TECNOLOGIAS, MÉTODOS E TEORIAS NA ENGENHARIA DE COMPUTAÇÃO

ERNANE ROSA MARTINS
(ORGANIZADOR)

Atena Editora
Ano 2020

TECNOLOGIAS, MÉTODOS E TEORIAS NA ENGENHARIA DE COMPUTAÇÃO

ERNANE ROSA MARTINS
(ORGANIZADOR)

Atena Editora
Ano 2020

Editora Chefe
Profª Drª Antonella Carvalho de Oliveira
Assistentes Editoriais
Natalia Oliveira
Bruno Oliveira
Flávia Roberta Barão
Bibliotecário
Maurício Amormino Júnior
Projeto Gráfico e Diagramação
Natália Sandrini de Azevedo
Camila Alves de Cremo
Karine de Lima Wisniewski
Luiza Alves Batista
Maria Alice Pinheiro
Imagens da Capa
Shutterstock
Edição de Arte
Luiza Alves Batista
Revisão
Os Autores

Atena Editora
Ano 2020

Atena Editora
Ano 2020

Atena Editora
Ano 2020

Atena Editora
Ano 2020

Atena Editora
Ano 2020

Atena Editora
Ano 2020

Tecnologias, métodos e teorias na engenharia de computação

Editora Chefe: Profª Drª Antonella Carvalho de Oliveira
Bibliotecário Maurício Amormino Júnior
Diagramação: Karine de Lima Wisniewski
Edição de Arte: Luiza Alves Batista
Revisão: Os Autores
Organizador: Ernane Rosa Martins

Dados Internacionais de Catalogação na Publicação (CIP)
(eDOC BRASIL, Belo Horizonte/MG)

T255 Tecnologias, métodos e teorias na engenharia de computação [recurso eletrônico] / Organizador Ernane Rosa Martins. – Ponta Grossa, PR: Atena, 2020.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-5706-361-3
DOI 10.22533/at.ed.613200409

1. Computação – Pesquisa – Brasil. 2. Tecnologia. I. Martins, Ernane Rosa.

CDD 004

Elaborado por Maurício Amormino Júnior – CRB6/2422

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

Atena Editora
Ano 2020

Ciências Exatas e da Terra e Engenharias

APRESENTAÇÃO

A Engenharia de Computação é a área que estuda as técnicas, métodos e ferramentas matemáticas, físicas e computacionais para o desenvolvimento de circuitos, dispositivos e sistemas. Esta área tem a matemática e a computação como seus principais pilares. O foco está no desenvolvimento de soluções que envolvam tanto aspectos relacionados ao software quanto à elétrica/eletrônica. O objetivo é a aplicação das tecnologias de computação na solução de problemas de Engenharia. Os profissionais desta área são capazes de atuar principalmente na integração entre software e hardware, tais como: automação industrial e residencial, sistemas embarcados, sistemas paralelos e distribuídos, arquitetura de computadores, robótica, comunicação de dados e processamento digital de sinais.

Dentro deste contexto, esta obra aborda os mais diversos aspectos tecnológicos computacionais, tais como: desenvolvimento de um método de verificação biométrica de indivíduos; uma abordagem para encontrar evidências de fraude aplicando técnicas de mineração de dados a bancos de dados públicos das licitações do governo federal brasileiro; o desenvolvimento de um método computacional para a classificação automática de melanomas; a aplicação de algoritmos recentes de aprendizagem de máquina, denominados XGBoost e Isolation Forest, para predição de irregularidades no consumo de energia elétrica; um modelo de receptor 5-HT2C humano que foi criado através de modelagem por homologia e estudos de acoplamento molecular com os ligantes ácido fúlvico, paroxetina, citalopram e serotonina; a análise do uso do Controlador Lógico Programável (CLP), apresentando sua composição (estrutura, programação e linguagem Ladder), montagem, vantagens e desvantagens, exemplo de tipos e fabricantes; uma sugestão de melhoria das etapas de análise de negócios e engenharia de requisitos, por meio do uso de conceitos viáveis de metodologias ágeis; a construção de um aplicativo, denominado QEnade, para a disponibilização de questões do ENADE para os estudantes; uma síntese conceitual do PC voltada para âmbito educacional referente à educação básica brasileira; um sistema de localização híbrido capaz de usar diferentes tecnologias para fornecer a localização interna e externa de robôs ou de outros dispositivos móveis; um sistema de sumarização multidocumento de artigos de notícias escritos em português do Brasil; o emprego de duas técnicas de aprendizado de máquinas para prever se parte do público infantojuvenil da cidade de Monte Carmelo esta suscetível a algum risco ou situação constrangedora nas redes sociais; a identificação das principais tecnologias que estão sendo utilizadas no contexto de Transformação Digital no cenário mundial; os elementos utilizados na construção de um sistema computacional, sem custo financeiro para a instituição e de fácil compreensão para o usuário, que utiliza os conhecimentos estatísticos para realizar a descrição, a apresentação e análise dos dados coletados; uma discussão acerca da confiabilidade das informações disseminadas na internet, para

entender os riscos e a importância da avaliação dos conteúdos encontrados no ambiente virtual; uma proposta de estratégia para a navegação de robôs semiautônomos baseada apenas em informações locais, obtidas pelos sensores instalados no robô e um planejador probabilístico que gera caminhos a serem seguidos localmente por ele, garantindo assim o desvio de obstáculos.

Sendo assim, está obra é significativa por ser composta por uma gama de trabalhos pertinentes, que permitem aos seus leitores, analisar e discutir diversos assuntos importantes desta área. Por fim, desejamos aos autores, nossos mais sinceros agradecimentos pelas significativas contribuições, e aos nossos leitores, desejamos uma proveitosa leitura, repleta de boas reflexões.

Ernane Rosa Martins

SUMÁRIO

CAPÍTULO 12

SUMARIZAÇÃO AUTOMÁTICA DE ARTIGOS DE NOTÍCIAS EM PORTUGUÊS USANDO PROGRAMAÇÃO LINEAR INTEIRA E REGRESSÃO

Data de aceite: 27/08/2020
Data de submissão: 05/06/2020

Hilário Tomaz Alves de Oliveira
Instituto Federal do Espírito Santo – Campus Serra
Serra - Espírito Santo
http://lattes.cnpq.br/8980213630090119

Laerth Bruno de Brito Gomes
Centro Universitário de João Pessoa – Unipê
Joao Pessoa – Paraíba

RESUMO: Sumarização Automática de Textos (SAT) tem por objetivo desenvolver métodos capazes de gerar resumos contendo as informações mais relevantes a partir de um ou mais documentos. Sistemas de SAT podem auxiliar no processo de identificar informações de interesse de maneira mais eficiente. Neste trabalho, apresentamos um sistema de sumarização multidocumento de artigos de notícias escritos em português do Brasil. A solução proposta inicialmente cria múltiplos resumidos candidatos usando uma abordagem baseada em conceitos com programação linear inteira. Posteriormente, um modelo de regressão é aplicado para estimar e selecionar o resumo candidato mais informativo. Experimentos realizados no corpus CSTNews demonstram que o sistema desenvolvido obteve resultados promissores comparados com outros sistemas da literatura, com base nas medidas de avaliação do ROUGE.

PALAVRAS-CHAVE: Sumarização automática de textos; Programação linear inteira; Regressão.

AUTOMATIC SUMMARIZATION OF NEWS ARTICLES IN PORTUGUESE USING INTEGER LINEAR PROGRAMMING AND REGRESSION

ABSTRACT: Automatic Text Summarization (ATS) aims to develop methods capable of generating summaries containing the most relevant information from one or more documents. ATS systems can assist in the process of identifying information of interest more efficiently. In this paper, we present a multi-document summarization system of news articles written in Brazilian Portuguese. The proposed solution initially creates multiple summary candidates using a concept-based approach with integer linear programming. Afterward, a regression model is applied to estimate and select the most informative candidate summary. Experiments carried out in the CSTNews corpus demonstrate that the developed system obtained encouraging results compared to other systems in the literature, based on the ROUGE evaluation measures.

KEYWORDS: Automatic text summarization; Integer linear programming; Regression.

1 I INTRODUÇÃO

Ao longo das últimas décadas, com o desenvolvimento tecnológico e o advento da era da informação, as pessoas passaram a ter a possibilidade de criar, compartilhar e

acessar informações rapidamente. Com isso, uma quantidade sem precedentes de novas informações é criada diariamente na forma de artigos de notícias, postagens em redes sociais, artigos científicos, entre outros. Essa vasta quantidade de dados também causa uma sobrecarga de informações, já que essa abundância excede a capacidade humana de processamento. Esse cenário tem se tornado um problema para as pessoas.

A Sumarização Automática de Textos (SAT) pode desempenhar um papel fundamental para mitigar os efeitos da sobrecarga de informação (MANI, 2001). O objetivo da SAT é criar sistemas computacionais capazes de gerar um resumo contendo as informações mais relevantes a partir de um ou mais documentos textuais (NENKOVA e MCKEOWN, 2012). Dessa forma, as pessoas podem ler o resumo que é uma representação condensada das fontes de entrada e, posteriormente, decidir se os documentos são relevantes para o seu interesse.

A SAT é uma tarefa bem conhecida pela comunidade de Processamento de Linguagem Natural (PLN) e tem demandado muita pesquisa ao longo dos anos. Em especial, um maior interesse tem sido dado nos últimos anos a tarefa de Sumarização Multidocumento (SMD) que consiste na criação de um único resumo a partir de vários documentos relacionados ao mesmo assunto. Muitos sistemas de SMD têm sido desenvolvidos adotando uma abordagem extrativa que identifica e seleciona o subconjunto de frases mais relevantes dos documentos de entrada para criação do resumo, evitando a inclusão de informações redundantes.

A maioria dos sistemas de SMD extrativos tem focado em estimar um escore de relevância para frases ou n-gramas pertences aos documentos a serem sintetizados. Para isso, uma grande variedade de técnicas supervisionadas e não supervisionadas tem sido explorada. Contudo, poucos trabalhos têm investigado a viabilidade de estimar a relevância de um resumo já construído (OLIVEIRA et al., 2017). Dessa forma, seria possível selecionar o resumo mais informativo a partir de um conjunto de resumos candidatos.

Neste trabalho, apresentamos um sistema para sumarização multidocumento de artigos de notícias escritos em português do Brasil. Inicialmente, diversos resumos candidatos são criados usando um método de sumarização baseado em conceitos via Programação Linear Inteira (PLI). Em uma segunda etapa, um modelo de regressão é usado para estimar a cobertura de informações relevantes (informatividade) dos resumos candidatos. Dez atributos baseados em indicadores de relevância, medidas de similaridade e divergência são extraídos a partir da análise do resumo e da coleção de documentos de entrada para construção do modelo. Por fim, o resumo estimado como mais informativo é selecionado.

Para avaliar a efetividade do sistema desenvolvido, experimentos foram executados utilizando o corpus CSTNews (CARDOSO et al. 2011) que é comumente utilizado para avaliar sistemas de sumarização de artigos de notícias escritos em português do Brasil. Como medida de informatividade dos resumos, analisamos as medidas de cobertura,

precisão e medida-F geradas a partir da ferramenta do *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) (LIN, 2004). Os resultados experimentais demonstram que a estratégia de gerar diversos resumos candidatos e, posteriormente, aplicar um modelo de regressão para estimar a informatividade dos candidatos é viável e resultou na criação de resumos mais informativos.

2 | TRABALHOS RELACIONADOS

Diversos trabalhos investigando o desenvolvimento de sistemas de sumarização multidocumento extrativos podem ser encontrados na literatura, especialmente para documentos escritos em inglês. Um dos primeiros sistemas de SMD para o português do Brasil foi o GistSumm (PARDO, 2005). Nesse sistema, cada sentença dos documentos de entrada recebe um escore de relevância que pode ser gerado com base na frequência de suas palavras ou usando um método baseado no tradicional *Term Frequency - Inverse Document Frequency* (TF-IDF). A frase com escore de relevância mais alto é selecionada como principal. Posteriormente, o sistema seleciona as frases que contêm pelo menos uma palavra em comum com a frase principal e têm uma relevância superior a um determinado limiar, que é a média de todos os escores das frases.

Castro Jorge e Pardo (2010) introduziram uma abordagem para SMD que explora relações semânticas e métodos de seleção de conteúdo com base no modelo da Teoria da Estrutura de Documentos Cruzados, do inglês *Cross-document Structure Theory* (CST). O sistema proposto, chamado *CSTSumm* (*CSTbased SUMMarizer*), aborda alguns dos principais problemas da tarefa de sumarização de múltiplos documentos, como redundância. Cardoso e Pardo (2016) desenvolveram uma pesquisa semelhante para criar métodos de sumarização multidocumento explorando modelos de CST e da Teoria da Estrutura Retórica, do inglês *Rhetorical Structure Theory* (RST). Ambos os modelos foram adotados para identificar informações relevantes nos documentos a serem sintetizados.

Sodré e Oliveira (2019) apresentaram um sistema supervisionado para sumarização de múltiplos documentos que combina diversos indicadores de relevância baseados em frequência, posição, centralidade, entre outros. Um modelo de regressão foi desenvolvido para estimar a relevância das frases. Posteriormente, aplicando um método de seleção guloso, as frases mais relevantes são utilizadas para geração dos resumos desde que elas não tenham uma similaridade maior que uma dado limiar com sentenças já selecionadas.

Recentemente, abordagens baseadas em conceito usando Programação Linear Inteira (PLI) têm alcançado um desempenho do estado da arte para tarefas de SMD (OLIVEIRA et al., 2018; GOMES e OLIVEIRA, 2019), especialmente para documentos escritos em inglês. Tais abordagens representam o processo de sumarização como um problema de otimização de máxima cobertura, cujo objetivo é extrair o subconjunto de frases que possuem o maior número de fragmentos textuais relevantes (chamados de

conceitos), respeitando o tamanho máximo do resumo desejado.

Gomes e Oliveira (2019) desenvolveram um sistema de sumarização multidocumento baseado em conceitos usando PLI para textos de notícias em português. O sistema proposto utiliza bigramas como conceitos e aplica métodos baseados na quantidade de documentos (frequência dos documentos) e na posição das frases que mencionam o bigrama para estimar a sua relevância.

Todos os trabalhos supracitados usam um conjunto de parâmetros e técnicas predefinido para mensurar a relevâncias das frases ou n-gramas e, posteriormente, construir o resumo selecionando as sentenças consideradas mais importantes. Conforme apontado por trabalhos recentes (OLIVEIRA et al., 2017), usar um único método de sumarização é uma significante limitação, já que, em geral, ele não consegue gerar resumos informativos para todos os documentos, mesmo quando eles pertencem ao mesmo domínio. O sistema desenvolvido neste trabalho busca mitigar essa limitação explorando a estratégia de gerar múltiplos resumos candidatos e, em uma segunda etapa, identificar e selecionar o resumo mais informativo utilizando um modelo de regressão.

3 I SISTEMA DE SUMARIZAÇÃO USANDO PLI E REGRESSÃO

O sistema desenvolvido neste trabalho é composto por três etapas principais, conforme apresentado na Figura 1. Dada uma coleção de artigos de notícias, inicialmente os documentos são processados para estruturá-los em um formato adequado. Posteriormente, diversos resumos candidatos são gerados aplicando uma extensão do método de sumarização baseado em conceitos usando PLI desenvolvido por Gomes e Oliveira (2019). Na última etapa, um modelo de regressão é usado para estimar a informatividade de cada resumo candidato, e aquele com maior relevância estimada é selecionado como o resumo mais representativo.

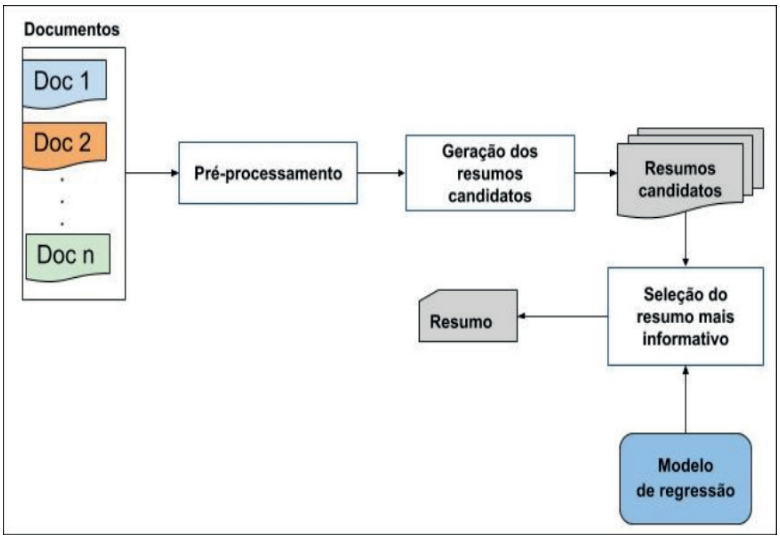

Figura 1. Visão geral das etapas do sistema proposto.

Os documentos de entrada estão escritos em linguagem natural, ou seja, estão em um formato não estruturado. Esse tipo de formato requer estruturação prévia para que eles possam ser usados para a aplicação dos algoritmos de sumarização. Portanto, na primeira etapa é realizado o pré-processamento dos documentos usando técnicas de PLN. As seguintes técnicas foram utilizadas: divisão de frases, tokenização, remoção de *stopwords* e *stemming*. Todas as técnicas foram implementadas usando as ferramentas *spaCy* e *Natural Language Toolkit* (NLTK), ambas na linguagem de programação Python.

Após a realização da etapa de pré-processamento, cada documento de entrada é dividido em uma ou mais frases, com cada frase possuindo uma ou mais palavras, e cada palavra possui associado o seu radical identificado pelo processo de *stemming* e uma marcação indicando se ela é uma *stopword*.

As etapas de geração dos resumos candidatos e da seleção do resumo mais informativo são explicadas nas subseções a seguir.

3.1 GERAÇÃO DOS RESUMOS CANDIDATOS

O objetivo desta etapa é gerar um conjunto de resumos candidatos que possam ser possíveis representações das informações mais relevantes dos documentos de entrada. Para isso, o sistema de sumarização multidocumento baseado em conceitos usando PLI desenvolvido por Gomes e Oliveira (2019) foi estendido. Esse sistema foi escolhido por ter obtido resultados competitivos em comparação com outros trabalhos na literatura, além de permitir a exploração de diferentes configurações visando gerar um conjunto diversificado de resumos candidatos.

A abordagem de sumarização baseada em conceitos usada neste trabalho considera

o processo de sumarização como um problema de máxima cobertura, cujo objetivo é selecionar o subconjunto de frases que maximize a seleção de elementos textuais (conceitos) considerados relevantes, respeitando o tamanho máximo do resumo desejado. Para isso, duas questões que precisam ser definidas é que elemento textual será usado para representar a noção de um conceito e como mensurar a sua relevância. A abordagem desenvolvida é realizada em três subetapas, descritas a seguir:

Extração dos conceitos. Nesta primeira subetapa, n-gramas são extraídos para cada frase dos documentos de entrada, como formas de representação de conceitos para serem usados no processo de sumarização. Com o objetivo de gerar diversos resumos candidatos, adotamos cinco formas de representações separadamente, sendo que cada uma delas é responsável pela geração de um resumo. As formas de representação utilizadas neste trabalho são: unigramas, bigramas, trigramas, quadrigramas e uma combinação de todas as anteriores em uma só representação.

O processo de construção dos n-gramas é feito a partir das palavras de cada sentença, sendo que cada palavra é representada pelo seu radical extraído processo de *stemming*. Como adotado em trabalhos anteriores (OLIVEIRA et al., 2018), conceitos formados somente por *stopwords* são removidos. Ao final desta subetapa, cada frase possui cinco listas de conceitos, sendo uma para cada forma de representação.

Ponderação dos conceitos. Após a extração dos conceitos, é necessário realizar o processo de ponderação visando gerar um escore (peso) que represente a sua relevância. Para isso, utilizamos o método não supervisionado de ponderação proposto por Oliveira et al. (2018) que é computado conforme apresentado na Equação 1. Nesse método, um conceito é considerado relevante se ele for mencionado em muitos dos documentos a serem sumarizados e se ele for mencionado pela primeira vez nas frases iniciais desses documentos.

$$\text{Peso}(c_i) = \text{FreqDoc}(c_i) \times SentPos(c_i) \quad (1)$$

$$SentPos(c_i) = \sum_{d_i \in D} 1 - \frac{I_{S_{c_i}}}{|S_{d_i}|} \quad (2)$$

Nessa equação, **Peso (c_i)** é o escore de relevância gerado para o conceito $\mathbf{c_i}$; **FreqDoc(c_i)** é o método de frequência dos documentos que retorna o total de documentos $\mathbf{d_i}$ da coleção de documentos de entrada **D** que mencionam o conceito $\mathbf{c_i}$ no seu texto; $|S_{d_i}|$ é o total de frases do documento $\mathbf{d_i}$; $I_{S_{c_i}}$ é o índice da primeira frase que contém o conceito $\mathbf{c_i}$ em cada documento $d_i \in D$, com a contagem das frases iniciando com zero.

Ao final desta subetapa, cada conceito presente nas cinco listas de conceitos possui um escore que representa a sua relevância.

Geração do resumo. Esta última subetapa é responsável pela seleção das frases para compor os resumos candidatos. Frases com contendo menos do que dez palavras são removidas por serem consideradas muito curtas para serem inseridas no resumo. Após essa filtragem, as frases remanescentes são usadas para construir o modelo de PLI baseado em conceitos, conforme descrito na Equação 3.

$$Maximize \sum_i w_i \times c_i \quad (3a)$$

$$s.t. \sum_j l_j \times s_j \leq L \quad (3b)$$

$$s_j \times Occ_{ij} \leq c_i \; \forall \, i, j \quad (3c)$$

$$\sum_j s_j \times Occ_{ij} \geq c_i \; \forall \, i, j \quad (3d)$$

$$c_i, s_j, Occ_{ij} \in \{0, 1\} \; \forall \, i, j \quad (3e)$$

Nessa equação, $\mathbf{w_i}$ representa o peso de um conceito $\mathbf{c_i}$ computado conforme apresentado na Equação 1. A variável $\mathbf{Occ_{ij}}$ indica a ocorrência de um conceito $\mathbf{c_i}$ na frase $\mathbf{s_j}$. A variável $\mathbf{l_j}$ representa a quantidade de palavras da sentença $\mathbf{s_j}$ e **L** é a quantidade máxima de palavras que o resumo a ser gerado pode ter. As Equações 3c e 3d são restrições que garantem a consistência do modelo, assegurando que, se uma frase for selecionada para a solução do modelo, todos os seus conceitos também devem ser selecionados, e um conceito só é selecionado se estiver presente em pelo menos uma frase selecionada. Um valor binário é atribuído a cada sentença, sendo o valor 1 atribuído a frases selecionadas para compor o resumo e 0, caso contrário. Por fim, as frases são ordenadas usando um índice gerado na etapa de pré-processamento, que indica a posição da frase no documento ao qual ela pertence.

O modelo baseado em conceitos é executado cinco vezes para gerar cinco resumos para cada uma das formas de representação de conceitos adotada neste trabalho. Ao final desta etapa, um conjunto com cinco resumos candidatos é gerado para a coleção de documentos de entrada. Esses candidatos representam diferentes possíveis resumos contendo as informações mais relevantes dos documentos a serem resumidos.

3.2 SELEÇÃO DO RESUMO MAIS INFORMATIVO

Nesta etapa, um algoritmo de regressão é aplicado para estimar a cobertura de informações relevantes (informatividade) de cada um dos resumos candidatos gerados na etapa anterior. Por ser uma tarefa de aprendizado supervisionado, a criação desse modelo de regressão requer a disponibilidade de um conjunto de treinamento contendo resumos

anotados com algum escore que represente a sua informatividade. Neste trabalho, nós avaliamos a utilização das medidas de cobertura, precisão e medida-F, como possíveis valores para representar a informatividade dos resumos candidatos. Essas medidas foram escolhidas por serem comumente usadas na literatura para avaliar sistemas de sumarização. Dessa forma, o objetivo do modelo desenvolvido é gerar um escore que represente uma estimativa da relevância de um dado resumo.

Identificar atributos para a construção do modelo de regressão é uma tarefa essencial para o seu desempenho. Por isso, selecionamos um conjunto com dez atributos que apresentaram bons resultado, conforme reportado em (OLIVEIRA et al., 2017) para sumarização de artigos de notícias escritos em inglês. Esses atributos são baseados em indicadores de relevância, como frequência e posição, sendo extraídos em nível de resumo, frase e conceitos. Alguns dos atributos adotados são gerados aplicando medidas de similaridade e divergência entre a lista de pesos de todos os conceitos presentes nos documentos de entrada e no resumo. Os dez atributos utilizados para construção do modelo são:

- **Palavras únicas:** computado pelo quociente entre a quantidade de palavras distintas no resumo, em relação ao total de palavras do resumo.
- **Palavras nos títulos dos documentos:** representa a proporção do total de palavras do resumo que estão nos títulos dos documentos, com *stopwords* sendo desconsideradas.
- **Palavras nas primeiras frases:** calculado pelo quociente entre as palavras do resumo que estão na primeira frase de cada um dos documentos de entrada, com *stopwords* sendo removidas.
- **Quantidade de frases:** é o total de frases do resumo.
- **Posição das frases:** computado pela média da medida de posição das sentenças (Equação 2) presentes no resumo.
- **Sobreposição dos pesos:** essa medida representa a sobreposição de pesos dos conceitos presentes no resumo e nos documentos de entrada, sendo computada conforme apresentado na Equação 4.

$$Sobreposicao(D,R) = \sum_{c_i \in D} Min(D_{c_i}, R_{c_i}) \qquad (4)$$

Nessa equação ***D*** é a lista de conceitos e seus pesos dos documentos de entrada, ***R*** é a lista de pesos dos conceitos presentes no resumo, $\mathbf{D}_{ci}$ é o peso do conceito $\mathbf{c}_i$ em **D**, e $\mathbf{R}_{ci}$ é o peso do conceito $\mathbf{c}_i$ em **R**.

- **Similaridade com outros resumos candidatos:** A medida de similaridade do cosseno é computada entre o resumo candidato e os outros resumos candidatos gerados.

- **Medidas de divergência:** As medidas de divergência Kullback-Leibler (KL) (KULLBACK, 1959), Jensen-Shannon (JS) (LIN, 2006) e Coeficiente de Divergência (SHIRKHORSHIDI; AGHABOZORGI; WAH, 2015) são computadas entre a lista de pesos dos conceitos presentes no resumo candidato e nos documentos de entrada.

Ao final desta fase, o resumo candidato com a maior escore de relevância estimado pelo modelo de regressão é selecionado como o resumo contendo mais informações relevantes dos documentos de entrada.

4 | EXPERIMENTOS

Nesta seção são apresentados e discutidos os experimentos realizados para avaliar o desempenho do sistema de SAT desenvolvido neste trabalho. Um primeiro experimento foi executado para avaliar cinco algoritmos de regressão para estimar a informatividade dos resumos candidatos. Por fim, um segundo experimento comparou o desempenho obtido pelo sistema desenvolvido em comparação com outros trabalhos da literatura.

Neste trabalho, utilizamos o corpus do CSTNews (CARDOSO et al. 2011) que foi criado para auxiliar pesquisas na área de SAT de artigos de notícias escritos em português do Brasil. Esse corpus possui cinquenta grupos de notícias, e em cada grupo existem aproximadamente três documentos de fontes diferentes abordando o mesmo assunto. Cada grupo de notícias possui cinco resumos extrativos criados para serem utilizados na avaliação de sistemas de sumarização multidocumento.

Para avaliar os resumos criados foram adotadas as medidas de cobertura, precisão e medida-F geradas a partir da ferramenta do ROUGE (LIN, 2004), usando uma variação chamada de ROUGE-1 (R-1). Essa variação computa a sobreposição de unigramas entre os resumos gerados automaticamente e os resumos de referência. Como cada grupo de documentos do corpus CSTNews possui mais de um resumo de referência, os valores das medidas de avaliação são gerados computando a média entre a comparação do resumo gerado automaticamente com todos os resumos de referência. O R-1 foi computado utilizando a ferramenta *py-rouge*, aplicando *stemming* e sem remoção de *stopwords*.

Os resumos gerados pelo sistema desenvolvido neste trabalho possuem no máximo de 110 palavras. Esse limiar foi escolhido para gerar resumos com tamanhos compatíveis com outros sistemas da literatura que geram resumos com aproximadamente de 100 palavras.

4.1 AVALIAÇÃO DOS ALGORITMOS DE REGRESSÃO

Neste primeiro experimento, avaliamos a aplicação de cinco algoritmos de aprendizado de máquina para realizar a estimação da informatividade dos resumos candidatos. Os seguintes algoritmos disponíveis na ferramenta *Scikit-learn* foram avaliados: Regressão Bayesiana, Regressão Linear, Regressão Ridge, Máquina de vetores de suporte do inglês *Support Vector Regression* (SVM) com kernel linear (SVR-Linear). Esses algoritmos foram usados com suas configurações padrões, ou seja, nenhuma calibração de parâmetros foi realizada.

Para avaliar o desempenho dos algoritmos utilizamos a metodologia de validação cruzada com k conjuntos (*k-fold cross validation*). O parâmetro *k* foi definido com o valor cinquenta (k=50) que é total de grupos de artigos de notícias do corpus CSTNews. Dessa forma, para cada grupo de documentos, o processo de geração do resumo é executado de acordo com as duas etapas a seguir:

- **Treinamento:** Nesta etapa, 49 grupos de artigos são usados para treinar o modelo de regressão. Como cada grupo possui 5 resumos candidatos, um total de 245 resumos são usados para treinamento.
- **Teste:** O grupo de documentos não selecionado na etapa de treinamento é usado para testar o modelo de regressão criado na etapa.

Na Tabela 1 são apresentados os resultados deste experimento considerando as medidas de cobertura, precisão e medida-F do R-1 dos resumos selecionados como mais informativos. Essas três medidas também foram avaliadas como possíveis escores de informatividade dos resumos (atributos alvos).

Algoritmos	Atributo Alvo	Cobertura	Precisão	Medida-F
R. Bayesiana	*Cobertura*	64,64 (9,35)	62,34 (8,93)	63,47
	Precisão	64,66 (9,36)	62,29 (8,89)	63,45
	Medida-F	64,57 (9,39)	62,29 (8,95)	63,41
R. Linear	*Cobertura*	**64,75** (9,56)	**62,40** (8,83)	**63,55**
	Precisão	64,59 (9,40)	62,24 (8,91)	63,40
	Medida-F	64,57 (9,39)	62,29 (8,95)	63,41
R. Ridge	*Cobertura*	64,73 (9,40)	62,36 (8,94)	63,52
	Precisão	63,67 (9,76)	61,29 (9,42)	62,46
	Medida-F	64,49 (9,51)	62,17 (9,07)	63,31
SVM-Linear	*Cobertura*	64,35 (9,40)	61,96 (8,99)	63,13
	Precisão	63,92 (9,70)	61,61 (9,28)	62,74
	Medida-F	64,34 (9,42)	61,95 (8,95)	63,12

Tabela 1. Resultados da avaliação (%) e desvio padrão (entre parênteses) dos algoritmos de regressão.

Em geral, os algoritmos apresentaram resultados similares, sendo que o algoritmo de Regressão linear usando a medida de cobertura do R-1 apresentou o melhor desempenho nas três medidas de avaliação. O resultado obtido na medida de cobertura demonstra que esse algoritmo conseguiu gerar resumos que possuem 64,75% das informações presentes nos resumos de referência. É importante ressaltar que o algoritmo de Regressão linear obteve 62,40% na medida de precisão e 63,55% na medida-F. Esses resultados indicam que os resumos gerados por esse algoritmo apresentam um bom equilíbrio entre a quantidade informações presentes nos resumos de referência e aquelas que o algoritmo selecionou.

Em geral, adotar a medida de cobertura do R-1 como escore de informatividade dos resumos foi a que levou os algoritmos de regressão a selecionarem melhor os resumos informativos. Tal resultado indica que essa medida é mais adequada para refletir informatividade dos resumos.

4.2 COMPARAÇÃO COM OUTROS TRABALHOS

Neste segundo experimento é realizado uma comparação dos resultados obtidos pelo sistema desenvolvido usando o algoritmo de Regressão linear com outros cinco sistemas identificados na literatura, sendo eles: GistSumm (PARDO, 2005), CSTSumm (CASTRO JORGE e PARDO, 2010), RC-4 (CARDOSO e PARDO, 2016), SentReg (SODRÉ e OLIVEIRA, 2019) e ILP (GOMES e OLIVEIRA, 2019). Além desses sistemas, identificamos o sistema de Oráculo que representa o limite máximo que pode ser obtido pelo sistema proposto neste trabalho, selecionando sempre dentre os cinco resumos candidatos de cada grupo de documentos, aquele que possui o maior valor real da medida de cobertura do R-1.

Na Tabela 2 são apresentados os resultados obtidos usando as medidas de cobertura, precisão e medida-F do R-1. O melhor sistema (com exceção do Oráculo) em cada medida de avaliação é destacado em negrito.

Sistemas	**Cobertura**	**Precisão**	**Medida-F**
CSTSumm	56,21 (11,34)	56,40 (10,17)	56,31
GistSumm	56,23 (12,34)	54,87 (10,76)	55,54
RC-4	57,66 (8,44)	58,33 (10,61)	57,99
SentReg	62,23 (9,28)	59,43 (8,08)	60,80
ILP	63,08 (8,56)	59,23 (7,66)	61,10
RegILP	**64,75** (9,56)	**62,40** (8,83)	**63,55**
Oracle	*66,30 (8,94)*	*63,79 (8,15)*	*65,02*

Tabela 2. Resultados da comparação (%) e desvio padrão (entre parênteses) com outros trabalhos da literatura.

O sistema desenvolvido neste trabalho obteve desempenho superior aos demais sistemas considerados em todas as medidas de avaliação do R-1. O sistema SentReg apresenta o segundo melhor resultado na medida de precisão, enquanto que o sistema ILP alcançou o segundo melhor desempenho nas medidas de cobertura e medida-F. É importante ressaltar que o sistema ILP utiliza uma abordagem similar ao sistema desenvolvido neste trabalho, mas sem a geração de múltiplos resumos candidatos e o módulo de seleção usando regressão. Dessa forma, é possível concluir que a estratégia de sumarização baseada em duas etapas adotada neste trabalho é viável e resultou na geração de resumos mais informativos do que os outros sistemas avaliados.

Apesar dos resultados encorajadores, ainda é existe muito espaço para melhorias, como pode ser observado comparando o resultado do sistema proposto com o sistema de Oráculo identificado. Tais resultados evidenciam que a etapa de geração dos resumos candidatos foi capaz de gerar resumos informativos para os documentos de entrada, mas esses não foram identificados na etapa de seleção. Tal problema ocorreu devido a erros de estimação cometido pelo algoritmo de regressão na seleção do resumo mais informativo.

5 | CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho foi apresentado um sistema de sumarização automático de artigos de notícias escritas em português. A solução proposta é baseada em uma estratégia de sumarização em duas etapas. Em uma primeira etapa, uma abordagem baseada em conceitos usando programação linear inteira é aplicada para gerar um conjunto de resumos candidatos. Posteriormente, um modelo de regressão foi construído para estimar a informatividade dos resumos candidatos visando selecionar o resumo mais informativo. Para isso, dez atributos baseados em medidas de similaridade, divergência e tradicionais indicadores de relevância foram utilizados.

Para avaliar o desempenho do sistema desenvolvido, experimentos executados usando o corpus CSTNews (CARDOSO et al., 2011) e as medidas de avaliação do ROUGE. Os resultados experimentais demonstraram que o algoritmo de Regressão linear obteve o melhor desempenho em comparação com outros algoritmos para a tarefa de estimar um escore de informatividade dos resumos candidatos. Além disso, o sistema proposto obteve resultados superiores em comparação com outros trabalhos da literatura.

REFERÊNCIAS

CARDOSO, P. C.; PARDO, T. A. S. **Multi-document summarization using semantic discourse models**. Procesamiento del Lenguaje Natural, (56): p. 57–64, 2016.

CASTRO JORGE, M. L. d. R.; PARDO, T. A. S. **Experiments with cst-based multidocument summarization**. In: Workshop on Graph-based Methods for Natural Language Processing, Stroudsburg, PA, USA, p. 74–82. 2010.

CARDOSO, P.; MAZIERO, E.; JORGE, M.; SENO, E.; FELIPPO, A. D.; RINO, NUNES, L. M.; PARDO, T. A. S. **CstNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese**. In: 3ª RST Brazilian Meeting. p. 88–105. 2011.

GOMES, Laerth; OLIVEIRA, Hilário. **A Multi-document Summarization System for News Articles in Portuguese using Integer Linear Programming**. In: Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2019, Salvador. Anais do XVI ENIAC. Porto Alegre: Sociedade Brasileira de Computação, p. 622-633. 2019.

KULLBACK, S. **Information theory and statistic**. Wiley, 1959.

LIN, C.-Y. **Rouge**: A package for automatic evaluation of summaries. In: Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain. Association for Computational Linguistics, 2004. p. 74–81.

LIN, J. **Divergence measures based on the shannon entropy**. IEEE Trans. Inf. Theor., IEEE Press, Piscataway, NJ, USA, v. 37, n. 1, p. 145–151, 2006.

MANI, I. **Summarization evaluation: An overview**. In: Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2, Tokyo, Japan, 2001.

NENKOVA, A.; McKeown, K. **A survey of text summarization techniques**. In Mining Text Data. Springer. p. 43-76. 2012.

OLIVEIRA, H.; LINS, R.D.; LIMA, R.; FREITAS, F. SIMSKE, S. J. **A regression-based approach using integer linear programming for single-document summarization**. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI). Boston, USA. p. 270-277. 2017.

OLIVEIRA, H.; LINS, R. D.; LIMA, R.; FREITAS, F.; SIMSKE, S. J. **A concept-based ILP approach for multi-document summarization exploring centrality and position**. In: 7th Brazilian Conference on Intelligent Systems (BRACIS). São Paulo, Brazil. p. 37–42, 2018.

PARDO, T. A. S. **Gistsumm-gist summarizer**: Extensões e novas funcionalidades. Serie de Relatórios do NILC. 2005.

SHIRKHORSHIDI, A. S.; AGHABOZORGI, S.; WAH, T. Y. **A comparison study on similarity and dissimilarity measures in clustering continuous data.** PLOS ONE, Public Library of Science, v. 10, n. 12, p. 1–20, 2015.

SODRÉ, Lucas; OLIVEIRA, Hilário. **Avaliando Algoritmos de Regressão para Sumarização Automática de Textos em Português do Brasil**. In: Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). Anais do XVI ENIAC. Porto Alegre: Sociedade Brasileira de Computação, p. 634-645. 2019.

ÍNDICE REMISSIVO

T

V

W

X

TECNOLOGIAS, MÉTODOS E TEORIAS NA ENGENHARIA DE COMPUTAÇÃO

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2020

TECNOLOGIAS, MÉTODOS E TEORIAS NA ENGENHARIA DE COMPUTAÇÃO

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2020