



TRABALHO DE CONCLUSÃO DE CURSO

**Sumarizador Abstrato de Texto
utilizando Redes Neurais Recorrentes
Corpus: Acórdãos do TCU**

Daniel Barbosa Beigelman

Brasília, dezembro de 2019

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

TRABALHO DE CONCLUSÃO DE CURSO

**Sumarizador Abstrato de Texto
utilizando Redes Neurais Recorrentes
Corpus: Acórdãos do TCU**

Daniel Barbosa Beigelman

*Trabalho de Conclusão de Curso submetido ao Departamento de Engenharia
Elétrica como requisito parcial para obtenção
do grau de Engenheiro Eletricista*

Banca Examinadora

Prof. Alexandre Ricardo Soares Romariz, Ph.D, _____

FT/UnB
Orientador

Prof. Thiago Alves Fonseca, FGA/UnB _____

Examinador interno

Vinícius Oliveira Silva _____

Examinador externo

FICHA CATALOGRÁFICA

BEIGELMAN BARBOSA, DANIEL

Sumarizador Abstrato de Texto utilizando Redes Neurais Recorrentes Corpus: Acórdãos do TCU [Distrito Federal] 2019.

xvi, 23 p., 210 x 297 mm (ENE/FT/UnB, Engenheiro, Engenharia Elétrica, 2019).

Trabalho de Conclusão de Curso - Universidade de Brasília, Faculdade de Tecnologia.

Departamento de Engenharia Elétrica

- | | |
|------------------------------|---------------------------------------|
| 1. Sumarização Abstrata | 2. Processamento de Linguagem Natural |
| 3. Redes Neurais Recorrentes | 4. Sequência-para-sequência |
| I. ENE/FT/UnB | II. Título (série) |

REFERÊNCIA BIBLIOGRÁFICA

BEIGELMAN, D. (2019). *Sumarizador Abstrato de Texto utilizando Redes Neurais Recorrentes Corpus: Acórdãos do TCU*. Trabalho de Conclusão de Curso, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 23 p.

CESSÃO DE DIREITOS

AUTOR: Daniel Barbosa Beigelman

TÍTULO: Sumarizador Abstrato de Texto utilizando Redes Neurais Recorrentes Corpus: Acórdãos do TCU.

GRAU: Engenheiro Eletricista ANO: 2019

É concedida à Universidade de Brasília permissão para reproduzir cópias deste Trabalho de Conclusão de Curso e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. Os autores reservam outros direitos de publicação e nenhuma parte desse Trabalho de Conclusão de Curso pode ser reproduzida sem autorização por escrito dos autores.

Daniel Barbosa Beigelman

Depto. de Engenharia Elétrica (ENE) - FT

Universidade de Brasília (UnB)

Campus Darcy Ribeiro

CEP 70919-970 - Brasília - DF - Brasil

AGRADECIMENTOS

À toda a minha família, que sempre me deu amor, confiança e suporte durante todo o processo da minha formação, em especial, aos meus pais, Ana e Paulo, que tornaram tudo isso possível quando investiram em minha educação desde pequeno.

À a minha namorada, Luíza, pelo seu apoio, amor, companheirismo e compreensão. E por me engajar e ser uma pessoa melhor, procurando a melhor versão de mim mesmo sempre. É um privilégio crescer ao seu lado.

Ao professor Alexandre Romariz, pelas oportunidades concedidas, suporte e confiança em me orientar e me guiar nesse projeto de fundamental importância na minha formação profissional.

Aos meus amigos Guilherme Raposo, Gustavo Leão, Marcos Diniz, Thiago Magalhães, Lucas Bamidele, Vitor Bordini e Pedro Sanches, que trilharam junto comigo essa jornada de graduação, por toda a rede de suporte que criamos, ao longo desses anos, e todas as boas lembranças que levaremos desse período.

À Universidade de Brasília, pelo suporte infraestrutura e oportunidade concedidas para a minha formação.

À ENETEC, por ser umas das maiores escolas pela qual já passei, por permitir meu desenvolvimento pessoal, abrir a minha perspectiva para outros mundos e pelas grandes amizades que lá fiz.

À todos que contribuíram de alguma forma com a minha formação, deixo aqui meu agradecimento.

RESUMO

Este trabalho apresenta a implementação de um modelo sequência-para-sequência de sumarização abstrata de texto usando Redes Neurais Recorrentes (RNN) configuradas em codificador e decodificador com um sistema de atenção-global implementado entre as duas redes neurais. O modelo foi treinado em cima do corpus de acórdãos do Tribunal de Contas da União (TCU), que conta com dados referentes aos acórdãos proferidos pelo TCU entre os anos de 1992 até 30/08/2019.

ABSTRACT

In this work, is presented an implementation of a framework for abstractive text summarization based on a sequence-to-sequence oriented Attentional Encoder-Decoder Recurrent Neural Networks model. The model was trained on the corpus of judgments of the Federal Court of Audit (TCU), which contains data related to the judgments handed down by TCU between 1992 and 30/08/2019.

SUMÁRIO

1	INTRODUÇÃO	1
1.1	CONTEXTUALIZAÇÃO	1
1.2	JUSTIFICATIVA	3
1.3	OBJETIVOS	3
2	REVISÃO BIBLIOGRÁFICA	4
3	METODOLOGIA	6
3.1	BANCO DE DADOS	6
3.2	LIMPEZA DOS DADOS	6
3.3	<i>Padding</i>	7
3.4	<i>Tokens</i> ESPECIAIS	8
3.5	<i>Tokenização</i>	8
3.6	ARQUITETURA	9
3.6.1	CODIFICADOR	9
3.6.2	DECODIFICADOR	10
3.6.3	CAMADA DE ATENÇÃO	11
3.7	ALGORITMO DE SELEÇÃO DA PRÓXIMA PALAVRA	11
3.8	TREINAMENTO DOS MODELOS	12
3.9	MÉTRICAS DE AVALIAÇÃO	13
3.9.1	BLEU	13
3.9.2	ROUGE	14
4	RESULTADOS E DISCUSSÃO	15
4.1	RESULTADOS OBJETIVOS	15
4.2	RESULTADOS SUBJETIVOS	16
4.3	DISCUSSÃO	20
5	CONCLUSÃO	21
	REFERÊNCIAS BIBLIOGRÁFICAS	22
	APÊNDICES	23
I.1	LINK PARA O GITHUB COM OS CÓDIGOS DO PROJETO	23

LISTA DE FIGURAS

1.1	Ilustração de um neurônio.....	2
1.2	Ilustração de redes neurais	2
1.3	Ilustração de uma rede LSTM.....	2
3.1	Ilustração do processo de tokenização	9
3.2	Figura ilustrativa do codificador atuando	10
3.3	Figura ilustrativa do decodificador atuando	10
3.4	Camada de Atenção.....	11

LISTA DE TABELAS

3.1	Resultados de treinamento do modelo 1	12
3.2	Resultados de treinamento do modelo 2	12
3.3	Resultados de treinamento do modelo 3	13
4.1	Comparativo de métrica de avaliação BLEU-1.....	15
4.2	Comparativo de métrica de avaliação ROUGE-1	15
4.3	Seleção de exemplos produzidos pelo Modelo 1	17
4.4	Seleção de exemplos produzidos pelo Modelo 2	18
4.5	Seleção de exemplos produzidos pelo Modelo 3	20

LISTA DE SÍMBOLOS

Siglas

ATL	Attention Layer
GPU	Graphics Processing Unit
GRU	Gate Recorrent Unit
LR	Learning Rate
LSTM	Long Short Term Memory
MT	Machine Translation
NLP	Processamento de Linguagem Natural
RN	Rede Neural
RNN	Rede Neural Recorrente
Seq2Seq	sequência-para-sequência
TCU	Tribunal de Contas da União

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

O crescimento exponencial de informações que é gerada diariamente, devido à popularidade e à democratização do ambiente web, trouxe consigo a necessidade de resumir as informações que chegam até nós, de forma a poupar o nosso tempo e esforço de encontrar informações concisas e relevantes que estamos procurando. Nesse contexto, a tarefa de diminuir o tamanho de um texto enquanto se preserva o seu conteúdo original tornou-se algo muito valioso para as pessoas. Como existem várias formas de escrever e reduzir um texto, esse tópico se tornou algo muito interessante entre os pesquisadores.

Sumarização automática de texto é o processo de gerar automaticamente resumos que retenham o maior valor semântico e os principais conteúdos de um documento original [1]. Esta tem sido umas das áreas mais pesquisadas entre a comunidade de Processamento Natural de Linguagem (NLP). As técnicas de sumarização, que tem como base a extração de frases exatas como aparecem no texto fonte ou que tem como base a geração de novas frases usando técnicas de NLP, são categorizadas, respectivamente, em técnicas extrativas e abstratas. A sumarização extrativa, que identifica a frase ou as frases importantes do texto fonte e as extrai dele, foi um tópico amplamente pesquisado e chegou ao seu estágio de maturidade nos últimos anos. Agora, a pesquisa mudou para a sumarização abstrata, que é a tarefa de gerar resumos através de paráfrases, podendo gerar sentenças que não se encontram no texto fonte. As complexidades subjacentes ao texto em linguagem natural tornam a sumarização abstrata uma tarefa bastante difícil e desafiadora.

Para a execução desta tarefa neste trabalho, redes neurais foram utilizadas como peça principal na construção do sumarizador abstrato. Pois isso, é importante explicarmos o que são e como funcionam essas redes. As redes neurais são modelos computacionais inspirados pelo sistema nervoso, em especial os neurônios, que são capazes de realizar o aprendizado de uma tarefa bem como o reconhecimento de padrões. O componente básico de uma rede neural se chama neurônio ilustrado na Figura 1.1. As entradas e saídas desse componente são números e cada conexão de entrada está associada com um peso que foi atribuído a ela. O neurônio calcula a soma ponderada das entradas pelos seus pesos e então aplica uma função de ativação, que vai atribuir características não lineares ao componente. Essa característica não linear permitirá que vários neurônios sejam agregados formando redes neurais profundas com múltiplas camadas, ilustrada na Figura 1.2, de maneira a possibilitar a modelagem de problemas e cálculos mais complexos.

No caso da nossa tarefa, será utilizada um tipo especial de rede neural chamada RNN (Rede Neural Recorrente). Essa rede tem como característica diferencial a capacidade de armazenar informações ao longo do tempo, o que permite que o nosso modelo consiga processar melhor as informações do texto para gerar um sumário mais coerente. Contudo, com a necessidade de armazenar informações do passado, nos deparamos com a dificuldade de selecionar quais informações

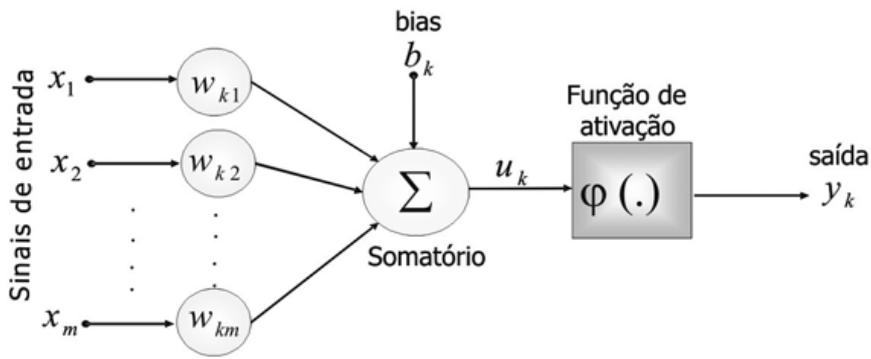


Figura 1.1: Ilustração de um neurônio

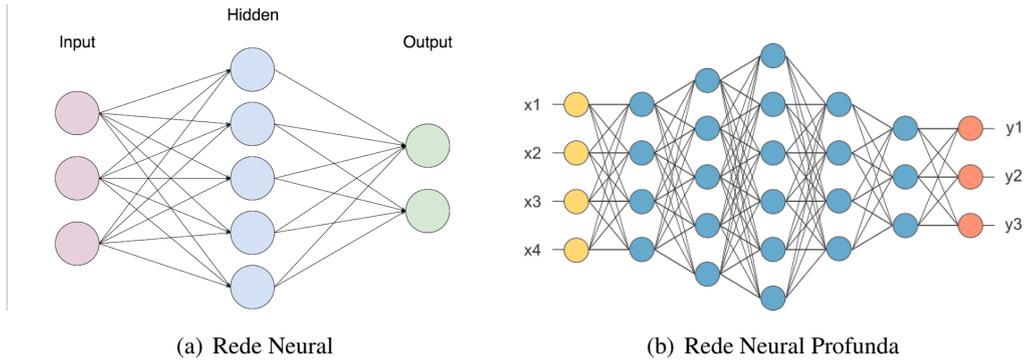


Figura 1.2: Ilustração de redes neurais

devemos guardar ou não. Para solucionar esse problema foi utilizada uma categoria específica de RNN chamada LSTM (*Long Short Term Memoria*), que por de alguns portões que ela possui, consegue aprender quais informações armazenar, quais esquecer e quais atualizar. Essa capacidade seletiva de informações se da através dos portões que atuam por meio da implementação das funções sigmóide e tangente hiperbólica.

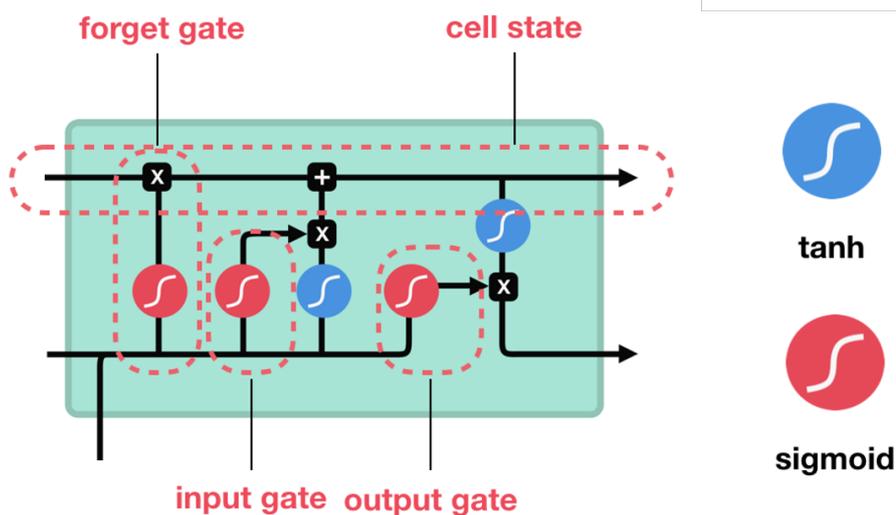


Figura 1.3: Ilustração de uma rede LSTM

Com os componentes básicos dos nosso modelo descritos, é importante entendermos como as redes neurais já vêm sendo usadas nesse ramo de pesquisa. Os modelos para sumarização abstrata de texto que utilizam RN recaem, em sua grande maioria, sobre uma categoria chamada sequência-para-sequência (seq2seq). Estes modelos mapeiam uma sequência de entrada, que pode ser um texto, para uma sequência alvo, que chamaremos de sumário. Em um passado recente, modelos seq2seq foram bem-sucedidos em tarefas como tradução automática [2], reconhecimento de fala [3] e legendas automáticas em vídeos [4]. Nos modelos seq2seq, uma arquitetura muito relevante é a rede neural “codificador-decodificador atenta” proposta em [2], que produziu desempenho estado-da-arte em tradução automática. Apesar das semelhanças, sumarização abstrata é um problema muito diferente de tradução automática. Ao contrário de tradução automática, o objetivo da sumarização abstrata (o sumário) é, tipicamente, um texto curto, que não depende do tamanho do texto fonte. Além disso, um dos principais desafios da sumarização é otimizar a compressão do texto fonte de maneira que os conceitos-chave desse texto sejam preservados, enquanto, na tradução automática, espera-se que a conversão semântica seja feita quase que de um para um sem alguma perda. Em tradução, existe uma forte noção de nível de alinhamento de palavras entre o texto origem e o texto destino, já em sumarização, essa relação é menos óbvia.

Embora modelos seq2seq tenham sido aplicados com sucesso nas tarefas citadas acima, ainda existe muito espaço para melhoria nesses modelos que a cada dia atingem novos patamares de desempenho na área de sumarização abstrata utilizando redes neurais.

1.2 JUSTIFICATIVA

Esse trabalho tem o propósito de conhecer e descobrir os desafios da área de sumarização abstrata de texto e criar bases de referência na língua portuguesa, já que existem poucos trabalhos na área com corpus brasileiros.

1.3 OBJETIVOS

Criar um sumarizador abstrato de texto seq2seq com a arquitetura codificador-decodificador e atenção-global, incluindo algumas modificações, de forma a gerar resultados para a comparação de futuros trabalhos que venham a ser feitos nesta área, na língua portuguesa.

Como resultado desse trabalho, espera-se desenvolver um sumarizador abstrato automático de texto treinado em cima do corpus de acórdãos do TCU, que tenha o desempenho próximo a outros algoritmos desenvolvidos para outras línguas.

2 REVISÃO BIBLIOGRÁFICA

Como explicado na Seção 1.1, a linha de pesquisa e desenvolvimento na área de sumarização automática vem focando em métodos abstratos já que os métodos extrativos atingiram o seu desempenho máximo por volta de 2016. A partir de 2015, os trabalhos que utilizam redes neurais com arquitetura codificador-decodificador, que será o foco desse trabalho, começaram a ganhar relevância e demonstrar resultados promissores na tarefa de sumarização abstrata, tornando-se uma alternativa interessante, em vez de métodos baseados em árvores de decisões ou grafos.

Em [5], os autores deram uma visão geral sobre o estado da arte das técnicas e artigos publicados na área de sumarização abstrata. Eles começam fazendo a distinção clara entre sumarização abstrata e sumarização extrativa. Em seguida, segmentam as áreas de pesquisa de sumarização abstrata em 4 (quatro) grandes blocos: baseado em estrutura, baseado em semântica, baseado em aprendizado profundo com redes neurais e baseado em discurso e retórica. Dentro de cada bloco, os autores descrevem as principais técnicas e trabalhos desenvolvidos e os resultados que estes atingiram, sempre levando em consideração a métrica de avaliação ROUGE-1. Por meio de tabelas informativas, os autores fizeram comparações elencando as vantagens e desvantagens das técnicas abordadas em cada bloco de forma a facilitar o entendimento do leitor das vantagens e desvantagens de cada um dos algoritmos. Os autores também fizeram uma análise e descrição das principais ferramentas usadas nos artigos dividindo-as em 3 (três) áreas: pré-processamento de texto, ferramentas de sumarização e ferramentas generalistas de processamento de linguagem natural. Por fim, eles elencaram os maiores desafios da área e direcionamentos para o futuro, sendo estes: a necessidade de indicadores quantitativos mais expressivos e que destaquem com maior relevância a qualidade dos sumários produzidos; a problemática de palavras raras dentro dos bancos de dados e como lidar com elas durante os treinamentos; a disponibilidade de banco de dados grandes e com maior diversidade de assuntos e contextos que possam ser usados para a tarefa de sumarização.

Em [6], os autores modelaram um sumarizador abstrato de texto usando uma arquitetura com redes neurais recorrentes (RNN) codificador-decodificador atentas e atingiram desempenho estado-da-arte em dois *corpora* de texto diferentes, Gigas corpus e DUC-Corpus. Eles abordaram diversas melhorias que não são modeladas adequadamente na arquitetura padrão, como modelar palavras chaves, capturar hierarquia nas estruturas de sentença para palavra e evidenciando palavras que são raras ou não vistas durante o período de treinamento. Para melhorar a representação das palavras, eles adicionaram à matriz de *word-embedding* elementos como: *tags* de partes do discurso, *tags* de nome de entidade e estatísticas de TF (*Term frequency*) e IDF (*Inverse Document Frequency*). Para modelar palavras raras ou nunca vistas pelo algoritmo, eles usaram uma estratégia de apontar para o texto original a posição dessa palavra e capturá-la do texto, ao invés de usar a abordagem tradicional de emitir um *token* “UNK” (*unknown* ou desconhecido). Para finalizar, eles também propuseram uma nova base de dados, CNN/Daily Mail corpus, que possuía

sumário de múltiplas sentenças e estabeleceram desempenho base para futuras comparações e *benchmark*. Com as melhorias proposta por eles, conseguiram ultrapassar as métricas de desempenho do estado-da-arte, na época, nos dois corpora. No Gigaword, eles atingiram o resultado de 34.6 em ROUGE-1 e, no DUC, 28.61 em ROUGE-1.

Em [7], os autores usaram um modelo base de sumarização de texto estado-da-arte com a intenção de explorar diferente mecanismos de atenção. Eles utilizaram o modelo criado pelo time da IBM Watson que possui uma estrutura de seq2seq utilizando redes neurais recorrentes. Com base nesse modelo, eles testaram 3 (três) mecanismos de atenção: atenção-global usando a pontuação do produto interno, a forma bilinear como proposta em [8] e um novo formato em que a pontuação é a projeção do estado oculto do codificador em escalar como uma métrica de “importância”. Para o treinamento de seus algoritmos eles utilizaram a base de dados Gigaword corpus. Durante o treinamento de seus algoritmos, eles perceberam que treinar uma matriz de *word-embedding* era extremamente custoso computacionalmente, devido ao tamanho do seu vocabulário, e decidiram usar uma matriz *word-embedding* previamente treinada para economizar tempo e recursos computacionais. Infelizmente, devido à falta de tempo e recurso computacional, eles, também, não conseguiram terminar o treinamento em cima de todo o corpus do Gigaword e então treinaram em cima de, apenas, uma pequena parcela do banco de dados, de forma que o resultado final foi um sumarizador que entrava em *loop* prevendo o mesmo *token* repetidas vezes. Apesar de não terem atingido grandes resultados, esse trabalho é crucial para ilustrar os desafios e dificuldades que podem ser encontradas na execução dessa tarefas e alguns caminhos alternativos que podem ser tomados.

Em [9], os autores propõem uma arquitetura de sumarizador abstrato de texto baseado no modelo codificador-decodificador equipado com um decodificador generativo. Com isso, informações estruturais implícitas nos sumários alvos são aprendidas baseadas em um modelo recorrente aleatório, que visa melhorar a qualidade da sumarização. Portanto, para a geração de um sumário, os autores utilizaram, informações estruturais latentes e variáveis discriminativas e determinísticas no processo. Para o treinamento de seu algoritmo, os autores usaram a corpora Gigaword, DUC-2004 e LCSTS, um banco de dados em chinês de sumarizações de texto curtas. Os resultados obtidos com esses modelos mostraram melhoras significativas em relação a um decodificador padrão. Em cima do banco de dados da Gigaword, o modelo sugerido teve um desempenho de 36.25 na métrica ROUGE-1 contra 32.69 do modelo padrão.

3 METODOLOGIA

Todos os procedimentos desenvolvidos nesse trabalho foram implementados por meio da programação em Python na IDE da *Google Colaboratory*, um ambiente online preparado com suporte a diversas bibliotecas de *Machine Learning* e acesso à GPU.

3.1 BANCO DE DADOS

O banco de dados utilizado foi disponibilizado por José Ferraz Neto na plataforma *Kaggle*. O banco de dados possui acórdãos proferidos pelo TCU entre os anos de 1992 até 30/08/2019, totalizando 269.198 (duzentos e sessenta e nove mil cento e noventa e oito) acórdãos dos quais, apenas, 93803 (noventa e três mil oitocentos e três) possuem seus respectivos sumários. Portanto, só foram utilizadas as observações que possuíam ambos, acórdão e sumário.

3.2 LIMPEZA DOS DADOS

Para realizar a limpeza dos dados, foram utilizados 3 (três) modelos de pré-processamento:

- Modelo 1) Eliminando todos os tipos de caracteres especiais e pontuação do texto.
- Modelo 2) Eliminando os caracteres especiais e mantendo as pontuações de final de texto junto às palavras.
- Modelo 3) Eliminando os caracteres especiais e mantendo as pontuações de final de texto como *tokens* separados.

É importante ressaltar que o pré-processamento de texto para os acórdãos e para os sumários são distintos. Como os acórdãos são textos maiores e com muito informações, o pré-processamento foi mais robusto visando diminuir a complexidade de informações mantendo apenas a mais importante, enquanto nos sumários tentou-se preservar o máximo possível do texto original. O procedimento de limpeza comum e o diferencial de cada modelo seguem a seguir:

Limpeza padrão:

Acórdão:

- Converter todos os caracteres para caixa baixa;
- Eliminar todos os caracteres especiais e os sinais de pontuações;
- Eliminar espaços duplos ou triplos;

- Eliminar os algarismos de 0 a 9;
- Eliminar palavras com 3 caracteres ou menos;
- Eliminar as “stopwords” do português adquiridas na biblioteca NLTK.

Sumário

- Converter todos os caracteres para caixa baixa;
- Eliminar espaços duplos ou triplos;
- Eliminar os algarismos de 0 a 9;
- Eliminar palavras com menos de 2 caracteres.

Modelo 1:

Sumário

- Eliminar todos os caracteres especiais e os sinais de pontuações;

Modelo 2:

Sumário:

- Eliminar todos os caracteres especiais;
- Mantém os sinais de pontuação

Modelo 3:

Sumário:

- Eliminar todos os caracteres especiais;
- Transformas os sinais de ponto final em *tokens*;

3.3 PADDING

Padding é o tratamento de texto que limita o comprimento do documento a um número de palavras ou *tokens* fixo. Isso ajuda no treinamento e na geração de sumários, pois determina o limite superior que o texto pode assumir.

Para determinar esse valor para os acórdãos e os sumários, foi feita uma análise estatística para que pelos menos 90% das observações do banco de dados se enquadrasse nesse limite. Com isso, cada modelo teve um número de *padding* definido individualmente para cada *feature*.

Modelo 1:

- Limite acórdão: 300 palavras;
- Limite sumário: 100 palavras.

Modelo 2:

- Limite acórdão: 300 palavras;
- Limite sumário: 102 palavras.

Modelo 3:

- Limite acórdão: 300 palavras;
- Limite sumário: 110 palavras.

3.4 TOKENS ESPECIAIS

Para que o nosso algoritmo saiba onde que um sumário começa e onde que ele termina, foram inseridos artificialmente nos sumários dois *tokens* para identificar essas posições. O *token* “xxbos” foi inserido no início para identificar o começo do sumário e o *token* “xxeos” foi inserido no final para identificar o fim do sumário. Dessa forma, o *token* “xxbos” é utilizado para iniciar a produção do sumário e se o algoritmo prever o *token* “xxeos” ele encerra o procedimento de produção.

Para designar uma palavra desconhecida do vocabulário do texto em que o algoritmo foi treinado ou que tem uma frequência muito baixa, foi utilizado o *token* especial “xxunk”. Com isso, o algoritmo poderá lidar com palavras desconhecidas ou, que devido a frequência tão baixa, durante o treinamento não poderiam ser aprendidas corretamente pelo modelo.

3.5 TOKENIZAÇÃO

Agora que nossos textos foram limpos e processados, devidamente limitados a um tamanho fixo e com os *tokens* especiais inseridos, está na hora de transformar os textos processados em uma linguagem que o computador entenda? números. Para isso, foi utilizado a função *Tokenizer* do Keras, que coleta todo o vocabulário do seu texto, cria um dicionário mapeando cada palavra com um número e depois substitui o texto por esse índice ou número criado. Dessa forma, o texto é transformado em um vetor de números em que cada número representa um *token* ou palavra do texto, como ilustrado na Figura 3.1

Esse procedimento foi realizado com os acórdãos e com os sumários, pois não necessariamente eles possuem o mesmo vocabulário: podem existir palavras que se encontram no acórdão, mas não existem no sumário e vice e versa.

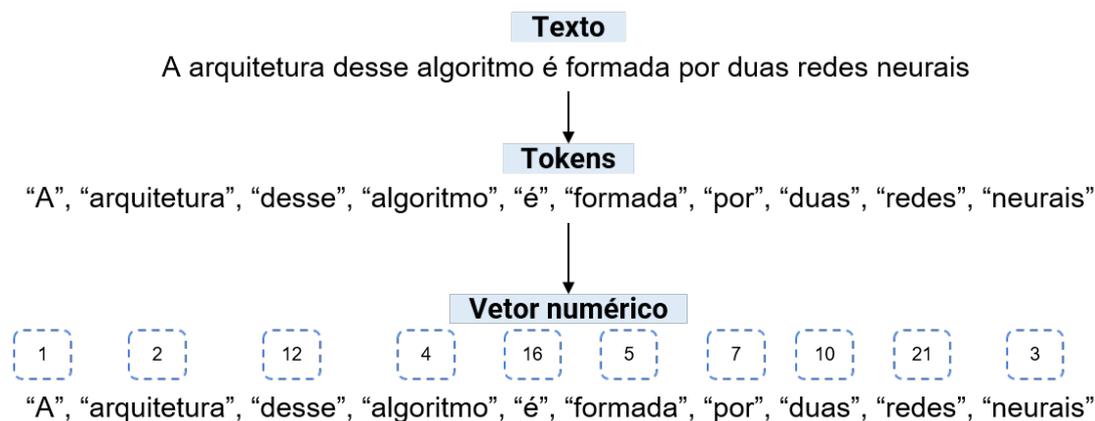


Figura 3.1: Ilustração do processo de tokenização

3.6 ARQUITETURA

A arquitetura desse algoritmo é formada por duas redes neurais, um codificador e um decodificador, que são treinados, respectivamente, para gerar um vetor de tamanho fixo que represente a semântica/mensagem do texto/acórdão que é usado para gerar uma distribuição de probabilidade em cima do vocabulário do sumário para que seja selecionada uma nova palavra, tendo como entrada o vetor gerado pelo codificador e a palavra anteriormente produzida ou inserida.

Para a implementação dessa arquitetura, as redes foram separadas em duas etapas: treinamento e inferência. Na etapa de treinamento, cada rede neural foi adaptando os parâmetros internos de sua rede a partir das repostas desejadas em cada observação do banco de dados, visando produzir sumários mais próximos e acurados em relação aos dados de treinamento e validação. Na fase de inferência, elas são conectadas para desempenhar o papel de gerar um sumário de um texto.

3.6.1 Codificador

O codificador é uma rede neural formada por 4 (quatro) camadas.

- 1ª camada: é uma camada *word-embedding* que tem a função de criar uma representação vetorial de cada palavra do vocabulário em um vetor próprio de tamanho fixo, esse processo se chama *word-to-vector*;
- 2ª, 3ª e 4ª camadas: são redes neurais LSTM's (*Long Short Term Memory*) que tem como função criar um vetor de representatividade da semântica do texto. Foram utilizadas 3 (três) para se ter uma representação mais completa do texto. Cabe esclarecer que esse número foi uma escolha arbitrária, quanto maior o número de LSTM's melhor a sua representação do texto e mais computacionalmente caro fica o treinamento da sua rede. Devido às limitações técnicas e de processamento foram escolhidas apenas 3 (três).

A Figura 3.2 ilustra o codificador processando cada palavra ou conjunto de palavras do acór-

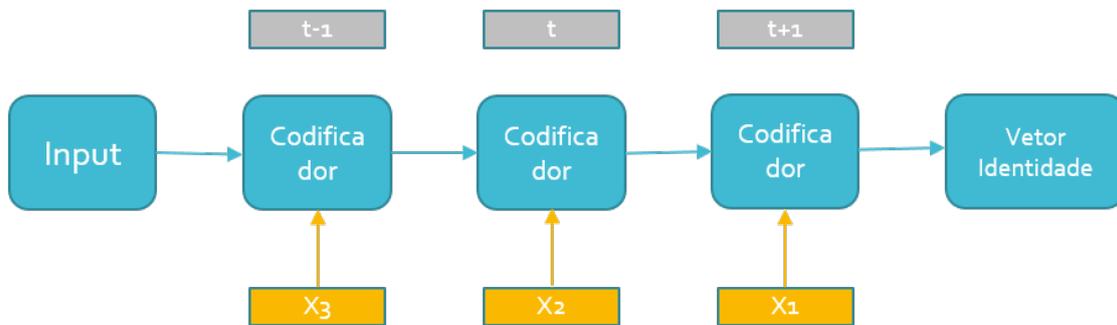


Figura 3.2: Figura ilustrativa do codificador atuando

dão em um tempo t , gerando, dessa maneira, o vetor identidade que será a representação desse texto.

3.6.2 Decodificador

O decodificador é uma rede neural formada por 3 (três) camadas.

- 1ª camada: é uma camada *word-embedding* como explicada em 3.6.1
- 2ª camada: é uma rede neural LSTM que é alimentada com os estados gerados pelo codificador e a representação de uma palavra do sumário, para então, gerar um vetor que vai ser processado por uma camada de atenção e posteriormente ser alimentada na última camada de distribuição de probabilidade;
- 3ª camada: última camada que gera uma distribuição de probabilidade em cima do vocabulário dos sumários através de uma função *softmax*.

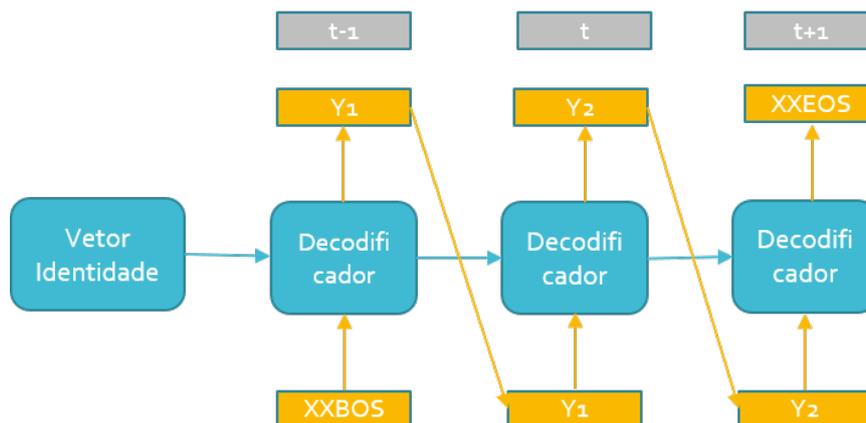


Figura 3.3: Figura ilustrativa do decodificador atuando

A Figura 3.3 ilustra o decodificador gerando uma nova palavra a partir do vetor identidade e da palavra previamente inserida ou gerada.

3.6.3 Camada de Atenção

Um das limitações desse tipo de arquitetura é que o codificador converte a entrada em um vetor de tamanho fixo e o decodificador prevê a saída a partir desse vetor. Isso funciona bem apenas para sentenças pequenas, à medida que o tamanho do resumo/sumário cresce, este começa a perder qualidade semântica.

Uma maneira encontrada para mitigar esse problema foi a implementação de uma camada de atenção que aumenta a importância de partes específicas da sequência fonte que resultaram na sequência alvo, que é ilustrada na Figura 3.4.

Essa camada, chamada Atenção-global, tem como entrada a saída da 2ª (segunda) camada do decodificador e a saída da 4ª (quarta) camada do codificador gerando com isso um novo vetor de relevância, que vai ser concatenado com a saída da 2ª (segunda) camada do decodificador, para que então esse novo vetor seja alimentado na 3ª (terceira) e última camada do decodificador e gerar a distribuição de probabilidade em cima do vocabulário dos sumários.

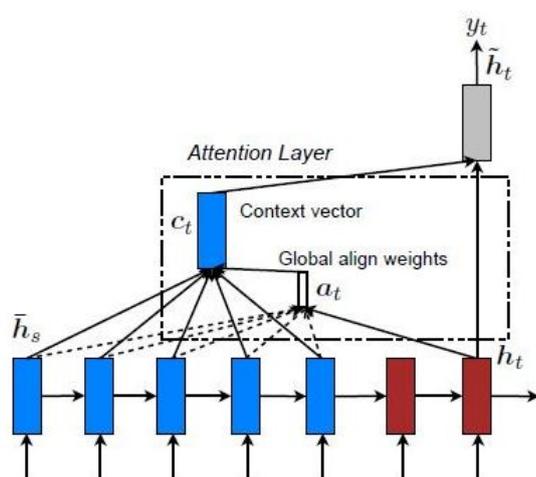


Figura 3.4: Camada de Atenção

3.7 ALGORITMO DE SELEÇÃO DA PRÓXIMA PALAVRA

Com a distribuição de probabilidade para a seleção da próxima palavra do nosso sumário em mãos, existem algumas estratégias para a seleção dessa palavra. A estratégia mais comum se chama *greedy search*, que se baseia em selecionar a palavra com maior probabilidade dentro do vocabulário. Embora esse algoritmo, na teoria, faça sentido, ele tem o problema de acabar entrando em *loops* onde ele fica prevendo a mesma palavra ou conjunto de palavras repetidas vezes e não gera aleatoriedade o suficiente para um discurso humano em linguagem natural.

Para mitigar esse problema e gerar um texto com maior aleatoriedade, e mais próximo da linguagem natural, foi utilizada a técnica escolha nuclear, que escolhe aleatoriamente uma palavra

entre um conjunto de palavras mais prováveis que possuem a probabilidade acumulada maior ou igual a um limiar determinado. Por exemplo, se determinarmos um limiar de 50% e as palavras mais prováveis têm, aproximadamente, de 9% a 10% de probabilidade de ocorrer, o algoritmo vai selecionar randomicamente uma palavra entre as 5 ou 6 que somam 50% de probabilidade dentro da distribuição, porém, se a palavra mais provável tem 60% de probabilidade, então ela é a escolha automática.

3.8 TREINAMENTO DOS MODELOS

Toda a modelagem e experimentação dos algoritmos foram feitas no Keras, uma biblioteca Python, que permite, de forma rápida e simples, testar diferentes modelos e adaptar o seu algoritmo. Para todos os 3 (três) modelos, os dados foram divididos em 20% validação e 80% treinamento. Em média foram necessárias de 12 (doze) a 15 (quinze) épocas para que o os algoritmos terminassem o seu treinamento tendo como indicador o erro de validação, quando esse parava de diminuir e começava a subir, ele interrompia o treinamento para evitar o *overfitting*.

O otimizador e a função de perda utilizados nos 3 (três) modelos foram *RSMprop* com LR de 0.001 e *Sparse Categorical Crossentropy*, respectivamente. O otimizador *RSMprop* foi escolhido como uma alternativa mais rápida e robusta que o tradicional *Gradient Descent*, pois ele altera a velocidade de aprendizado de maneira independente em cada dimensão do banco de dados, permitindo uma convergência mais rápida. A função de perda, *Sparse Categorical Crossentropy*, foi selecionada por ser uma opção em que o cálculo é feito mais rapidamente, já que ela trabalha com matrizes esparsas, levando em consideração que as matrizes do projeto são muito grandes devido à grande quantidade de *tokens* em cada texto, com isso a proposta é diminuir o tempo total de treinamento dos modelos.

A seguir, é descrito o número de épocas de cada modelo, bem como seus erros de validação e treinamento e suas acurácias de validação e treinamento.

Modelo 1

Número de épocas	Erro de treinamento	Erro de validação	Acurácia de treinamento	Acurácia de validação
12	0.8560	0.9951	82.84%	81.13%

Tabela 3.1: Resultados de treinamento do modelo 1

Modelo 2

Número de épocas	Erro de treinamento	Erro de validação	Acurácia de treinamento	Acurácia de validação
14	0.8582	1.0614	82.52%	79.98%

Tabela 3.2: Resultados de treinamento do modelo 2

Modelo 3

Número de épocas	Erro de treinamento	Erro de validação	Acurácia de treinamento	Acurácia de validação
15	0.7614	0.9473	84.26%	81.68%

Tabela 3.3: Resultados de treinamento do modelo 3

3.9 MÉTRICAS DE AVALIAÇÃO

Avaliar a qualidade de uma tradução ou de uma sumarização de texto, objetivamente, pode ser uma tarefa difícil, pois existem mais de uma maneira de resumir ou traduzir um texto. Porém é preciso ter um indicativo se a sua tradução ou sumarização está boa ou ruim. Para isso foram utilizadas duas métricas de avaliação, uma chamada BLEU que seria o paralelo da precisão e a outra, ROUGE, que seria o paralelo do *recall*.

A métrica precisão, dentro do contexto de aprendizado de máquina, busca dizer: daqueles que classifiquei como corretos, quantos efetivamente eram? Traduzindo em fórmula temos:

$$Precision = \frac{VP}{VP + FP} \quad (3.1)$$

onde:

$$VP = VerdadeiroPositivo$$
$$FP = FalsoPositivo$$

A métrica *recall*, dentro do contexto de aprendizado de máquinas, é a frequência em que o seu classificador encontra os exemplos de uma classe, ou seja, “quando realmente é da classe X, o quão frequente você classifica como X?”. Traduzindo para fórmula:

$$Recall = \frac{VP}{VP + FN} \quad (3.2)$$

onde:

$$VP = VerdadeiroPositivo$$
$$FN = FalsoNegativo$$

3.9.1 BLEU

BLEU (*bilingual evaluation understudy*) [10] é um algoritmo criado para avaliação da qualidade de textos que foram traduzidos através de uma máquina usando algoritmos de processamento

de linguagem natural. O valor de BLEU é um número entre 0 e 1. Esse valor indica o quão similar o texto é comparado a sua referência: textos com BLEU próximos de 1 são mais parecidos com as suas referências.

O BLEU usa uma forma modificada da precisão para comparar o texto candidato com a sua referência ou referências. Para o cálculo de uma precisão padrão, basta contar o número de ocorrências das palavras no texto referência (*unigrams*) e depois dividir pelo número total de ocorrências das palavras no texto candidato. Infelizmente, os sistemas de geração podem gerar palavras "razoáveis" demais, resultando em traduções improváveis, mas de alta precisão. Intuitivamente, o problema é claro: uma palavra de referência deve ser considerada esgotada após a identificação de uma palavra candidata correspondente. Formalizamos essa intuição com a precisão *n-gram* modificada. Para calcular esta, primeiro conta-se o número máximo de vezes que uma palavra ocorre em qualquer texto único de referência. Em seguida, fixa-se a contagem total de cada palavra candidata pelo máximo de sua contagem na referência, adiciona-se essas contagens cortadas e divide-se pelo número total (não cortado) de palavras no texto candidato.

A precisão *n-gram* modificada é calculada da mesma forma para qualquer *n*: todas as contagens de *n-grams* dos candidatos e suas as contagens máximas de referência correspondentes são coletadas. As contagens de candidatos são cortadas por seus valores máximos de referência correspondente, somados e divididos pelo número total de *n-grams* candidatos.

Os textos candidatos mais longos que as suas referências já são penalizados pelo algoritmo de precisão modificada, por isso, não há necessidade de penalizá-los novamente. Contudo, texto candidatos mais curtos que a sua referência não são penalizados, logo, é introduzido um fator multiplicativo de penalidade por brevidade. Com esta penalidade de brevidade em lugar, uma tradução candidata com alta pontuação deve agora corresponder às traduções de referência em comprimento e em palavras escolhidas.

3.9.2 ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [11] é um algoritmo, assim como o BLEU, criado para a avaliação automática da qualidade de sumário ou traduções comparando-as com a suas referências criadas por humanos. O ROUGE conta o número de unidades sobrepostas de *n-grams*, sequências de palavras e pares de palavras entre o sumário gerado por computador a ser avaliado e os sumários ideais criados por seres humanos. Fica claro dessa maneira que o ROUGE é uma medida relacionada ao recall porque o denominador da equação é a soma total do número de *n-grams* ocorrendo no seu conjunto de referências, diferentemente do BLEU que conta a percentagem de *n-grams* no texto candidato que se sobrepõem com suas referências.

4 RESULTADOS E DISCUSSÃO

A seguir, serão apresentadas as análises dos resultados objetivos e subjetivos de cada um dos 3 (três) modelos. As análises objetivas vão descrever o desempenho de cada um dos modelos nas métricas estabelecidas no capítulo anterior, fazendo comparação também com outros trabalhos. As análises subjetivas vão ser feitas pela curadoria do autor do TCC baseado na sua opinião dos sumários que jugou bons e ruins, juntamente com desempenho desses sumários nas métricas estabelecidas.

4.1 RESULTADOS OBJETIVOS

As métricas obtidas para avaliação foram a BLEU e a ROUGE, sendo que apenas a ROUGE tem parâmetro de comparação com outros trabalhos, pois nenhuma deles usou a BLEU. Os desempenhos das métricas são apresentados em intervalos devido à aleatoriedade do algoritmo de seleção de próxima palavra. Os resultados apresentados apresentam o intervalo de resultados de 10 interações sobre o dados de teste.

Métrica BLEU

Modelo avaliado	Corpus	Score BLEU-1
Modelo 1	Acórdãos TCU	0.3034 – 0.3366
Modelo 2	Acórdãos TCU	0.2685 – 0.2935
Modelo 3	Acórdãos TCU	0.3539 – 0.3699
feats-lvt2k-2sent-ptr [6]	Giga-words	-
DRGN [9]	Giga-words	-

Tabela 4.1: Comparativo de métrica de avaliação BLEU-1

Métrica ROUGE

Modelo avaliado	Corpus	Score ROUGE-1
Modelo 1	Acórdãos TCU	0.4006 – 0.4279
Modelo 2	Acórdãos TCU	0.3744 – 0.4497
Modelo 3	Acórdãos TCU	0.3461 – 0.4110
feats-lvt2k-2sent-ptr [6]	Giga-words	0.362
DRGN [9]	Giga-words	0.364

Tabela 4.2: Comparativo de métrica de avaliação ROUGE-1

Embora alguns modelos tenham sido aplicados em corpus diferentes, eles servem de base para um parâmetro de comparação. Os modelos propostos nesse trabalho atingiram performances

próximos do estado-da-arte em 2016 e 2017, no que se diz respeito a métrica ROUGE, lembrando que os modelos foram treinados em corpus diferentes de tamanhos diferentes e línguas diferentes.

4.2 RESULTADOS SUBJETIVOS

Dentro do espectro de sumários gerados pelos 3 (três) modelos, alguns ficaram bem próximos do seu original ou com valor semântico bem preservado, enquanto outros fugiram da proposta do texto ou perderam o seu sentido original. Dessa forma, foram selecionados em cada um dos modelos alguns exemplos que tiveram um bom resultado do ponto de vista de quem lê e interpreta o texto e alguns exemplos que divergiram da proposta inicial.

Modelo 1

Sumários produzidos de boa qualidade
<p>Sumário original: tomada de contas especial acordo lba superintendência do distrito federal omissão na prestação de contas responsável revel contas irregulares débito</p> <p>Sumário gerado: tomada de contas especial acordo lba superintendência do distrito federal omissão na prestação de contas responsável revel contas irregulares débito</p>
<p>Sumário original: tomada de contas especial omissão no dever de prestar contas revelia ausência de comprovação da aplicação dos recursos contas irregulares débito multa a ausência de comprovação da aplicação dos recursos em decorrência da omissão no dever de prestar contas do responsável importa no julgamento pela irregularidade das contas na condenação em débito e na aplicação de multa</p> <p>Sumário gerado: tomada de contas especial omissão no dever de prestar contas revelia ausência de comprovação da aplicação dos recursos contas irregulares a ausência de comprovação da aplicação dos recursos em decorrência da omissão no dever de prestar contas do responsável importa no julgamento pela irregularidade das contas na condenação em débito e na aplicação de multa</p>
<p>Sumário original: tomada de contas especial royalties prefeitura municipal de guaratinga ba omissão na prestação de contas responsável revel contas irregulares débito</p> <p>Sumário gerado: tomada de contas especial royalties prefeitura municipal de políticas salarial de goiás omissão na prestação de contas responsável revel contas irregulares débito</p>
Sumário produzidos de má qualidade
<p>Sumário original: representação rdc contratação integrada de empresa para elaboração do projeto básico e executivo de engenharia e execução das obras para duplicação de rodovia e restauração da pista existente e obras de arte especiais na rodovia br es execução da obra em desconformidade ao que estabelece o edital de licitação medida cautelar suspensão parcial da execução do contrato referendo do plenário</p> <p>Sumário gerado: representação obras de construção do contorno rodoviário no estado do Ceará concessão de medida cautelar medida cautelar suspensão do certame revogação da cautelar adotada</p>

Sumário original: pensão civil sf mapa mg pensões concedidas a beneficiários sob a condição de filhos maiores inválidos oitivas de defesa por dois interessados legalidade e registro do ato em favor de bahia mendes renúncia tácita ao direito de defesa por antônio de brito carvalho nesta fase processual ato ilegal negativa de registro elementos de defesa apresentados por antônio da silva sem os requisitos definidos na súmula tcu necessidade de diligências saneadoras para o ato de antônio da silva ciência

Sumário gerado: pessoal pensão civil decorrente de aposentadoria proporcional considerada ilegal pagamento integral da vantagem pecuniária individual vpi da lei n a gratificação de desempenho de atividade da seguridade social e do trabalho gdasst instituída pela lei n ilegalidade do ato negativa de registro determinações

Sumário original: tomada de contas especial oriunda de conversão de denúncia ilicitude na aplicação de recursos do fundef e do fnde irregularidade das contas débito e multa não provimento de sucessivos recursos recurso de revisão conhecimento não provimento oposição de embargos de declaração rejeição oposição de novos embargos declaratórios conhecimento ausência de omissões e de contradições na deliberação embargada rejeição ciência

Sumário gerado: embargos de declaração conhecimento inexistência de omissão contradição ou obscuridade rejeição ciência aos interessados

Tabela 4.3: Seleção de exemplos produzidos pelo Modelo 1

Modelo 2

Sumários produzidos de boa qualidade

Sumário original: tomada de contas especial. descentralização de verbas federais. fundo nacional de desenvolvimento da educação fnde. programa nacional de alimentação escolar. omissão quanto ao dever legal de prestar contas. citação. revelia. irregularidade das contas e condenação em débito. multa. fixação de prazo. autorização para cobrança judicial se necessária. remessa de cópia dos autos ao ministério público da união. ciência ao responsável.

Sumário gerado: tomada de contas especial. descentralização de verbas federais. fundo nacional de desenvolvimento da educação fnde. programa nacional de alimentação escolar pnae. omissão quanto ao dever legal de prestar contas. citação. revelia. irregularidade das contas e condenação em débito. multa. fixação de prazo. autorização para cobrança judicial se necessária. remessa de cópia ao ministério público da união. ciência da deliberação.

Sumário original: tomada de contas especial. funasa. recursos destinados a melhorias sanitárias. omissão no dever de prestar contas. revelia. ausência de comprovação da aplicação dos recursos. contas irregulares. a ausência de comprovação da aplicação dos recursos em decorrência da omissão no dever de prestar contas do responsável importa no julgamento pela irregularidade das contas na condenação em débito e na aplicação de multa.

<p>Sumário gerado: tomada de contas especial. omissão no dever de prestar contas. citação. revelia. ausência de comprovação da aplicação dos recursos. contas irregulares. a ausência de comprovação da aplicação dos recursos em decorrência da omissão no dever de prestar contas do responsável importa no julgamento pela irregularidade das contas na condenação em débito e na aplicação de multa.</p>
<p>Sumário original: tomada de contas especial. acordo. lba. superintendência do df. omissão na prestação de contas. responsável revel. contas irregulares. débito.</p> <p>Sumário gerado: tomada de contas especial. acordo. lba. superintendência do df. omissão na prestação de contas. responsável revel. contas irregulares. débito.</p>
<p>Sumário produzidos de má qualidade</p>
<p>Sumário original: denúncia. concurso público para preenchimento de cargos de magistério. aplicação de prova didática . decreto n° . . conhecimento. procedência parcial. emissão de alerta. retirada da chancela de sigilo. comunicações. arquivamento. . a gravação de provas xxunk prevista no art. §° do decreto n° . é impositiva e deve ser realizada diretamente pela administração não se constituindo em faculdade a ser xxunk ao alvedrio de cada candidato às suas xxunk</p> <p>Sumário gerado: denúncia. denúncia contra o tribunal regional eleitoral do brasil trt região. inspeção. audiência dos responsáveis. procedência parcial. razões de justificativa parcialmente acolhidas. procedência parcial. determinação. arquivamento.</p>
<p>Sumário original: tomada de contas especial. convênio. ausência de comprovação da boa e regular aplicação dos recursos. espólio. solidariedade de terceiro contratado. citação. contas irregulares. débito. . julgam se irregulares as contas com condenação em débito do espólio do responsável falecido em face da não aprovação da prestação de contas. . o contratado responderá solidariamente pelo débito apurado quando constatado o recebimento integral dos recursos e a não conclusão do objeto contratado</p> <p>Sumário gerado: tomada de contas especial. convênio. ministério da integração nacional. reconstrução de casas populares. execução apenas parcial do objeto pactuado. citação. revelia. contas irregulares. débito. multa. a não comprovação da regular aplicação dos recursos transferidos ao município de forma integral. . a ausência de comprovação da regular aplicação dos recursos conveniados importa no julgamento pela irregularidade das contas do responsável na sua condenação em débito e na aplicação de multa</p>
<p>Sumário original: tomada de contas especial. irregularidades praticadas no exercício da função de gerência. prejuízo à instituição bancária. contas irregulares. multa. o prejuízo decorrente de irregularidades praticadas por empregado de instituição bancária federal justifica o julgamento pela irregularidade das contas a condenação em débito e a aplicação de multa ao responsável.</p> <p>Sumário gerado: tomada de contas especial. apropriação indevida de valores. citação. revelia. contas irregulares. débito. multa. julgam se irregulares as contas com imputação de débito e aplicação de multa em virtude da apropriação de recursos da agência física em agência bancária.</p>

Tabela 4.4: Seleção de exemplos produzidos pelo Modelo 2

Modelo 3

Sumários produzidos de boa qualidade

Sumário original: tomada de contas especial . omissão no dever de prestar contas . revelia . ausência de comprovação da aplicação dos recursos . contas irregulares . a ausência de comprovação da aplicação dos recursos em decorrência da omissão no dever de prestar contas do responsável importa no julgamento pela irregularidade das contas na condenação em débito e na aplicação de multa .

Sumário gerado: tomada de contas especial . omissão no dever de prestar contas . ausência de comprovação da aplicação dos recursos . contas irregulares . a ausência de comprovação da aplicação dos recursos em decorrência da omissão no dever de prestar contas do responsável importa no julgamento pela irregularidade das contas na condenação em débito e na aplicação de multa .

Sumário original: tomada de contas especial . convênio . lba . superintendência do distrito federal . omissão na prestação de contas . responsável revel . contas irregulares . débito .

Sumário gerado: tomada de contas especial . acordo . lba . superintendência do distrito federal . omissão na prestação de contas . responsável revel . contas irregulares . débito .

Sumário original: atos de admissão . prorrogação da validade de concurso público por tempo indeterminado estabelecida judicialmente . negativa de registro . possibilidade de subsistência das admissões enquanto mantida a decisão judicial . determinações .

Sumário gerado: atos de admissão . prorrogação da validade de concurso público por tempo indeterminado estabelecida judicialmente . negativa de registro . possibilidade de subsistência das admissões enquanto mantida a decisão judicial . determinações .

Sumário produzidos de má qualidade

Sumário original: convênio . desvio de objeto multa . embargos . conhecer . não prover . . cabem embargos de declaração para corrigir omissão obscuridade ou contradição na decisão embargada . . os embargos de declaração não são a via apropriada para a rediscussão de questões de mérito já apresentadas e devidamente apreciadas e refutadas por ocasião de exame anterior não sendo suficiente para ensejar a reforma da deliberação recorrida em sua essência ou substância . . rejeitam se os embargos declaratórios quando não se confirmam ou ausentes as alegações de existência de omissões contradições ou obscuridades no julgado .

Sumário gerado: embargos de declaração . ausência de omissão obscuridade ou contradição . conhecimento . rejeição .

Sumário original: representação de empresa contratada . não pagamento de notas fiscais relativas à aquisição de equipamentos . descumprimento da ordem cronológica das datas de exigências para pagamento de obrigações . procedência . o pagamento de obrigações relativas ao fornecimento de bens locações obras e prestação de serviços deve ser xxunk de prévio empenho efetivado conforme a ordem cronológica das datas das respectivas exigências e suportado por disponibilidade orçamentária comprovada a teor do que dispõem o art . § inciso iii da lei n . . e arts . e § inciso xxunk da lei n . . .

Sumário gerado: representação . conhecimento . precedente . determinações .

Sumário original: representação sobre a ocorrência de irregularidades na aplicação de recursos federais repassados à prefeitura municipal de cerejeiras ro por meio de convênios acordos e ajustes . conhecimento por preencher os requisitos de admissibilidade . comunicações das irregularidades apuradas às autoridades competentes com vistas à instauração das respectivas tomadas de contas especiais . aplicação de multa aos responsáveis . determinações à prefeitura municipal de cerejeiras . ciência da decisão a ser adotada bem como do relatório e voto que a fundamentarem às autoridades competentes ao secretário da secretaria federal de controle interno ao presidente da câmara municipal de cerejeiras ao atual prefeito e ao vereador geraldo xxunk pereira .

Sumário gerado: representação . convênio . não atendimento de diligência . multa .

Tabela 4.5: Seleção de exemplos produzidos pelo Modelo 3

4.3 DISCUSSÃO

Analisando os resultados objetivos e subjetivos dos 3 (três) modelos propostos, pode-se dizer que estes foram satisfatórios para a proposta inicial desse trabalho, entregando sumários, em grande parte legíveis e bem conectados ao texto original. Observando as métricas de avaliação, percebe-se que as pontuações obtidas foram relativamente acima dos outros trabalhos, isso se deve, em parte, pelo uso de um banco de dados mais restrito e com um vocabulário mais específico, o que ajudou no desempenho da tarefa de sumarização.

Observando a produção dos sumários, é notado que, apesar da implementação de uma camada de atenção entre as redes, o problema de gerar sumários maiores continua sendo uma barreira e ponto de melhoria dentro do algoritmo, pois textos maiores acabam tendo seus sumários prejudicados, como mostram os exemplos expostos na Seção 4.2.

Comparando os modelos entre si, percebe-se que as métricas BLEU e ROUGE favorecem os modelos de formas diferentes. O modelo 3, na média, teve o melhor desempenho na métrica BLEU, o modelo 1, na média, teve o melhor desempenho na métrica ROUGE e o modelo 2 apresentou um meio termo entre os 3 (três). Contudo, olhando para a compreensão e a sintaxe dos sumários gerados, percebe-se que o modelo 1, apesar de ter uma das melhores métricas, apresenta um nível de compreensão e sintaxe reduzidos devido às simplificações que foram aplicadas a ele. Enquanto o modelo 2, apresentou o melhor resultado do ponto de vista sintático e de compreensão do leitor.

5 CONCLUSÃO

Com a conclusão do trabalho e a análise dos resultados, percebe-se que, embora com resultado promissores, a área de pesquisa em sumarização abstrata de texto é relativamente nova e apresenta grandes desafios a serem superados, como a limitação do tamanho do sumário e criar sumarizadores cada vez mais genéricos e menos específicos a uma aplicação. Com grande potencial de crescimento, a área possibilitará futuramente a criação de novas ferramentas e utilidades para o ser-humano.

Tendo em vista o trabalho desenvolvido, sua continuidade e a diversidade de temas relacionados de grande relevância que necessitam de aprofundamento, são deixadas como propostas para trabalhos futuros:

- Implementação de um decodificador generativo [9] aos modelos apresentados;
- Implementação de melhorias na modelagem do texto através das sugestões aplicadas em [6];
- Fazer uma análise quantitativa visando analisar qual é o tamanho máximo que o sumário pode ter antes de perder qualidade semântica e sintática;
- Implementar um modelo sem nenhuma espécie de pré-processamento de texto e comparar com os modelos aqui apresentados;
- Utilizar uma matriz de *word-embedding* previamente treinada em um banco de dados robusto da língua portuguesa e testar nos modelos apresentados, comparando os resultados obtidos.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] H. P. Edmundson, “New methods in automatic extracting,” *Journal of the ACM (JACM)*, p. 16(2):264–285, 1969.
- [2] Y. B. Dzmitry Bahdanau, Kyunghyun Cho, “Neural machine translation by jointly learning to align and translate,” *arXiv*, 2014.
- [3] D. S. P. B. Y. B. Dzmitry Bahdanau, Jan Chorowski, “End-to-end attention-based large vocabulary speech recognition,” *IEEEExplor*, 2015.
- [4] J. D. R. M. T. D. K. S. Subhashini Venugopalan, Marcus Rohrbach, “Sequence to sequence – video to text,” *arXiv*, 2015.
- [5] S. G. Som Gupta, “Abstractive summarization: An overview of the state of the art,” *Expert Systems With Applications 121 (2019) 49–65*, 2019.
- [6] C. N. d. s. C. G. B. X. Ramesh Nallapati, Bowen Zhou, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” *IBM Watson*, 2016.
- [7] E. Jobson and A. Gutiérrez, “Abstractive text summarization using attentive sequence-to-sequence rnns,” *STANFORD*, 2017.
- [8] M.-T. L. H. P. C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv*, 2015.
- [9] L. B. Z. W. Piji Liy, Wai Lamy, “Deep recurrent generative decoder for abstractive text summarization,” *Sub-Lab, CUHK*, 2017.
- [10] T. W. Kishore Papineni, Salim Roukos and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” *ACL*, 2002.
- [11] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *University of Southern California*, 2004.

I.1 LINK PARA O GITHUB COM OS CÓDIGOS DO PROJETO

Para acessar o github com os códigos utilizados nesse trabalho clique no [link](#).