2013, p. 37) sustentam que, em seu processo de produção, "deve-se considerar que há um texto escrito por A (i.e., autor do texto-base) que é retextualizado por B (i.e., produtor do resumo)". Essa retextualização é perceptível ao leitor do resumo a partir do momento em que o seu produtor faz a distinção entre as vozes que sempre estarão presentes em textos desse gênero, sendo elas a do autor do texto original e a do produtor do resumo. Separar essas duas vozes é essencial no processo de escrita, pois, assim, evita-se a atribuição de ideias resumidor quando, na verdade, elas pertencem ao autor do texto original. Nesse contexto, são os verbos *dicendi* que irão desempenhar um papel decisivo para indicar a autoria de cada ideia apresentada no resumo.

De acordo com Machado et al. (2007, p. 49), com o intuito de separar essas vozes, apresentamos as ideias do autor do texto-base como se ele "estivesse realizando vários tipos de atos". Esses "atos" do autor do texto original devem ser interpretados pelo produtor do resumo, pois é a partir dessa interpretação que ele irá escolher o verbo *dicendi* mais adequado ao mencionar as ações de comunicação do autor do texto-base. Em uma situação hipotética, se o objetivo é resumir um artigo de opinião, escolher um verbo *dicendi* como "fala" ou "diz" para apresentar um argumento do autor não é adequado nesse caso, sendo preferível optar por verbos como "critica", "opina" ou "argumenta", por exemplo. Na mesma situação, de texto argumentativo, parece incoerente utilizar o verbo "informar" para indicar o objetivo do texto. Logo, podemos dizer que o uso dos verbos *dicendi* não se trata de algo meramente estilístico, pois, assim como lembrado por Bragagnollo (2017), eles representam o nível de compreensão e de interpretação do produtor do resumo acerca do texto original, o que faz com que esse tipo de verbo seja, na verdade, uma ferramenta linguística.

Por ter esse caráter de separar as vozes do autor do texto-base e do produtor do resumo, os verbos *dicendi* são extremamente importantes para esse gênero, tanto a ponto de descaracterizá-lo se não forem utilizados na produção do texto, visto que "sem esse aspecto o leitor não consegue visualizar a autoria das informações apresentadas no resumo" (Bragagnollo, 2011, p. 126). À vista disso, sabe-se que esses verbos ajudam a esclarecer a origem das informações apresentadas, evitando, dessa forma, ambiguidades e garantindo que as ideias expostas no resumo sejam

atribuídas corretamente ao autor do texto original. Assim, o uso de verbos *dicendi* em textos do gênero resumo sugere a postura do autor do texto original, tanto em relação àquilo que textualmente se lê, quanto àquilo que se depreende da interpretação geral do texto, como o "objetivo do autor".

No recorte de uma situação de produção de resumo para o ambiente de avaliação (i.e., vestibular 2024 da UFPR), há alguns elementos apontados por Costa & Silva (2013) que poderíamos considerar aqui. Primeiro, é preciso que haja uma coerência entre as informações do texto-fonte e o resumo, sem que este acrescente ideias ou opiniões àquele. Depois, o avaliador analisa a pertinência entre o que diz o texto-fonte e o resumo, o que apenas uma sumarização não faz. Em ambos os casos, as autoras consideram que é preciso se considerar uma coerência global na produção, para que a avaliação seja realizada.

Nesta pesquisa, consideramos que os verbos *dicendi* sejam cruciais para a métrica de avaliação de resumos, seja por humanos, seja por LLMs. Assim, além de critérios básicos exigidos para a boa avaliação de um resumo, como objetividade, clareza e, sobretudo, respeito ao conteúdo do texto original, consideramos que o emprego dos verbos *dicendi* é uma estratégia linguística fundamental, pois o seu uso garante que a voz do autor original seja devidamente representada e que o leitor compreenda as ideias principais do texto. Consequentemente, as chances de o resumo produzido apresentar coesão e coerência são maiores conforme o nível de adequação dos verbos *dicendi* escolhidos para referenciar as ações do autor do texto-base.

2.4. Modelos de Linguagem de Grande Escala

Atualmente, modelos de linguagem de grande escala (do inglês, *Large Language Model* - LLM) como o *Generative Pre-trained Transformer*(GPT) (Radford et al., 2018) estendem as capacidades dos modelos de linguagem na geração de texto. LLMs são treinados em grandes quantidades de dados de texto e são capazes de gerar texto semelhante ao humano (Kasneci et al., 2023). Usar esses modelos, por meio de *prompts* apropriados, pode ser uma estratégia eficaz para orientar a avaliação e melhoria da coerência do texto.

O termo "prompt" se refere a uma instrução ou entrada dada a um modelo de linguagem solicitando uma resposta ou continuação de texto. Um *prompt* é usado para direcionar o modelo de linguagem para gerar texto relevante ou executar uma tarefa posterior (Radford et al., 2018). Os *prompts* podem consistir em perguntas, descrições de problemas, fragmentos de texto ou qualquer tipo de entrada que ajude a moldar a resposta gerada pelo modelo. (Radford et al., 2018). No processamento de linguagem natural (PLN), esse novo paradigma segue o processo de "pré-treinamento, *prompt* e predição" (Liu et al., 2023). Usar *prompts* permite que o comportamento do LLM seja guiado pelo fornecimento de informações contextuais ou exemplos relevantes relacionados à tarefa.

Apesar da enorme repercussão desses modelos, sabe-se que eles não são criativos, justamente por terem base probabilística (Paes & Freitas, 2023). A manipulação de hiperparâmetros, como a temperatura, levanta a questão de até que ponto a criatividade pode ser influenciada. No contexto da presente pesquisa, podemos entender "criatividade" como a capacidade de gerar diferentes respostas adequadas ao ambiente de aplicação, o que corresponde à noção de diversidade lexical (Martins, 2016). Além disso, a relação entre a alteração de hiperparâmetros e a manifestação de comportamentos criativos é um tema ainda em aberto, demandando investigações mais profundas que extrapolem os experimentos com *prompts*. Nesse sentido, este trabalho levanta a hipótese de que esses modelos darão respostas e farão avaliações aquém daquelas esperadas por humanos, uma vez que é mais provável se limitarem ao léxico relativo ao contexto.

2.5. Trabalhos relacionados

Embora a literatura considere fundamental o emprego de verbos *dicendi* na escrita de um resumo escolar, pelas razões já mencionadas, os trabalhos sobre o tema na área não se debruçam sobre o assunto. Aqueles que encontramos: focam sobretudo nas dificuldades dos estudantes em realizar a escrita desse gênero (Silva & Nascimento, 2016; Biral, 2003; Souza, 2017); investigam a concepção de professores sobre o gênero resumo (Lima & Alves, 2017); apresentam algumas intervenções no ambiente escolar, as quais podem melhorar a escrita desse tipo de texto (Rocha, 2015); e analisam o gênero no contexto de vestibular (Bicudo & Hila, 2015).

Desses mencionados, Souza (2017) desdobra um pouco a importância dos verbos *dicendi* ao apresentar um modelo sociorretórico para a escrita desse gênero. Para o autor, refletir sobre a escolha do verbo adequado a se usar é uma característica essencial para o gênero. Depois dele,

apenas (Bicudo & Hila, 2015) fazem um apanhando do uso de verbos *dicendi* das 50 melhores provas de um vestibular no Paraná. Os autores mostram que, nesse corpus, os candidatos conseguiram encontrar os verbos adequados para a indicação dos atos do autor do texto original. Nos demais textos pesquisados, os verbos *dicendi* aparecem como parte do gênero, mas sem grande destaque. Por isso, o presente trabalho é também uma contribuição para o estudo do gênero, à medida que foca em um aspecto pouco explorado desse conhecido gênero escolar.

Nesta pesquisa, não temos o objetivo de analisar resumos gerados por modelos de linguagem (e.g., BERT), LLMs (e.g., LLaMA, GPT, Gemini e Sabiá) ou outras técnicas, tal como se vê em outros trabalhos (Inácio, 2021; Souza, 2022; Paiola, 2022; Paes & Freitas, 2023). Aliás, a principal razão para não apresentarmos trabalhos sobre resumo relacionados à computação é que em PLN o resumo é tratado em termos de sumarização. Mas, em se tratando de resumo escolar, foco deste trabalho, a sumarização é somente uma parte do processo de escrita.

3. Metodologia

Nos experimentos, utilizamos um texto original e 24 resumos deste texto. O texto original tem origem no vestibular da UFPR de 2024.[2] Já os 24 resumos são fruto de um projeto de extensão que a UTFPR desenvolve em parceria com o Cursinho Solidário, em Curitiba.[3] A Universidade oferece aulas e correção de redações. Em troca, os alunos são convidados a ceder os textos para pesquisas. Esses textos formam o projeto Digitus, em construção desde 2018 na UTFPR, mas ainda sem recursos para organização do banco de textos. Do projeto, além de docentes da UTFPR, participam estudantes bolsistas e voluntários da UTFPR e da PUCPR, os quais recebem treinamento para correção das redações, focando no *feedback* aos estudantes do Cursinho, com vistas a estimular a reescrita.[4]

Com o corpus em mãos, desenvolvemos um fluxo de trabalho que partiu do texto-base. Em seguida, solicitamos a avaliadores humanos que indicassem um verbo relativo ao objetivo do texto. Essa mesma tarefa foi solicitada a quatro LLMs (i.e., LLaMA, GPT, Gemini e Sabiá). Finalmente, foi atribuída uma nota geral (por humanos e pelos LLMs) a um conjunto de resumos, considerando, sobretudo a coerência dos verbos *dicendi* sugeridos anteriormente.

A Figura 1 ilustra o fluxo de trabalho de avaliação do texto-base e dos resumos. O Fluxo está dividido em duas partes: avaliação do texto-base e avaliação do texto resumo. A primeira parte se inicia com a análise do texto original, tanto por avaliadores humanos quanto por modelos de linguagem. Ambos os métodos visam extrair o verbo principal que define o objetivo do texto original. Na segunda parte, esses verbos são utilizados como referência para avaliar os textos resumos, tanto por meio de uma análise humana, automática e humana/automática. A comparação entre os verbos do texto original e dos resumos permite avaliar a qualidade e a pertinência dos resumos gerados. A análise final dos resultados, incluindo métricas de desempenho, possibilita a avaliação da eficácia do fluxo de trabalho e a identificação de áreas para melhoria.

O fluxo de trabalho da Figura 1 nos permitiu executar 3 experimentos distintos para avaliar o texto-base e os resumos combinando avaliação humana e LLMs. No primeiro experimento na tarefa "Avaliação humana do texto-base" pedimos a 5 avaliadores humanos, da área de Letras, que indicassem o verbo (verbo') que melhor indicasse o objetivo geral do texto. Entendemos que esse número de avaliadores ofereceria uma dimensão maior de possibilidades, uma vez que a transposição de um enunciado pode ser feita de diferentes maneiras e todas de forma correta. Depois na "Avaliação humana dos textos dos resumos", pedimos a duas avaliadoras, vinculadas ao projeto de coleta de dados, que corrigissem os 24 resumos. Solicitamos que corrigissem o texto de forma global, mas com um foco maior no uso adequado de verbos *dicendi*. Destacamos que essas avaliadoras fazem parte do projeto de correção e coleta de dados, o que garante uma qualidade na correção. Um terceiro avaliador, o coordenador do mesmo projeto, fez uma última correção nos textos, com o intuito de dirimir as possíveis divergências. Ao selecionar pessoas diferentes para as tarefas, nossa intenção era tornar o processo o mais flexível, considerando as múltiplas possibilidades de interpretação e de escrita que se verifica no resumo.

No segundo experimento na tarefa "Avaliação automática do texto-base"pedimos a LLMs para analisar o texto-base com o "Prompt de avaliação do texto-base". O *prompt* tem a finalidade de identificar o verbo (verbo") que melhor repre-

[2]Texto disponível em: https://servicos.nc.ufpr.br/documentos/ps2024/provas/2fase/001-Compreensao_e_Producao_de_Texto.pdf

[3]Resumos disponíveis em: https://github.com/osmarbraz/cohewl/tree/main/dataset/coheatbee

[4]O projeto para coleta recebeu parecer favorável do Comitê de Ética em Pesquisa da Instituição, número CAAE 88328218.6.0000.5547.

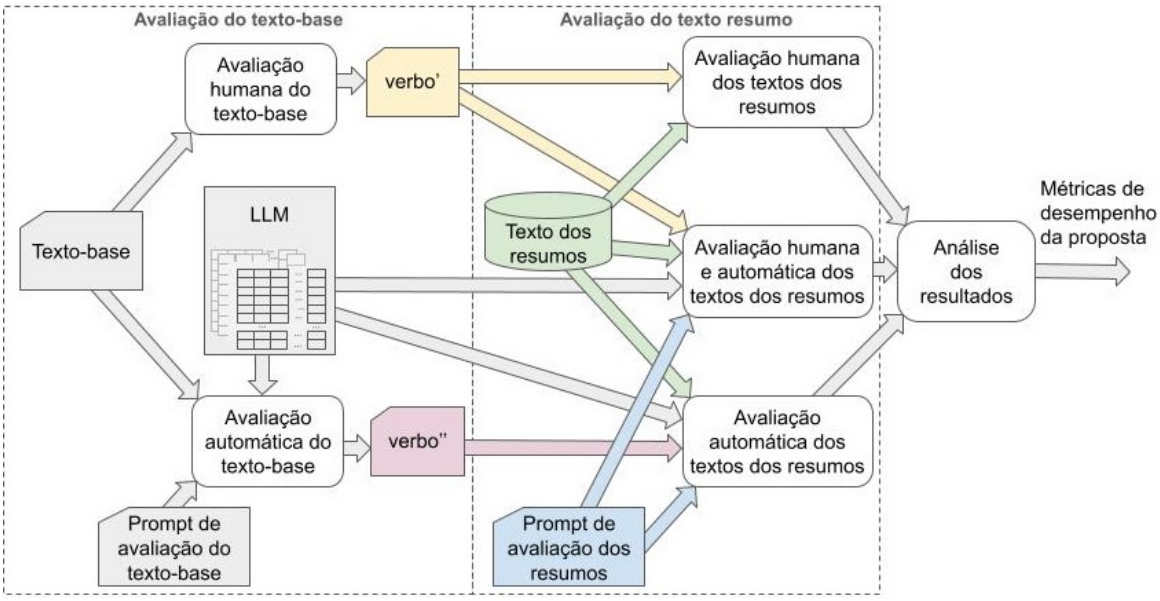

Figura 1: Fluxo de trabalho proposto

senta o objetivo do texto-base. A fim de captar uma parte da flexibilidade de interpretação e escrita, solicitamos a cada LLM uma lista cinco sinônimos do verbo identificado. Estes sinônimos deveriam ser semanticamente próximos e capazes de substituir o verbo original sem alterar significativamente o sentido do texto. Depois a tarefa "Avaliação automática dos textos dos resumos" analisamos a qualidade do resumo, comparando os verbos utilizados nele com os verbos relacionados ao objetivo do texto original. A ideia foi verificar se: *i*) os resumos capturaram a essência do texto original por meio os verbos escolhidos; e *ii*) se eles foram coerentes com o objetivo indicado. Inicialmente solicitamos ao LLM para identificar os verbos presentes tanto no texto original quanto nos resumos. Para cada verbo identificado, solicitamos que os LLMs atribuíssem uma nota quanto a sua coerência com o texto. Por fim, solicitamos que os LLMs atribuíssem uma nota para cada resumo, considerando a relevância e a adequação dos verbos utilizados.

Por fim no terceiro experimento combinamos a tarefa "Avaliação humana do texto-base" com a tarefa "Avaliação automática dos textos dos resumos", desta forma utilizamos os verbos identificados por seres humanos para auxiliar na avaliação automática dos resumos pelos LLMs.

Os resultados dos 3 experimentos são comparados na tarefa "Análise dos resultados" utilizando métricas de correlação (Coeficiente Pearson) entre os avaliadores humanos e entre a avaliação humana e automática. Essa métrica permitem avaliar a qualidade dos resumos gerados e a eficácia do LLM utilizado pela correlação das notas.

Além dos 3 experimentos distintos, realizamos experimentos automatizados e combinados utilizando 4 LLMs. Para estes LLMs usamos as configurações de hiperparâmetros padrão, incluindo *temperature* e *top_p*. Esses LLMs incluem LLaMA-3.1 70B Instruct[5] (*temperature* = 0,75 e *top_p* = 0,9), GPT-3.5[6] (*temperature* = 0,7 e *top_p* = 1,0), Gemini[7] (*temperature* = 1 e *top_p* = 0,95) e Sabiá-3 [8] (parâmetros *temperature* e *top_p* não divulgados). Os LLMs possuem uma limitação intrínseca: a capacidade de processar uma quantidade finita de tokens em uma única janela de contexto (i.e., limite de informação que o LLM consegue processar). Um token, a unidade básica de processamento de texto para um LLM, pode corresponder a uma palavra, parte de uma palavra ou um caractere. Considerando essa restrição, optamos por avaliar os modelos em resumos, textos mais concisos nos quais a coerência entre as partes é mais evidente. As capacidades de processamento em tokens dos modelos são: GPT-3.5 (16.385 tokens), Gemini e Sabiá-3 (32.000 tokens) e LLaMA 3.1 70B Instruct (128.000 tokens).

4. Experimentos e Resultados

Nesta seção, adicionamos detalhes aos três experimentos e apresentamos os seus resultados. Ao final, analisamos e discutimos os resultados.

[5] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
[6] https://www.chatgpt.com/
[7] https://gemini.google.com/
[8] https://www.maritaca.ai/

4.1. Avaliação humana

No primeiro experimento na tarefa "Avaliação humana do texto-base" solicitamos a 5 avaliadores humanos que indicassem verbos coerentes ao objetivo central do texto, estes verbos sem repetição são: *apresentar*, *destacar*, *explicar* e *informar*. Esta lista é coerente com o texto-base, que é uma notícia de divulgação científica do Jornal da Unicamp.

Na sequência, na tarefa "Avaliação humana dos textos dos resumos" avaliadores humanos atribuíram uma nota de 0,0 até 10,0 aos resumos escolares. A nota considera o uso dos verbos da identificados na tarefa anterior ou de seus sinônimos. A Tabela 1 apresenta a matriz de correlação de Pearson entre as notas dos 3 avaliadores. Essa matriz quantifica a força e a direção da relação linear entre os pares de avaliadores, com base na nota atribuída por eles ao texto resumido considerando o verbo central identificado de no passo 1. Números mais próximos de 1 revelam uma correlação maior.

A Tabela 1 mostra que os avaliadores humanos apresentam uma correlação linear forte (maior que 0,70), sugerindo pouca discrepância entre si.

4.2. Avaliação automática

No segundo experimento, utilizamos somente LLMs para avaliar o texto-base e os textos resumidos. Na tarefa de "Avaliação automática do texto-base", um LLM identifica o verbo central do texto. Na sequência, na tarefa "Avaliação automática dos textos dos resumos", o verbo central identificado é utilizado para avaliação dos resumos. Para cada tarefa desenvolvemos um *prompt* específico para a avaliação do texto-base e do resumo. Todos os *prompts* são do tipo *zero-shot*, ou seja os modelos não recebem exemplos ou definições prévias de treinamento ao contrário do tipo *few-shot* que são expostos a exemplos e definições.

O *prompt* usado para solicitar ao LLM para avaliar o texto-base é transcrito a seguir. O LLM ao analisar este *prompt* deve considerar somente o texto compreendido entre os marcadores "####", substituindo <TEXTO> pelo texto-base. A seguir o *prompt* é divido em 4 passos. O passo 1 identifica a fonte e autoria do texto para auxiliar o passo 2. O passo 2 identifica o verbo principal do texto considerando a origem e fonte do texto do passo 1. O passo 3 apresenta os verbos sinônimos ao verbo principal do passo 2. Isso foi feito para simular as diferentes possibilidades de ocorrência, tal como pode ocorrer com humanos. O passo 4 instrui o LLM a avaliar a coerência semântica dos cinco verbos sinônimos em relação ao verbo principal que representa o objetivo central do texto. A coerência semântica nesse contexto se refere a proximidade de significado entre os verbos. Ou seja, analisa se os sinônimos propostos no passo 3 realmente expressam a mesma ideia central do texto, com a mesma intensidade e nuance. Em "Resposta:" será apresentado o resultado do LLM.

```
Considerar para a tarefa a seguir somente o texto
que está entre #### e ####.
####
"<TEXTO>"
####
Tarefa: Você é um assistente útil responsável
   pela análise de coerência semântica de textos.
   Sua tarefa é analisar um texto seguindo os
   passos abaixo:
Passo 1. Apresente a fonte e a autoria do texto
   acima.
Passo 2. Considerando a fonte e autoria do passo
   1, apresente um verbo no infinitivo que
   represente o objetivo central do texto.
Passo 3. Considere que sinônimos são palavras
   semanticamente próximas. Apresente 5 (cinco)
   verbos sinônimos do verbo no infinitivo
   apresentado no passo 2.
Passo 4. Agora, atribua uma nota para os verbos
   sinônimos apresentados no passo 3,
   considerando a coerência deles com o verbo do
   objetivo da autoria apresentado no passo 2.
   Atribua notas de 1 a 10, sendo 1 (um) para o
   verbo menos coerente e 10 (dez) para o mais
   coerente.
Resposta:
```

No experimento com quatro LLMs LLaMA-3.1 70-B Instruct, GPT 3.5, Gemini e Sabiá-3, foram identificados os seguintes verbos predominantes em suas respectivas respostas: "analisar" (LLaMA e GPT), "informar' (Gemini) e "alertar" (Sabiá). Dos verbos identificados, apenas o Gemini apresentou um dos verbos indicados pelos humanos. Já o LLaMa e o GPT extraíram um verbo que aparece no texto-base, enquanto o Sabiá apresenta um verbo com alguma relação com o texto, mas que não é seu objetivo.

O *prompt* usado para solicitar ao LLM para avaliar o texto resumido é transcrito a seguir. O verbo previamente identificado no *prompt* anterior será reaproveitado. Este verbo será inserido no local marcado por <VERBO>. Ao analisar este *prompt*, o LLM deve considerar somente o texto entre os marcadores "####", substituindo o marcador <TEXTO> pelo texto do resumo a ser analisado. A seguir o *prompt* é divido em 3 passos. Em "Resposta:" será apresentado o resultado do LLM.

Avaliadores	Avaliador 1	Avaliador 2	Avaliador 3
Avaliador 1	-	0,84	0,74
Avaliador 2	0,84	-	0,82
Avaliador 3	0,74	0,82	-

Tabela 1: Matriz de correlação de Pearson da nota atribuída pelos avaliadores humanos aos resumos

```
Considerar para a tarefa a seguir somente o
texto que está entre ####  e ####.
####
"<TEXTO>"
####
Tarefa: Você é um assistente útil responsável
  pela análise de coerência semântica de textos.
  Sua tarefa é analisar um texto seguindo os
  passos abaixo:
Passo 1. Liste a ocorrência do verbo <VERBO> ou
  de seus sinônimos.
Passo 2. Agora, atribua uma nota para os verbos
  listados no passo 1, considerando a coerência
  semântica das palavras entre si. Atribua uma
  nota de 1.0 a 10.0, sendo 1.0(um) para o verbo
  menos coerente e 10.0(dez) para o mais
  coerente.
Passo 3. Considerando as notas dos verbos do
  passo 2, atribua uma nota ao texto, sendo 1.0
  (um) nota mais baixa e 10.0 (dez) a mais alta.
Resposta:
```

A Tabela 2 apresenta a matriz de correlação de Pearson entre as notas geradas de forma automática pelos quatro LLMs: LLaMA-3.1 70-B Instruct, GPT 3.5, Gemini e Sabiá-3.

Observa-se na Tabela 2 uma correlação positiva de fraca para moderada (0,37) entre LLaMa-3.1 e GPT-3.5, a maior correlação entre os modelos. Depois deles, a correlação entre o Gemini e o GPT-3.5 foi a segunda mais forte (0,33). No restante, nota-se uma correlação negativa moderada entre o LLaMA-3.1 70B e o Sabiá-3, sugerindo que, em geral, esses modelos avaliaram os resumos de forma oposta. No entanto, a correlação negativa entre o GPT-3.5 e o Sabiá-3, assim como entre o Gemini e o Sabiá-3, sugere que este último modelo apresentou um padrão de avaliação distinto dos demais. Em resumo, os resultados indicam que os LLMs possuem diferentes critérios para avaliar a qualidade dos resumos e que a concordância entre eles é divervariável.

4.3. Avaliação combinada humana e automática

O terceiro experimento combina a saída da tarefa "Avaliação humana do texto-base" do primeiro experimento com a tarefa "Avaliação automática dos textos dos resumos" do segundo experimento. O *prompt* da avaliação automática dos resumos possui uma lista de verbos ao invés de somente um.

O *prompt* usado para solicitar ao LLM para avaliar o texto dos resumos combinado com humano foi construído a partir do *prompt* utilizado na tarefa "Avaliação automática dos textos de resumos". No *prompt* transcrito a seguir o LLM deve considerar somente o texto entre os marcadores "####" e a palavra <TEXTO> deve ser substituída pelo texto do resumo a ser analisado. Os verbos da "Avaliação humana do texto-base" (i.e., *apresentar*, *destacar*, *explicar* e *informar*) são utilizados para preencher <LISTA_VERBOS>. Isso foi feito para combinar a avaliação humana com a automática dos LLMs. A seguir o *prompt* é divido em 3 passos. Em "Resposta:" será apresentado o resultado do LLM.

```
Considerar para a tarefa a seguir somente o texto
que está entre #### e ####.
####
"<TEXTO>"
####
Tarefa: Você é um assistente útil responsável
  pela análise de coerência semântica de textos.
  Sua tarefa é analisar um texto seguindo os
  passos abaixo:
Passo 1. Liste a ocorrência dos verbos
  <LISTA_VERBOS>.
Passo 2. Agora, atribua uma nota para os verbos
  listados no passo 1, considerando a coerência
  semântica das palavras entre si. Atribua uma
  nota de 1.0 a 10.0, sendo 1.0(um) para o verbo
  menos coerente e 10.0(dez) para o mais
  coerentes.
Passo 3. Considerando as notas dos verbos do
  passo 2, atribua uma nota ao texto, sendo 1.0
  (um) nota mais baixa e 10.0 (dez) a mais alta.
Resposta:
```

A Tabela 3 apresenta a matriz de correlação de Pearson entre as notas geradas de forma combinada entre humanos e automática pelos quatro LLMs: LLaMA-3.1 70-B Instruct, GPT 3.5, Gemini e Sabiá-3. Essa matriz quantifica a força e a direção da relação linear entre os pares de LLMs, com base na nota atribuída por eles ao texto resumido considerando o verbo central identificados por seres humanos.

Na Tabela 3 a maioria dos valores de correlação na matriz são próximos de zero ou possuem um valor absoluto baixo. Isso sugere que, de forma geral, não há uma forte relação linear entre as notas atribuídas pelos diferentes LLMs.

LLMs	*LLaMA-3.1*	*GPT 3.5*	*Gemini*	*Sabiá-3*
LLaMA-3.1	-	0,37	-0,02	-0,36
GPT-3.5	0,37	-	0,33	-0,53
Gemini	-0,02	0,33	-	-0,26
Sabiá-3	-0,36	-0,53	-0,26	-

Tabela 2: Matriz de correlação de Pearson das notas atribuídas pelos LLMs aos resumos

LLMs	*LLaMA-3.1*	*GPT 3.5*	*Gemini*	*Sabiá-3*
LLaMA-3.1	-	-0,13	0,08	-0,17
GPT-3.5	-0,13	-	0,33	0,19
Gemini	0,08	0,33	-	-0,11
Sabiá-3	-0,17	-0,19	-0,11	-

Tabela 3: Matriz de correlação de Pearson das notas atribuídas pelos LLMs com os verbos identificados por humanos

Isso significa que, nesses pares de LLMs, quando um deles atribui uma nota mais alta, o outro tende a atribuir uma nota um pouco mais baixa. Uma única correlação positiva mais significativa que se vê 0,33 é entre o Gemini e o GPT-3.5. Ainda assim, é uma correlação fraca para os padrões dessa medida estatística.

4.4. Discussão

Comecemos a análise dos dados recuperando a pouca divergência entre os avaliadores humanos observada na Tabela 1. Como se pontuou na Seção 3, esses avaliadores participam de um mesmo projeto de apoio de correção de textos. Tal fato pode garantir que haja uma maior correlação nas notas atribuídas, o que deveria ser comum em processos avaliativos (Rassi & Lopes, 2023; Cândido & Webber, 2018).

Quanto ao experimento de indicação do objetivo, o resultado cumpriu com a hipótese levantada na Seção 2.4. Naquele trecho, sugerimos que os modelos poderiam ter um desempenho abaixo dos humanos, justamente pela falta de criatividade e sua estrutura probabilística. Como se viu, apenas o Gemini apresentou um verbo igual (*informar*) àqueles que os humanos mais sugeriram para o objetivo do texto-base, enquanto os demais extraíram do conteúdo do texto-base para apresentar suas sugestões. E essa mesma falta de criatividade parece explicar os dados da Tabela 2. Ao contrário do que ocorreu com os avaliadores humanos, que divergiram menos entre si, os LLMs tiveram uma baixíssima correlação relativa à atribuição de notas. A Tabela 3 sugere que a inserção de dados que não estão nos textos (i.e., verbos indicados por humanos) piora a correlação entre os LLMs. Isso nos faz supor que, quando eles apresentam um resultado e isso é tomado como *prompt* para a próxima tarefa o resultado é melhor (Tabela 2).

A Tabela 4 apresenta um comparativo detalhado das estatísticas de notas obtidas por diferentes avaliadores humanos, LLMs de forma automática ("Auto") e combinada ("Comb"). As métricas analisadas incluem média, desvio padrão, moda, valor máximo e mínimo, permitindo uma análise aprofundada do desempenho de cada modelo e avaliador.

Na Tabela 4 percebe-se que os LLMs, em geral, atribuem notas maiores em comparação aos avaliadores humanos. Seu desempenho também é superior, com médias de notas mais altas e menor dispersão dos dados. O Gemini se destacou com a maior média e o menor desvio padrão. Chama a atenção que, no experimento combinado, GPT-3.5 e Sabiá-3 apresentaram médias mais próximas dos avaliadores humanos, os quais parecem realizar uma avaliação mais global (quando se observa as médias, as notas maiores e as menores). Por outro lado, também nesses casos, esses modelos tiveram um alto desvio padrão. Apesar disso, o GPT-3.5 se mostrou mais próximo das médias dos humanos, sobretudo do Avaliador 3 (i.e., moderador).

A primeira hipótese que levantamos era que de o LLM que identificasse o verbo de objetivo correto corrigiria melhor as redações. Como se vê na Tabela 4, isso não se confirma com o Gemini, que foi o único a sugerir o verbo *informar*, igual os humanos. Aliás, depois da péssima análise do Sabiá-3, o Gemini foi o LLM que apresentou as maiores divergências de média, moda e nota mínima em relação aos avaliadores humanos. Logo, a previsão não foi corroborada pelos dados.

Avaliadores e LLMs	Média	Desvio Padrão	Moda	Máximo	Mínimo
Avaliador 1	5,38	2,08	5,00	9,00	2,00
Avaliador 2	5,67	2,10	6,00	9,00	2,00
Avaliador 3	6,04	2,10	8,00	9,50	1,00
LLaMA-3.1 Auto	8,37	1,72	8,50	10,00	1,00
GPT-3.5 Auto	8,55	0,96	8,50	10,00	7,00
Gemini Auto	9,30	0,49	9,50	10,00	8,50
Sabiá-3 Auto	2,60	4,15	0,00	9,50	0,00
LLaMA-3.1 Comb	8,02	1,95	8,50	10,00	0,00
GPT-3.5 Comb	6,73	3,09	9,00	9,50	0,00
Gemini Comb	8,71	0,72	9,00	10,00	7,50
Sabiá-3 Comb	6,33	3,79	0,00	10,0	0,00

Tabela 4: Estatísticas das notas atribuídas por Humanos e LLMs

A segunda hipótese levantada era de que, se o LLM realizasse sua análise a partir dos verbos indicados pelos humanos, o resultado seria melhor. Nesse caso, de fato o que se vê na mesma Tabela 5, no experimento combinado, é uma aproximação maior dos LLMs em relação às notas dos humanos. Logo, a hipótese foi parcialmente corroborada, necessitando de mais investigação para sua checagem completa.

A Tabela 5 apresenta o percentual de acerto das notas atribuídas pelos LLMs de forma automática ("Auto") e combinada ("Comb") ao serem comparadas ao terceiro avaliador, considerando uma variação nas notas. Os resultados foram medidos em função da igualdade e de quatro diferentes intervalos de tolerância (i.e., ±0,5, ±1,0, ±1,5 e ±2,0) em relação à nota do avaliador.

Na Tabela 5, tomamos as notas do moderador (Avaliador 3) e verificamos o percentual de acerto dos LLMs. Sabe-se que a divergência é comum em processos avaliativos e os trabalhos em PLN que se dedicam à avaliação automática têm na discrepância um ponto nevrálgico (Cançado et al., 2020; Rassi & Lopes, 2023; Cândido & Webber, 2018). Os dados da Tabela 5 revelam alguns pontos cruciais e que jogam luz sobre a possibilidade de os verbos *dicendi* serem, de fato, um elemento interessante para a análise de resumos.

No experimento combinado, o percentual de acertos dos modelos foi superior ao experimento original (Auto), inclusive na igualdade de notas "=". Tal fato corrobora a hipótese levantada na introdução, de que o emprego de verbos indicados por humanos na análise traria melhores resultados. Além disso, à medida que a variação de notas aumenta há uma significativa melhoria já no experimento original, chegando a 54% de acertos para o GPT-3.5 e o Sabiá-3. O experimento combinado, no entanto, traz números surpreendentes de acertos já na variação de 1,5 ponto. Quando se observa a divergência de até dois pontos, os modelos (i.e., com exceção do Sabiá-3) têm um desempenho bastante satisfatório, atingindo quase 71% de acerto das notas. No cômpito das notas atribuídas, a diferença de 2 pontos é equivalente a 20%. Tais dados sugerem que a participação humana na elaboração dos critérios de correção (i.e., *prompts*) é indispensável, uma vez que o experimento combinado apresentou melhores resultados. Isso é equivalente ao treinamento que se faz também com corretores humanos, uma vez que, quanto mais treinados, menos tendem a divergir (Rassi & Lopes, 2023).

Por fim, voltando aos pontos iniciais do artigo, podemos destacar que:

- o enfoque aos verbos *dicendi* na coerência de resumos pode ser um tema de maior investigação na área; os resultados mostram que é um elemento linguístico a se investir, quer no ensino, na pesquisa e na avaliação de resumos;
- a análise propriamente dita dos textos, comparando versões humanas com versões de LLMs mostrou que a avaliação humana é a mais adequada para o tipo de texto analisado;
- os resultados ajudam na reflexão sobre as capacidades e limitações do emprego de IAs na produção e avaliação de textos, visto que, além das avaliações humanas terem superado a dos LLMs, a participação humana na elaboração dos critérios ainda tem um peso enorme nos resultados apresentados pelos LLMs; e
- há um potencial impacto das IAs no campo da educação, mas algumas tarefas podem ser repensadas para que elas sejam utilizadas mais como ferramentas do que substitutas da função humana (Paes & Freitas, 2023). Nesse

LLMs	**Intervalos de tolerância das notas**				
	=	**±0,5**	**±1,0**	**±1,5**	**±2,0**
LLaMA-3.1 Auto	0,00%	20,83%	29,17%	33,33%	45,83%
GPT-3.5 Auto	4,17%	16,67%	29,17%	37,50%	54,17%
Gemini Auto	0,00%	4,17%	8,33%	29,17%	33,33%
Sabiá-3 Auto	4,17%	16,67%	29,17%	37,50%	54,17%
LLaMA-3.1 Comb	4,17%	20,83%	45,83%	54,17%	62,50%
GPT-3.5 Comb	4,17%	16,67%	37,50%	50,00%	70,83%
Gemini Comb	0,00%	12,50%	33,33%	45,83%	50,00%
Sabiá-3 Comb	8,33%	16,67%	25,00%	29,17%	29,17%

Tabela 5: Percentual de acerto das notas dos LLMs com o avaliador 3 considerando igualdade e 4 intervalores de tolerância (0,5, 1,0, 1,5 e 2,0)

sentido, cabe aos pesquisadores investir tempo de análise para se refletir sobre as potencialidades e limitações dessas aplicações, sem o deslumbre ou o pânico que elas parecem causar na sociedade.

5. Conclusão e Trabalhos Futuros

Este estudo investigou a viabilidade de se empregar LLMs para avaliar a coerência de resumos escolares, com foco na identificação e análise de verbos *dicendi*. Ao combinar a expertise humana com a capacidade de PLN de quatro LLMs diferentes, foi possível desenvolver um método combinado, que se pretende robusto e preciso para avaliar a qualidade de resumos.

A pesquisa demonstrou que, embora os avaliadores humanos tenham exibido um alto grau de confiabilidade entre si, os LLMs exibiram níveis variados de concordância. Embora os LLMs tenham demonstrado capacidade de identificar verbos relevantes e atribuir pontuações, seu desempenho foi influenciado pela complexidade da tarefa e pelo modelo específico empregado. Notavelmente, o modelo Gemini superou consistentemente os outros em termos de exatidão e precisão.

Além disso, a pesquisa contribuiu para o avanço do campo da avaliação automática de textos, ao demonstrar o potencial dos LLMs para analisar a coerência. Percebeu-se a melhoria significativa no desempenho do LLM quando verbos fornecidos por humanos foram incorporados aos *prompts* de avaliação. Isso sugere que uma abordagem híbrida, combinando experiência humana e aprendizado de máquina, pode aumentar a confiabilidade e a objetividade da avaliação automatizada. Os resultados obtidos podem ser utilizados para desenvolver ferramentas automatizadas de avaliação de resumos, que podem auxiliar professores e alunos no processo de ensino-aprendizagem.

A presente metodologia de trabalho que compara os conhecimentos dos falantes com aqueles dos LLMs poderia ser estendida em trabalhos a nível de ensino básico. Nesse ponto, consideramos que a escola pode ser o local em que a busca pela reflexão sobre o impacto das tecnologias ocorra de modo mais amplo, sem um dualismo rígido (i.e., desejar ou rejeitar as tecnologias digitais). Nesse sentido, considerar o quanto essas tecnologias são capazes (ou não) de (re)produzirem textos é algo que pode ser investigado desde muito cedo no ensino de Língua Portuguesa, sobretudo na produção textual.

Trabalhos futuros relacionados com a coerência do texto e os resultados apresentados neste artigo incluem: (*i*) ampliar a base de dados com resumos de diferentes gêneros e complexidades para melhorar a generalização dos resultados, (*ii*) desenvolver métricas de avaliação mais específicas para resumos coerentes, considerando aspectos linguísticos e semânticos além dos verbos *dicendi*, (*iii*) fornecer feedback imediato aos alunos sobre suas habilidades de escrita e (*iv*) considerar a coerência em textos maiores, devido a limitação dos LLMs em relação à quantidade de tokens processados.

Agradecimentos

Este trabalho foi apoiado pela bolsa Universal do CNPq 2022, pela FAPESC sob o processo 2021TR1510, pela Universidade Estadual de Santa Catarina (UDESC), pelo Projeto Print CAPES-UFSC Automação 4.0 e indiretamente pelo projeto Céos, financiado pelo Ministério Público do Estado de Santa Catarina (MPSC), que tem contribuído significativamente para o aprimoramento do nosso Laboratório (LISA) e da

infraestrutura de processamento de alto desempenho da UFSC. Agradecemos também aos avaliadores humanos da UTFPR que nos ajudaram nas tarefas e aos avaliadores da Linguamática pelos apontamentos no texto.

Contribuições dos autores

Osmar de Oliveira Braz Junior: Conceitualização, Metodologia, Software, Validação, Análise formal, Investigação, Recursos, Curadoria de dados, Redação – rascunho original, Redação – revisão e edição, Visualização. **Ana Julia Araujo Sanchuki**: Validação, Análise formal, Investigação, Recursos, Redação – rascunho original, Redação – revisão e edição. **Roberlei Alves Bertucci**: Conceitualização, Metodologia, Validação, Análise formal, Investigação, Recursos, Redação – rascunho original, Redação – revisão e edição. **Renato Fileto**: Conceitualização, Metodologia, Validação, Análise formal, Investigação, Recursos, Redação – rascunho original, Redação – revisão e edição, Visualização, Supervisão, Administração do projeto.

Referências

Bicudo, Cíntia & Cláudia Valéria Doná Hila. 2015. O bom resumo em situação de vestibular. *Claraboia* 2(2). 102–114.

Biral, Josete. 2003. *Operações recorrentes na produção de resumos*: Universidade Federal do Paraná. Tese de Mestrado.

Bragagnollo, Rubia Mara. 2011. *O gênero resumo acadêmico na formação docente inicial*: Universidade Estadual de Maringá. Tese de Mestrado.

Bragagnollo, Rubia Mara. 2017. A produção textual do gênero resumo escolar. Em Juliano Desiderato Antonio Pedro Navarro (ed.), *Gêneros textuais em contexto de vestibular*, Eduem. 10.7476/9788576287414

Braz Jr, Osmar Oliveira & Renato Fileto. 2021. Investigando coerência em postagens de um fórum de dúvidas em ambiente virtual de aprendizagem com o BERT. Em *32nd Simpósio Brasileiro Informática na Educação*, 749–759. 10.5753/sbie.2021.217397

Campos, Claudia Mendes & Josélia Ribeiro. 2013. Gêneros. Em Iara Bemquerer Costa Maria José Foltran (ed.), *A tessitura da escrita*, 23–44. Editora Contexto

Cançado, Márcia, Luana Amaral, Evelin Amorin, Adriano Veloso & Heliana Mello. 2020. Subjetividade em correções de redações: detecção automática através de léxico de operadores de viés linguístico. *Linguamática* 12(1). 63–79. 10.21814/lm.12.1.313

Cândido, Thiago Gaglietti de & Carine Geltrudes Webber. 2018. Avaliação da coesão textual: Desafios para automatizar a correção de redações. *Revista Novas Tecnologias na Educação* 16(1). 103–112. 10.22456/1679-1916.86013

Clark, Herbert H. & Richard J. Gerrig. 1990. Quotations as demonstrations. *Language* 66(4). 764–805. 10.2307/414729

Costa, Iara Bemquerer & Luciana Pereira da Silva. 2013. Coerência. Em Iara Bemquerer Costa Maria José Foltran (ed.), *A tessitura da escrita*, 64–81. Editora Contexto

De Beaugrande, Robert-Alain & Wolfgang U. Dressler. 1981. *Introduction to text linguistics*. Longman

Inácio, Marcio Lima. 2021. *Sumarização de opinião com base em abstract meaning representation*: Universidade de São Paulo. Tese de Mestrado.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn & Gjergji Kasneci. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences* 103. 102274. 10.1016/j.lindif.2023.102274

Koch, Ingedore Grunfeld Villaça & Luiz Carlos Travaglia. 2021. *A coerência textual*. Editora Contexto 18th edn.

Lima, Juliana Guedes & Maria de Fátima Alves. 2017. O gênero resumo escolar: o que nos dizem os professores do ensino médio? Em *4th Simpósio Nacional de Linguagens e Gêneros Textuais*,

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9). 1–35. 10.1145/3560815

Machado, Anna Rachel, Lília Abreu-Tardelli & Eliane Lousada. 2007. *Resumo*. Parábola Editorial 5th edn.

Machado, Anna Rachel, Eliane Lousada & Lília Santos Abreu-Tardelli. 2005. O resumo escolar: Uma proposta de ensino de gênero. *Signum: Estudos da Linguagem* 8(1). 89–101. 10.5433/2237-4876.2005v8n1p89

Martins, Mário. 2016. A diversidade lexical na escrita de textos escolares. *Fórum Linguístico* 13(1). 1068–1082. 10.5007/1984-8412.2016v13n1p1068

Meira, Ricardo Radaelli, Augusto Weiand, Eliseo Reategui, Marcio Bigolin & Regina Motz. 2023. A analítica da escrita para identificação de indicadores de qualidade textual. *Revista Novas Tecnologias na Educação* 21(2). 342–351. 10.22456/1679-1916.137756

Nunes, Paula Ávila. 2024. Escrever não é útil. *Revista da ABRALIN* 23(2). 192–213. 10.25189/rabralin.v23i2.2190

Paes, Aline & Cláudia Freitas. 2023. ChatGPT, MariTalk e outros agentes de conversação. Em Maria das Graças Volpe Caseli, Helena de Medeiros Nunes (ed.), *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em Português*, BPLN.

Paiola, Pedro Henrique. 2022. *Sumarização abstrativa de textos em português utilizando aprendizado de máquina*: Universidade Estadual Paulista. Tese de Mestrado.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever et al. 2018. Improving language understanding by generative pre-training. Documento de Trabalho (preprint).

Rassi, Amanda Pontes & Priscilla de Abreu Lopes. 2023. Capítulo 19 correção automática de redação. Em Maria das Graças Volpe Caseli, Helena de Medeiros Nunes (ed.), *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em Português*, BPLN.

Rocha, Édila Regina da Silva. 2015. *O gênero resumo de texto no livro didático do nono ano: análise e proposta de intervenção*: Universidade Estadual de Maringá. Tese de Mestrado.

Silva, Andresa Dantas da & Erivaldo Pereira do Nascimento. 2016. A produção textual do gênero resumo acadêmico/escolar: um estudo mediado por sequências didáticas. Em *3rd Congresso Nacional de Educação (CONEDU)*,

Souza, Cinthia Mikaela de. 2022. *Proposta de uma abordagem para sumarização extrativa de textos científicos longos*: Universidade Federal de Minas Gerais. Tese de Mestrado.

Souza, Rita Rodrigues de. 2017. Modelo de estrutura retórica para leitura e escrita de resumo escolar no ensino médio técnico. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 33(3). 911–943. 10.1590/0102-445046525302137346

Van Dijk, Teun Adrianus & Walter Kintsch. 1983. *Strategies of discourse comprehension*. Academic Press

Wang, Yuan & Minghe Guo. 2014. A short analysis of discourse coherence. *Journal of Language Teaching and Research* 5(2). 460–465. 10.4304/jltr.5.2.460-465