Sumarização Automática de Contos Para Uma Plataforma Virtual de Divulgação de Machado de Assis

João Vitor Minarello[1], **Carlos R. Santos Jr.**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia – Campus Hortolândia (IFSP)
CEP 13183-250 — Hortolândia — SP — Brasil

j.minarello@aluno.ifsp.edu.br, carlos.rsantos@ifsp.edu.br

Abstract. *With the continuous advancements in language processing and text generation technologies such as ChatGPT, Bard, Bing Chat, and various other significant initiatives within the expansive field of artificial intelligence, there has been a notable surge in public interest in these areas. However, it is important to recognize that despite the promising advancements and efficiency of these tools, they still have certain limitations. Considering this, the objective of this project is to explore the potential of neural network-based language processing techniques, specifically focusing on the summarization of literary texts, particularly short stories. For this purpose, the algorithm developed for the experiments employed the BART model, which has demonstrated effectiveness in various natural language processing tasks.*

Resumo. *Considerando os avanços em ferramentas de processamento de linguagem e geração de texto como ChatGTP, Bard e Bing Chat, além de numerosas outras grandes iniciativas que envolvem o campo da inteligência artificial em geral, o interesse público em volta de tal área cresce. Apesar disso, deve-se considerar que estas ferramentas, apesar de demonstrarem grandes avanços e eficiência, ainda possuem limitações. Visto isso, este trabalho se propõe a observar as possibilidades do processamento de textos por redes neurais, propondo explorar uma abordagem para a geração de sumários para textos literários do gênero conto. O algoritmo desenvolvido para testar a criação de sumários foi baseado no modelo BART.*

1. Introdução

No ano de 2021, o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação estimou que existiam 59 milhões de domicílios no Brasil com acesso a internet, número equivalente a 82% do total [CETIC 2021]. Além disso, 52% da população brasileira se considera leitora [Pró-Livro 2019], e este é um número que tende a crescer [Rios 2018].

Dadas estas informações, considerou-se que a criação de uma plataforma de divulgação literária *on-line* tinha potencial de se mostrar uma empreitada proveitosa, possibilitando um crescimento do acesso da população (que já frequenta ambientes *on-line*) à cultura, podendo alcançar públicos que tal material não alcançaria por vias comuns, promovendo a diversidade e inclusão literária.

Para a realização deste trabalho, os autores optaram por abordar também um tema que permitisse explorar a integração das tecnologias de inteligência artificial com o campo

da literatura. A proposta do projeto, nesse ponto, foi desenvolver um algoritmo que utilizasse técnicas de sumarização automática, baseadas em inteligência artificial ou, mais especificamente, processamento de linguagem natural (PLN) para criação de sumários de textos do gênero literário conto. Desta forma, criou-se uma oportunidade de pesquisar informações sobre as possibilidades da área no contexto atual, considerando as ferramentas e técnicas selecionadas durante o desenvolvimento do projeto.

A partir da junção das ideias apresentadas, definiu-se que no desenvolvimento do projeto seria criada uma plataforma digital onde fossem divulgadas informações sobre a vida e obra do autor Machado de Assis, que além de ter como mote a fomentação do consumo literário nacional, se mostraria como oportunidade para utilizar de forma prática os sumários gerados automaticamente com técnicas da PLN.

O restante do artigo está dividido em seis seções, sendo as mesmas:

- Trabalhos Correlatos, onde são discutidos três trabalhos relacionados aos temas do projeto;
- Referencial Teórico, onde são apresentados conceitos chave sobre os quais as tecnologias utilizadas no desenvolvimento baseiam-se;
- Desenvolvimento, onde estão detalhados os dois artefatos desenvolvidos durante a evolução do projeto, o algoritmo de sumarização e a plataforma de divulgação do autor Machado de Assis;
- Resultados, onde são discutidos os resultados dos testes dop algoritmo de sumarização criado;
- Conclusão, onde são apresentadas conclusões da pesquisa, assim como possíveis trabalhos futuros;
- Referências, que contém referências a todos os artigos, páginas e livros consultados para o desenvolvimento da pesquisa e escrita do artigo.

2. Trabalhos Correlatos

Esta seção apresenta três trabalhos correlatos ao tema explorado nessa pesquisa. Os mesmos foram encontrados e selecionados a partir de pesquisas no Google Acadêmico, utilizando termos como: "sumarização automática", "processamento de linguagem natural", "sumarização extrativa", "sumarização abstrativa"e "plataforma digital de divulgação literária".

O primeiro [Souza et al. 2017], tem como proposta a criação de uma solução capaz de produzir sumários de textos notadamente diminutos em relação aos textos originais, de forma automática. Prezou-se no trabalho pela produção de sumários finais que não sofressem perdas semânticas consideráveis e que se mantivessem de acordo com as regras gramaticais.

Para alcançar tal proposta, foi desenvolvido um protótipo para captação de textos da *web* de forma automática, e o mesmo foi utilizado para captar 100 textos sobre o tema “informação sobre energia limpa”, que constituiu o corpus comentado utilizado para alimentar um protótipo para a geração automática de sumários, também desenvolvido. O protótipo foi desenvolvido utilizando a linguagem Java SE e em módulos, a saber:

- CAPTURA, que captava automaticamente informações da *web* sobre o tema, em horário predefinido;

- DICIONARIO, que insere palavras presentes no texto em um banco de dados;
- SUMARIO, que gera automaticamente os sumários.

O módulo de sumarização automática apresentado na pesquisa divide-se, de forma simplificada, nas seguintes etapas:

- Identificação de palavras relevantes no texto;
- Eliminação de frases menos relevantes;
- Reescrita das frases restantes, traduzindo-as em frases menores ;
- Aplicação de correção gramatical estatística.

O segundo [Costa and Martins 2015], visou comparar diferentes métodos de sumarização automática, utilizando a bancada *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) como guia na medição de qualidade dos sumários produzidos automaticamente. Tal bancada determina parâmetros para avaliação de resultados para dois domínios diferentes; a geração de títulos para textos jornalísticos e a geração de sumários para textos jornalísticos, e é considerada uma norma em termos de métricas para avaliação de sistemas de sumarização automática.

Foi verificado por Costa que para a geração de sumários de apenas uma frase, ou títulos, que métodos baseados na seleção da primeira frase de cada documento alcançaram os melhores resultados, com base na convenção de que as informações mais relevantes de textos jornalísticos são dispostas nos primeiros parágrafos. No caso da geração de sumários, verificou-se que a abordagem baseada na decomposição de matrizes em valores singulares teve melhores resultados.

No terceiro trabalho [Rios 2018], os autores se propuseram a criar uma plataforma para divulgação literária no Distrito Federal. O trabalho baseou sua justificativa em pesquisas que demonstram que Brasília liderava o ranking de utilização de internet no ano de 2018, e que demonstravam que em 2015 ocorreu um crescimento de 6% no número de leitores no país.

Considerando essas informações, os autores decidiram criar uma plataforma onde leitores e escritores pudessem se comunicar, os últimos podendo inclusive divulgar suas obras ou eventos envolvendo a literatura local. Utilizaram tecnologias disponíveis no o ambiente *on-line* de criação e edição de websites Wix, disponibilizando durante o desenvolvimento uma versão inicial da plataforma.

3. Referencial Teórico

Esta seção apresenta os principais conceitos da área de Inteligência Artificial relacionados as técnicas de Sumarização Automática de Textos explorados no desenvolvimento do projeto.

Como mencionado na seção anterior, a maior parte da pesquisa teórica do trabalho utilizou como plataforma o Google Acadêmico. No caso do referêncial teórico, as pesquisas envolveram termos como: "sumarização automática", "*summarization*", "processamento de linguagem natural", "*natural language processing*", "*extractive summarization*", "*abstractive summarization*", "*neural network*".

3.1. Redes Neurais Artificiais

Uma rede neural artificial é um algoritmo de aprendizado de máquina baseado no neurônio humano [Su-Hyun et al. 2018].

A menor parte de uma rede neural artificial é um neurônio, e os mesmos são organizados em camadas. Os neurônios são regras de processamento matemáticas que calculam os pesos a serem atribuídos aos valores de entrada. Durante o treinamento, o modelo da rede se adapta aos pesos ideais [Suzuki 2011].

Existem dois tipos de treinamentos; supervisionado e não supervisionado. No supervisionado, a rede treinada tem acesso aos dados esperados como saída para os dados de entrada, e se adapta a partir da comparação entre o resultado gerado e o esperado. No não supervisionado a rede não tem dados esperados para sua saída, e sua adaptação se baseia em um critério interno criado para a rede. [Anderson and McNeill 1992].

3.2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área de estudos que investiga e propõe ferramentas para tornar a linguagem humana processável, ou inteligível, para computadores.

Na definição proposta por Elizabeth D. Liddy [Liddy 2001], a PLN é uma série de técnicas computacionais teoricamente motivadas para analisar e representar textos de ocorrência natural em um ou mais níveis de análise linguística, com a finalidade de alcançar uma proficiência no processamento de linguagem comparável ao humano. A mesma declara em sua pesquisa que existem muitos métodos e técnicas para explorar um tipo específico de análise linguística, e justifica assim a generalização na passagem "uma série de técnicas computacionais" de sua definição.

Neste artigo, foi proposta a exploração do campo específico da sumarização automática de textos, que se relaciona ao PLN no tocante do processamento inicial do texto, momento no qual cria-se uma representação computacional do mesmo que é mais apropriada para aplicações computacionais.

3.3. Sumarização Automática de textos

A sumarização automática de textos consiste na automação de processos para gerar sumários a partir de textos originais. Uma ferramenta de sumarização automática tem o objetivo de produzir textos condensados que apresentem as ideias principais de um texto original de forma inteligível [Maybury 1995]. Alguns desafios são comumente enfrentados na empreitada da criação de uma ferramenta para geração automática de sumários, como por exemplo, segundo El-Kassas [Wafaa et al. 2020]:

- Identificação das passagens mais pertinentes do texto;
- Sumarização de longas obras, como livros;
- Sumarização multidocumento;
- Avaliação de sumários gerados por computadores sem a necessidade de um resumo criado por humanos para comparação;
- Geração de um sumário abstrativo equivalente a um sumário criado por humanos;

Algumas abordagens foram definidas para atingir esse objetivo, as quais dividem comumente a tarefa em três etapas: pré-processamento, onde o texto é traduzido para um formato estruturado e mais apto a ser processado por um computador, processamento, onde uma das abordagens é aplicada ao texto estruturado para obter o sumário, e pós-processamento, onde o sumário gerado pode ser reorganizado e melhorado, com a finalidade de torná-lo mais compreensível.

3.3.1. Abordagens da Sumarização Automática de Textos

Dentre as possíveis abordagens para o problema, três têm sido mais frequentemente pesquisadas e utilizadas; a extrativa, a abstrativa e um híbrido das duas [Wafaa et al. 2020]. Esta seção apresenta de forma resumida cada uma dessas abordagens.

3.3.1.1. Abordagem Abstrativa

Na abordagem abstrativa, o conteúdo do sumário final provém de fontes externas, não contidas no texto original. É tecnicamente mais desafiadora, pois depende de complexos processamentos de linguagem natural [Gambhir and Gupta 2017].

Para criar sumários abstrativos, o texto original deve ser parafraseado a partir de uma compreensão dos principais conteúdos apresentados no texto original, o que é feito a partir da utilização de métodos da PLN [Al-Abdallah and Al-Taani 2017]. Em suas implementações, a técnica utiliza o pré-processamento e utiliza a representação computacional do texto criada para gerar um sumário a partir de técnicas de PLN [Chitrakala et al. 2018].

A abordagem abstrativa tem a seu lado a possibilidade de gerar sumários mais apropriados, a partir do parafraseamento, fusão ou compressão, o que pode aproximar o sumário gerado de um feito por humanos [Liwei et al. 2018]. Além disso, a produção de sentenças que não reutilizem partes ou frases do texto original apresenta a possibilidade de maior condensação do texto, pois possibilita evitar redundâncias [Y. et al. 2018].

Porém, a dificuldade para criar uma ferramenta de geração de sumários automática ainda é alta. O próprio campo do processamento de linguagem natural é emergente [Shao et al.,2017], o que pode atrapalhar o desenvolvimento de pesquisas que dependam do mesmo.

Por fim, a abordagem abstrativa tende a se limitar ao domínio do modelo no qual se baseia. Apesar de poder ter resultados positivos no domínio em que foram inicialmente treinados, as modelos tendem a ser falhos em um contexto mais generalizado [Gupta and Lehal 2010]. Normalmente os domínios sobre os quais os modelos são treinados tem linguagem e estilos de textos limitados e bem definidos, e isso os torna inflexíveis para demais estilos ou gêneros literários.

3.3.1.2. Abordagem Extrativa

Na abordagem extrativa, o sumário gerado contém apenas informações que já estavam no documento original, ou seja, apenas frases ou palavras contidas no texto a ser resumido. Por ser mais simples de implementar e aprimorar, espera-se que a abordagem extrativa apresente melhores resultados em boa parte dos casos.

Para implementar uma ferramenta de sumarização extrativa, costuma-se seguir um padrão que divide a tarefa em três etapas :

- Criação da representação computacional;
- Classificação das frases por importância inferida [Nenkova and McKeown 2012];
- Com base na classificação anterior, extração de sentenças com pontuações mais altas;

- Ordenação das sentenças, gerando sumário final.

A abordagem extrativa determina que a prioridade das sentenças é definida a partir da recorrência dos termos presentes na mesma no resto do texto, e assim é tomada a decisão sobre a sua permanência ou não no sumário final. Alguns dos problemas recorrentes e passíveis de exploração dentro do campo da sumarização extrativa são:

- Redundância ou repetição de frases;
- Incoerência, causada pela extração a partir de documentos diferentes, ordenação errônea de frases ou repetição de palavras;
- Má priorização dos termos encontrados no texto, que pode acarretar na exclusão de informações importantes para a compreensão de textos com assuntos múltiplos ou que se contradigam.

Pesquisas têm sido desenvolvidas aplicando a abordagem extrativa em conjunto com outras técnicas, como redes neurais, que vem sendo pesquisada com maior profundidade recentemente [Widyassari et al. 2020].

3.3.1.3. Abordagem Híbrida

A abordagem híbrida é uma combinação das abordagens abstrativa e extrativa, e frequentemente implementa os seguintes passos [Wafaa et al. 2020]:

- Pré-processamento;
- Extração das sentenças mais importantes do texto (fase extrativa);
- Geração de sumário utilizando as sentenças extraídas anteriormente (fase abstrativa);
- Pós processamento, onde as frases geradas são ordenadas.

A ideia por trás da abordagem é que os sumários finais tenham maior qualidade, por fazerem uso de duas técnicas amplamente pesquisadas e desenvolvidas. Porém, como o texto utilizado para fazer a etapa abstrativa é composto por frases extraídas que não são necessariamente consecutivas, o sumário resultante pode ser desconexo e de difícil compreensão.

4. Desenvolvimento

O desenvolvimento do trabalho dividiu-se em duas etapas:

- Criação de um algoritmo para sumarização automática de textos do gênero conto, de autoria de Machado de Assis;
- Desenvolvimento de uma plataforma de divulgação para disponibilição de informações sobre o autor, sumários gerados e contos originais.

A seguir, as duas são detalhadas com maior profundidade.

4.1. Algoritmo de Sumarização

Desenvolveu-se um algoritmo para a sumarização de contos, utilizando o modelo pré-treinado BART, da biblioteca Transformer. Tal modelo, de propriedade do Facebook, tem como vantagem a disponibilização de tecnologias no estado da arte, chegando a alcançar classificações 6.0 pontos mais altas na tarefa de sumarização do que o trabalho

anterior com maior classificação, BERT (de acordo com a avaliação na bancada ROUGE) [Lewis et al. 2020].

A definição da utilização dessa tecnologia baseou-se no fato da mesma gerar sumários o partir da técnica abstrativa, o que poderia ser interessante para textos do gênero conto, partindo da pressuposição de que reutilizar a linguagem original dos mesmos pode afetar a coerência dos sumários gerados. Como gênero literário de livre criação, contos tendem a possuir linguagem e estrutura mais variáveis, diminuindo as possibilidades de basear-se apenas nas palavras e construções linguísticas contidas no mesmo para gerar um sumário conciso.

Apesar da escolha ser justificada, deve-se manter em mente que a mesma foi feita em caráter exploratório. A técnica abstrativa também pode tender a demonstrar limitações com textos de um domínio onde o modelo utilizado não tenha sido anteriormente treinado, como é o caso do BART para textos do gênero conto de autoria de Machado de Assis. O modelo foi pré-treinado a partir de artigos da Wikipédia e do dataset BookCorpus, ambos em inglês [Lewis et al. 2020].

Modelos da biblioteca Transformers, como o BART, requerem grande poder de processamento, o que inviabiliza a utilização dos mesmos fora de servidores ou ambientes especializados. Dada essa limitação, os pesquisadores decidiram utilizar o Google Colaboratory para desenvolver o algoritmo de processamento de textos. Tal ferramenta é disponibilizada gratuitamente para fins de estudos que envolvam processamento de linguagem natural. Nos "*notebooks*" (ambientes que podem ser criados na suíte de computação em nuvem Google Cloud Platform), é possível executar códigos desenvolvidos em Python [Google 2023].

O algoritmo (Figura1) dividiu-se em três principais partes, a serem detalhadas na sequência:

- Tradução do conto original para inglês, necessária para o processamento do texto na mesma linguagem com a qual o modelo utilizado foi treinado;
- Sumarização do texto traduzido;
- Tradução do sumário gerado em inglês para português brasileiro, a fim de que o mesmo possa ser disponibilizado na plataforma.

A tradução do conto original (Figura 2), assim como a posterior tradução do sumário final, foi feita por meio da utilização da biblioteca Translators, disponível para uso gratuito em aplicações desenvolvidas na linguagem Python. A biblioteca utiliza vários serviços de tradução amplamente difundidos virtualmente, como Google, Bing e Deepl [Uliontse 2023]. Um dos serviçcos de tradução deve ser indicado via parâmetro à biblioteca durante a implementação.

Para o desenvolvimento do algoritmo em questão, utilizou-se o serviço Google Translate, mais especificamente nessa etapa, para tradução do português para o inglês. Como a biblioteca limita a quantia de caracteres do texto a ser processado a 5000, os contos originais foram divididos em parágrafos via código.

Uma vez que os parágrafos são traduzidos, os mesmos são processados utilizando o pipeline de sumarização do modelo BART (Figura 3), também disponibilizado de forma gratuita.

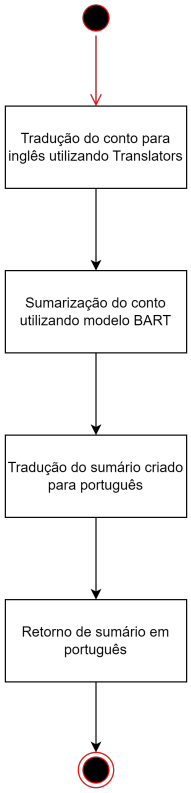

Figura 1. Diagrama do algoritmo criado

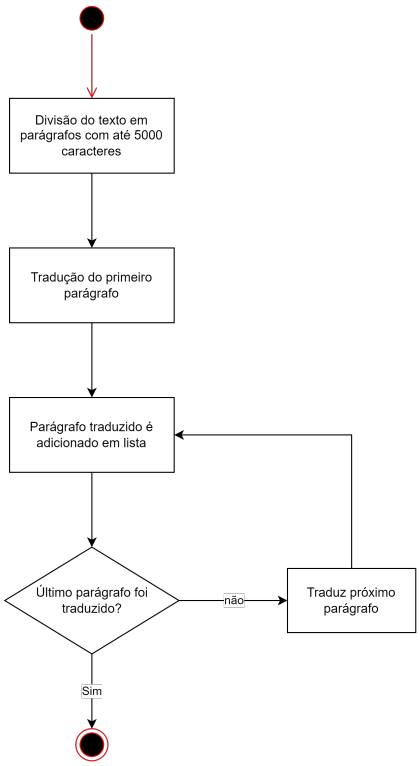

Figura 2. Diagrama do processo de tradução de textos

O pipeline possui uma limitação de quantia de *tokens* (unidades individuais de texto que são definidas a partir do esquema de divisão do texto utilizado pelo modelo) do texto a ser sumarizado, assim como do sumário gerado. A limitação padrão do modelo BART é de 1024 *tokens*, número que não comportou o tamanho dos contos testados. Uma das alternativas para a sumarização de textos longos é sua divisão em textos menores. Como o mesmo processo já havia sido implementado para a tradução dos contos do português para o inglês, o protótipo reutilizou essa funcionalidade na etapa do processamento.

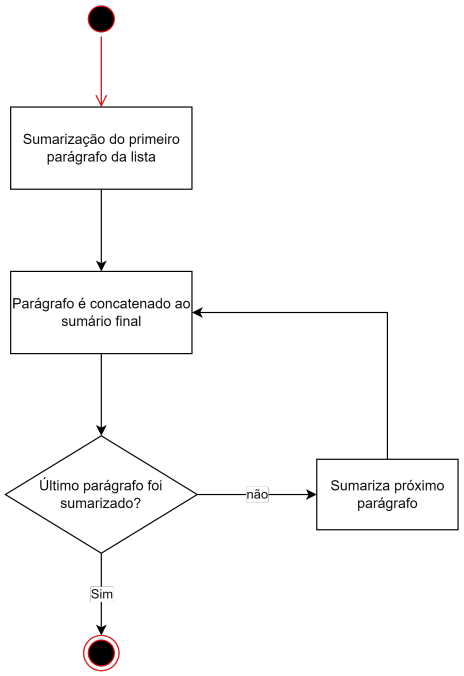

Figura 3. Diagrama do processo de sumarização de textos

Por fim, o sumário gerado a partir do modelo BART precisou ser traduzido de volta para o idioma original do conto, visto que o mesmo seria disponibilizado para um público brasileiro através da plataforma. Nessa etapa, utilizou-se novamente a biblioteca Translators para gerar um sumário traduzido, ainda utilizando o serviço Google Translate, dessa vez do inglês para o português.

4.2. Plataforma de Divulgação Virtual do Autor Machado de Assis

A plataforma foi dividida em quatro seções, cada uma contida em uma página. O conteúdo das mesmas está disposto da seguinte maneira:

- Sumários (Figura 4): contém os sumários gerados a partir do processamento dos contos, efetuado utilizando o algoritmo descrito na seção anterior;

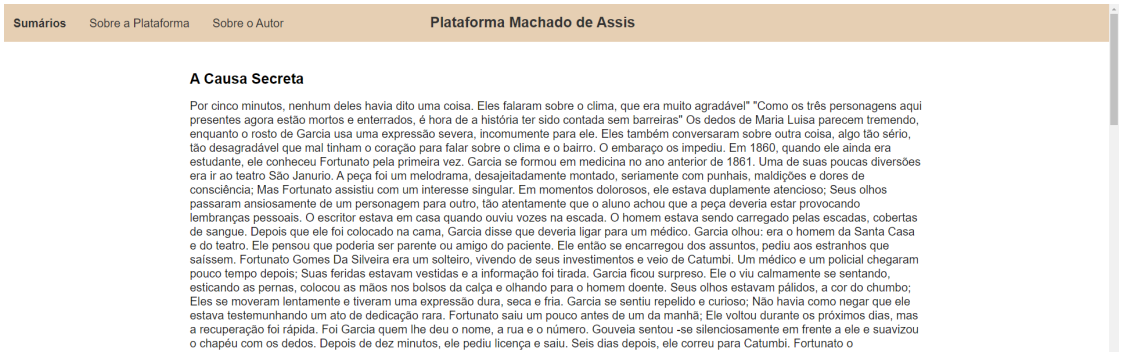

Figura 4. Página Sumários

- Páginas de Contos (Figura 5), onde os contos processados foram disponibilizados em seu formato original, podendo ser acessados através da lista de sumários contida na Página Sumários.

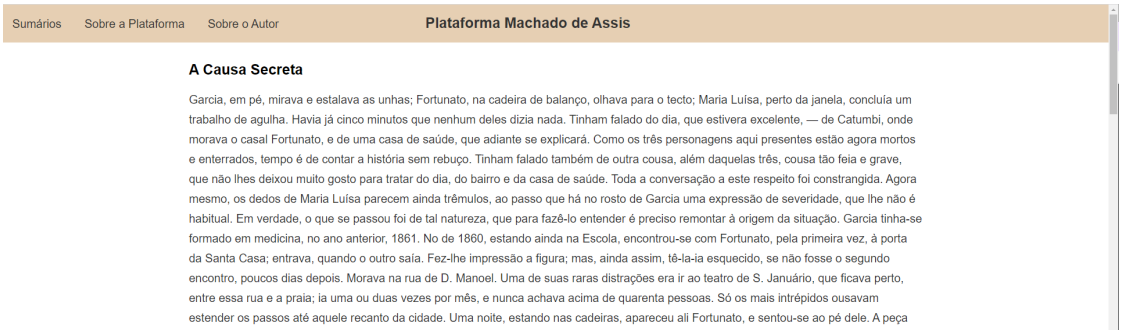

Figura 5. Página do conto A Causa Secreta

- Sobre a Plataforma (Figura 6), onde as ideias por trás da criação do site foram disponibilizadas, acessível pela barra de navegação principal;

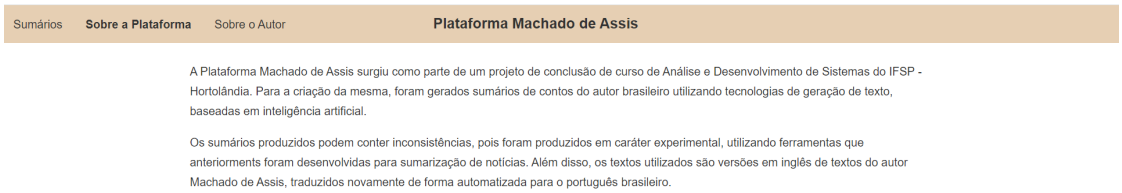

Figura 6. Página Sobre a Plataforma

- Sobre o Autor, onde disponibilizou-se uma breve biografia de Machado de Assis (Figura 7), assim como a sua bibliografia (Figura 8), acessível pela barra de navegação principal;

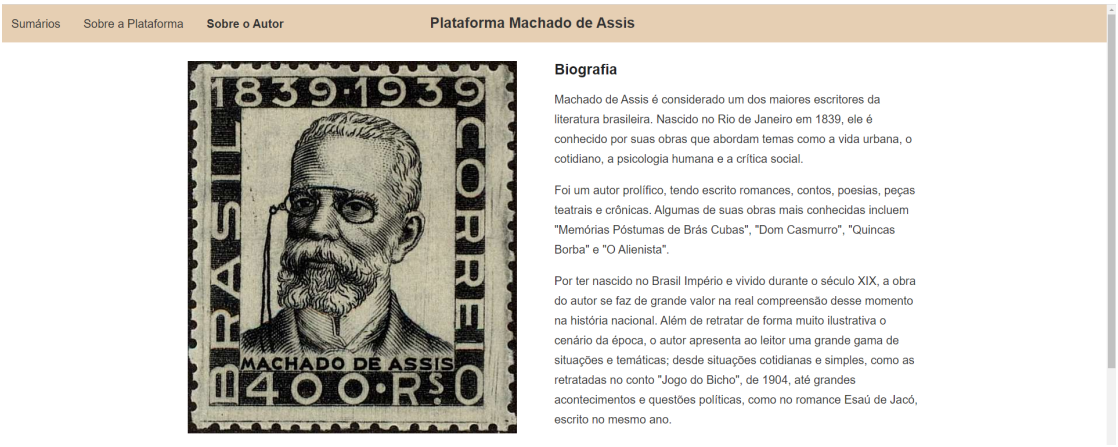

Figura 7. Página Sobre o Autor

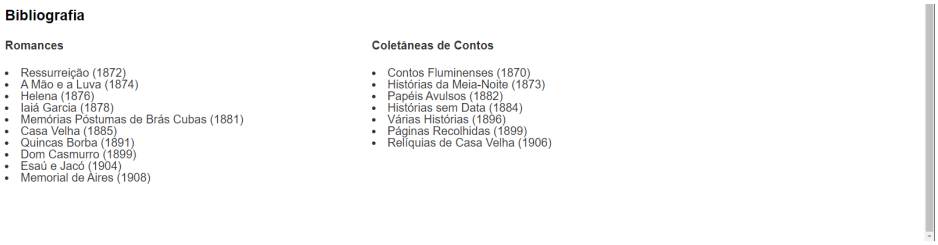

Figura 8. Bibliografia na Página Sobre o Autor

Acessando o portal, o usuário é direcionado para a página contendo os sumários do autor, onde poderá selecionar um conto da lista para acessar a obra completa ou clicar em um dos títulos de páginas listados na barra de navegação superior: Sobre o Autor e Sobre a Plataforma, sendo direcionado assim para as respectivas páginas. Abaixo está um mapa do site, ilustrando as páginas conforme podem ser acessadas.

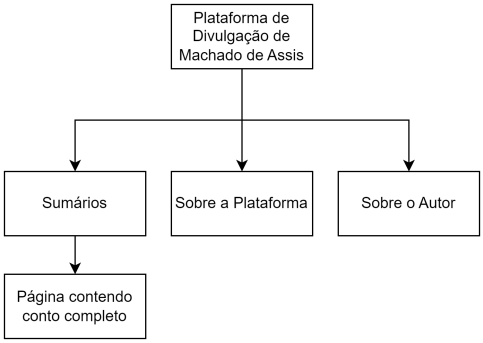

Figura 9. Mapa do Site da Plataforma

5. Resultados

Em um primeiro momento, pode-se considerar que os sumários gerados pelo algoritmo criado satisfazem suas finalidades iniciais, considerando que seu objetivo primário seja

apresentar o conteúdo do texto de forma reduzida. Por outro lado, ao efetuar uma avaliação mais aprofundada dos textos criados, é possivel discernir em alguns casos passagens mal redigidas ou pouco organizadas, decorrentes de limitações das tecnologias utilizadas. A seguir, dois exemplos ilustram os tipos de construções problemáticas mais frequentes nos resumos gerados a partir dos testes efetuados.

Na frase inicial do sumário gerado para o conto "A Causa Secreta", podemos observar a seguinte construção:

"Por cinco minutos, nenhum deles havia dito uma coisa."

No texto original, a parte mais provável de estar sendo sumarizada aqui é a seguinte:

"Havia já cinco minutos que nenhum deles dizia nada."

Podemos enxergar uma provável causa para essa ressignificação confusa da frase ao observarmos a tradução da frase original para o inglês:

"It had been five minutes since any of them had said anything."

Quando traduzida de volta para o idioma original (português), se tal tradução não considerar um contexto e possível adaptação da mesma, a sentença acaba se transformando na que foi inserida no sumário final.

Outra passagem onde podemos observar uma das limitações do algoritmo, no sumário gerado para o conto "O Enfermeiro", é a seguinte:

"Eu disse meu nome. Dificilmente eu o havia proferido quando ele fez um gesto de espanto. Então ele me perguntou meu nome.
Seu nome é Colombo?
Sim, Colombo.
Qual é o seu nome?
Colombo", diz ele.
Eu não sei qual é o seu nome. Colombo.
"O que é?"ele pergunta.
"É Colombo", responde o homem."

Essa passagem demonstra limitações do modelo ao trabalhar com textos estilizados ou com ideias conflituosas. Na passagem original, um dos personagens narra a interação de forma indireta, onde um segundo personagem estranha seu sobrenome, "Valongo". Depois, pergunta, confirmando se ouviu corretamente: "Colombo?":

"Em seguida, perguntou-me pelo nome: disse-lho e ele fez um gesto de espanto. Colombo? Não, senhor: Procópio José Gomes Valongo. Valongo? achou que não era nome de gente, e propôs chamar-me tão-somente Procópio, ao que respondi que estaria pelo que fosse de seu agrado."

No sumário, a classificação do modelo provavelmente teve problemas entre diferenciar o nome real do incorreto, além de gerar uma narrativa confusa da ordem das falas, adaptando-as parcialmente para um discurso direto.

Outro ponto que é possível observar nessa última passagem gerada é o "O que é?", que reforça a limitação do sistema de tradução na forma em que foi utilizado. Neste caso, o falante de português utilizaria, mais recorrentemente, algo como "Qual é?"em referência ao nome sobre o qual está inquirindo.

6. Conclusão

Ao fim do desenvolvimento do projeto, haviam sido criadas versões iniciais do algoritmo de sumarização de contos e da Plataforma Virtual de Divulgação de Machado de Assis, ambos cumprindo, até certo ponto, as propostas iniciais definidas para o projeto.

Existem melhorias que podem ser trabalhadas nas duas partes do projeto, tanto por questões de expansão do conteúdo mesmo quando por aperfeiçoamentos no algoritmo para geração de sumários (como discutido anteriormente, na seção Resultados) e na estrutura do *back-end*.

No portal, é possível continuar o trabalho planejando e desenvolvendo um sistema mais robusto e de manutenção mais usual. Uma proposta possível é trabalhar com a criação de um banco de dados, onde as informações disponibilizadas nas páginas podem ser mantidas.

Outra possibilidade é a utilização da API do Google Colaboratory, possibilitando a geração de sumários a partir de textos que possam ser submetidos por um mantenedor da plataforma ou pelo público consumidor da mesma.

Mais propostas que podem ser trabalhadas envolvem melhorias nos sumários gerados pela ferramenta, sendo possível fazer melhorias no treinamento do modelo BART, o que pode gerar sumários mais próximos do que um humano seria capaz.

Como os contos precisaram ser divididos e traduzidos para os processamentos, os sumários gerados continham alguns problemas de coerência, ou apresentavam construções não utilizadas no português brasileiro. Visto isso, também apresenta-se a oportunidade de alterar operações do algoritmo, pesquisando ferramentas ou fluxos mais apropriadas para as etapas de tradução.

Ademais, diálogos inclusos nos contos também afetaram os sumários gerados, pois o estilo do texto tende a mudar nesses trechos. É possível explorar maneiras de contornar essas limtações, fazendo um pré-processamento do conto onde trechos irregulares são removidos. Porém, em tal caso, deve-se lembrar que a remoção de diálogos ou outros trechos deve ser feita de forma atenta, para que a coerência do sumário final não seja ainda mais prejudicada.

7. Referências

Referências

Al-Abdallah, R. and Al-Taani, A. (2017). Arabic single-document text summarization using particle swarm optimization algorithm. *Procedia Computer Science*, 117:30–37.

Anderson, D. and McNeill, G. (1992). Artificial neural networks technology. *Kaman Sciences Corporation*, 258(6):1–83.

CETIC (2021). Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: Tic domicílios 2020. Disponível em: `https://cetic.br/media/docs/publicacoes/2/TIC_Domicilios_2020_LivroEletronico.pdf`.

Chitrakala, G., N., M., B., R., C., R., and B., D. (2018). Concept-based extractive text summarization using graph modelling and weighted iterative ranking. pages 149–160.

Costa, M. A. A. and Martins, B. (2015). Uma comparação sistemática de diferentes abordagens para a sumarização automática extrativa de textos em português. *Linguamática*, 7(1).

Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47.

Google (2023). Google colaboratory. https://colab.research.google.com/notebooks/welcome.ipynb?hl=pt-BR. Acesso em 29 jun. 2023.

Gupta, V. and Lehal, G. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liddy, E. D. (2001). Natural language processing. https://www.bibsonomy.org/bibtex/24864613c44e0472d987b3933e14ffc54/zromero. Acesso em 29 jun. 2023.

Liwei, H., Po, H., and Chao, B. (2018). *Abstractive Document Summarization via Neural Model with Joint Attention*, pages 329–338.

Maybury, M. T. (1995). Generating summaries from event data. *Information Processing & Management*, 31(5):735–751. Summarizing Text.

Nenkova, A. and McKeown, K. (2012). *A Survey of Text Summarization Techniques*, pages 43–76.

Pró-Livro, I. (2019). Retratos da leitura no brasil - 5ª edição. Disponível em: `http://prolivro.org.br/home/images/attachments/38.pdf`.

Rios, M. L. (2018). Quadrado literário: uma plataforma de divulgação da literatura brasiliense.

Souza, O., Tabosa, H. R., Oliveira, D. M., and Oliveira, M. H. S. (2017). Um método de sumarização automática de textos através de dados estatísticos e processamento de linguagem natural. *Informação & Sociedade: Estudos*, 23(3).

Su-Hyun, H., Woon, K. K., SangYun, K., and Chul, Y. Y. (2018). Artificial neural network: Understanding the basic concepts without mathematics. *dnd*, 17(3):83–89.

Suzuki, K. (2011). *Artificial Neural Networks - Methodological Advances and Biomedical Applications*.

Uliontse (2023). Translators. https://pypi.org/project/translators. Acesso em 29 jun. 2023.

Wafaa, E.-K., Cherif, S., Ahmed, R., and Hoda, M. (2020). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Widyassari, A., Rustad, S., Shidik, G., Noersasongko, E., Syukur, A., Affandy, Setiadi, and Moses, D. R. I. (2020). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 34.

Y., S. D., Raj, K., and Sudiksha, J. (2018). Development of embedded platform for sanskrit grammar-based document summarization. In *Speech and Language Processing for Human-Machine Communications*, pages 41–50, Singapore. Springer Singapore.

Documento Digitalizado Público

Anexo I (Artigo) – João Vitor Minarello - HT1520199

Assunto: Anexo I (Artigo) – João Vitor Minarello - HT1520199
Assinado por: Carlos Junior
Tipo do Documento: Relatório
Situação: Finalizado
Nível de Acesso: Público
Tipo do Conferência: Documento Original

Documento assinado eletronicamente por:

- **Carlos Roberto dos Santos Junior**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 07/07/2023 17:06:19.

Este documento foi armazenado no SUAP em 07/07/2023. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1372188
Código de Autenticação: 97f4df3742