relato de pesquisa

UM MÉTODO DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS ATRAVÉS DE DADOS ESTATÍSTICOS E PROCESSAMENTO DE LINGUAGEM NATURAL

Osvaldo de Souza[*]
Hamilton Rodrigues Tabosa[**]
Davi Martins de Oliveira[***]
Mayra Helena de Souza Oliveira[****]

RESUMO Este artigo discute a mediação da informação em relação à sumarização automática de textos, examina técnicas de processamento de linguagem natural (PLN), e analisa o uso de técnicas de processamento de texto baseadas em métodos estatísticos de ocorrência de palavras do português brasileiro. Contextualiza o termo *sumarização* à Ciência da Informação. Propõe e apresenta um método de produção automática de sumários de textos baseado em técnicas de PLN e métodos estatísticos de uso de palavras. Para cada uma dessas técnicas, analisa e exemplifica, e oportunamente, apresenta as equações que governam o uso de tais técnicas. Como resultados obtidos na pesquisa, destaca-se um inédito *corpus* anotado, composto por aproximadamente meio milhão de palavras do português brasileiro, além dos resultados médios obtidos com os testes empíricos da ferramenta de sumarização, que indicam uma redução da dimensionalidade, para textos com até 500 palavras, da ordem de 53%. A análise geral dos achados da pesquisa indica que os resultados são promissores quanto à capacidade de redução e a preservação do valor semântico dos textos.

Palavras-chave: Sumarização automática de textos. Acessibilidade Informacional. Processamento de Linguagem natural. Mediação da Informação.

* Doutor em Engenharia de Teleinformática pela Universidade Federal do Ceará, Brasil. Professor do Departamento de Ciências da Informação da Universidade Federal do Ceará, Brasil. Coordenador do Grupo de Pesquisa CNPq Aplicações em Tecnologias Assistivas e Usabilidade.
E-mail: osvsouza@gmail.com.

** Doutor em Ciência da Informação pela Universidade Federal da Paraíba, Brasil. Professor do Departamento de Ciências da Informação da Universidade Federal do Ceará, Brasil.
E-mail: hrtabosa@gmail.com.

*** Graduando no Bacharelado em Biblioteconomia pela Universidade Federal do Ceará, Brasil. Trabalha na Universidade Federal do Ceará, Brasil.
E-mal: cartmandreamer@gmail.com.

**** Graduanda em Biblioteconomia pela Universidade Federal do Ceará, Brasil.
E-mail: helena.azuos@gmail.com.

1 INTRODUÇÃO

Com exceção do texto acadêmico, é comum a ausência de um resumo e de palavras-chave como elementos pré-textuais a auxiliarem o leitor em seu primeiro contato com um texto na Web, ambiente onde é comum haver uma grande quantidade de textos com conteúdos semelhantes. Nesse contexto, uma forma de contribuir para uma melhoria da representação, mediação, recuperação e uso da informação seria prover as páginas Web de um importante metadado: o resumo, o que permitiria um primeiro contato do leitor com as páginas Web muito mais produtivo, pois a partir dele poderíamos decidir ler o documento na íntegra ou não.

Nesse contexto, o resumo é efetivamente um elemento de acessibilidade: um caso particular de acessibilidade informacional.

Considerando-se a heterogeneidade da audiência da Ciência da Informação (CI) (MORAES; CARELLI, 2016), torna-se necessário evidenciar que, neste artigo, os termos

sumarização, condensação e resumo de textos são tratados como equivalentes, significando a exposição abreviada, precisa e sucinta do assunto ou matéria de um determinado documento, compreendendo as ideias principais do texto.

Justifica-se esse posicionamento devido à natureza interdisciplinar do tema, considerando o posicionamento de autores como Hutchins (1987) citado por Martins et al. (2001), que classifica sumários científicos em três tipos (indicativos, informativos e sumários de crítica), referindo-se exatamente às mesmas ideias definidas pela Associação Brasileira de Normas Técnicas (ABNT) (2003) como tipologias de resumo: indicativo, informativo e crítico. Pardo, Rino e Nunes (2003) também utilizam o termo sumarização e resumo para se referirem à mesma ideia. Considerou-se também o fato de que os trabalhos na área de Processamento de Linguagem Natural (PLN) usam exaustivamente os termos resumo e sumário como sinônimos.

A questão que inevitavelmente se coloca é: como gerar sumários, em tempo hábil, para tão grande massa de textos? A partir dessa questão, justificou-se empreender todo o trabalho relativo à pesquisa apresentada neste artigo, inclusive os resultados parciais alcançados.

Assim, pretende-se criar uma solução tecnológica capaz de produzir, sem intervenção humana e automaticamente, sumários de textos com considerável nível de redução da dimensionalidade dos documentos, sem perda semântica significativa e dotado de correção gramatical, utilizando recursos do PLN e métodos estatísticos baseados na ocorrência e uso das palavras no português brasileiro.

Para alcançar a construção de tal solução tecnológica, empreendeu-se a pesquisa visando aos seguintes objetivos:

a) o desenvolvimento de um protótipo para a captação automática de textos na Web;
b) a construção de um *corpus* anotado, de textos coletados na Web, de tema geral, e de um subtema específico relacionado à "energia limpa";
c) a construção de um *corpus* anotado de palavras do português brasileiro;
d) o desenvolvimento de um protótipo para a geração automática de sumários de textos.

2 BASES CONCEITUAIS DA SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

No processo de sumarização automática de textos é necessária uma etapa semelhante ao processo de indexação. Quanto à indexação automática, o grande desafio é, conforme Lancaster (2004), a extração de termos representativos do conteúdo dos documentos.

Borges (2009) afirma que os sistemas baseados em indexação por extração automática realizam, basicamente, as seguintes tarefas: 1) contar palavras num texto; 2) cotejá-las com uma lista de palavras proibidas; 3) eliminar palavras não significativas (artigos, preposições, conjunções, etc.) e 4) ordenar as palavras de acordo com sua frequência. O autor adverte que esse tipo de indexação, por ser baseado unicamente em critérios estatísticos, apresenta limitações.

Semelhante a esse processo, porém com uma preocupação quanto aos aspectos semânticos do texto dos documentos indexados, na indexação por atribuição automática é possível agregar outros conceitos aos termos (a partir da adoção de um instrumento de controle terminológico), ampliando a capacidade de representação temática do conteúdo do documento e agregando novo valor à indexação automática feita em primeira instância.

As técnicas de indexação automática por extração e por atribuição podem ser combinadas para a produção dos sumários dos textos. Tais sumários podem ser produzidos pela simples identificação e extração inalterada, das partes relevantes do texto, as quais passariam a compor um texto de menor dimensionalidade, sob isso nos fala Sparck Jones, (1993) que denominou essa categoria de resumo de "extratos", acrescentando ainda um segundo tipo, denominado por ele como "*abstract*" o qual por sua vez seria construído a partir de partes, ou mesmo do texto completo, reescrito, havendo, portanto uma modificação de partes do texto para a composição do resumo.

Um desafio que geralmente os sumarizadores automáticos têm de enfrentar, nem sempre com muito sucesso, é o fato de o texto resumido ficar tão truncado que dificulta uma leitura linear e coerente, como se as frases

não fizessem tanto sentido quando reunidas. A solução proposta e desenvolvida na pesquisa e relatada neste artigo procura diminuir a dificuldade supracitada.

O método desenvolvido é direcionado para a detecção do tema mais relevante do texto, e a partir da frase que representa esse tema relevante, construir o sumário com ele e os demais parágrafos que estejam associadas a essa ideia principal. Nessa abordagem, o texto produzido quase sempre é tão legível quanto o original, conforme demonstram os resultados apresentados mais adiante.

As pesquisas no campo do PLN têm procurado soluções para tais dificuldades de manipulação da linguagem, com vistas não só à sumarização de textos, mas também à tradução e à busca de informações em textos, por exemplo.

Pereira (2011, p. 2) afirma que o PLN, embora envolva diversas áreas do conhecimento, "consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em uma língua natural". Seguindo esse entendimento, conforme Gonzalez e Lima (2003, p. 3), o PLN "trata computacionalmente os diversos aspectos da comunicação humana, tais como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos".

Compreendemos que a produção de um sumário de um documento necessariamente prescinde de algum nível de compreensão do texto. Isto significa que é preciso detectar os maiores valores semânticos nele contidos.

Alinhado a essa compreensão, no escopo da pesquisa realizada, trabalhou-se buscando capacitar a ferramenta tecnológica na identificação do conteúdo semântico do texto, lançando mão de elementos do PLN e da estatística.

Portanto, para a consecução dos objetivos de nossa pesquisa, concordamos com Pereira (2011, p. 3), com relação aos seguintes aspectos do PLN, que são de interesse para o desenvolvimento de um sumarizador de textos:

> Morfologia: reconhece uma palavra em termos de unidades básicas (morfemas).
>
> Sintaxe: define a estrutura de uma frase com base na forma como as palavras dessa frase se relacionam entre si (categorias gramaticais).
>
> Semântica: associa significado às estruturas sintáticas, em função do significado das palavras que a compõem.
>
> Pragmática: adequa o significado de uma frase ao contexto em que ela é usada. (PEREIRA, 2011, p.3)

Além disso, o PLN abrange também outros temas sobre os quais desenvolvem-se estudos e pesquisas, tais como: o processamento morfossintático e semântico de sentenças, as representações de variações linguísticas e ambiguidades, a etiquetagem de texto, a eliminação de palavras não funcionais, como artigos, conetivos e preposições sem prejuízo da coerência textual, a representação do conhecimento e a recuperação de informação, entre outros.

No processamento dos textos, utilizando-se abordagens baseadas em PLN, geralmente são procedidas as seguintes ações básicas: a) normalização da grafia; b) redução da palavra à forma raiz da expressão escrita.

A normalização da grafia ocorre pela reescrita da palavra com a adoção de letras apenas maiúsculas ou apenas minúsculas. Na presente pesquisa adotou-se a escrita em letras maiúsculas. A redução da palavra à forma raiz reescreve as palavras removendo-se acentuação, flexões verbais, etc. Este processo é normalmente denominado de *Stemming*, todavia, o mesmo não foi adotado na pesquisa.

Após as ações básicas, deve ser aplicada alguma técnica para a classificação das palavras quanto a sua relevância nos textos. Uma dessas técnicas é baseada no *Text Frequency-Inverse Document Frequency*) (TF-IDF). O TF–IDF é caracterizado como um metadado avaliativo, de natureza estatística, que avalia o quanto um determinado termo, ou palavra, é importante em um documento específico pertencente a um determinado *corpus* de textos correlacionados. A primeira menção à avaliação de relevância de um termo em um texto é atribuída a Luhn (1957) sem que houvesse, contudo, uma postulação matemática do cálculo dessa avaliação. Uma formulação matemática foi proposta por Salton e Buckley (1988) que prevê o cálculo da frequência do termo, da frequência inversa no documento

e, da composição entre a frequência do termo e a frequência inversa no documento, conforme as equações 1, 2 e 3.

$$TF = \frac{t}{tp} \quad (1)$$

Na qual *t* é a quantidade de vezes que um termo ocorreu no texto, e *tp* é o total de termos existente no documento. O cálculo da frequência inversa no documento foi proposto conforme a equação 2.

$$IDF = log\frac{N}{n} \quad (2)$$

Na equação 2 *N* representa o total de documentos no *corpus*, e *n* é o total de vezes em que um determinado termo ocorreu em algum documento desse *corpus*. Por fim o cálculo do *TF-IDF* é obtido pela equação (3), que em palavras nos diz que o *TF-IDF* é a frequência de um determinado termo em um documento, multiplicado pela frequência inversa desse termo no corpus documental considerado.

$$TF\text{-}IDF = TF * IDF \quad (3)$$

A avaliação da importância de um termo no contexto de um determinado texto pode ser inferida a partir da frequência de ocorrência desse termo no texto. Termos que ocorrem com maior frequência em geral não representam bem o documento, de fato, aumentam o nível de ruído nas respostas em um sistema de recuperação da informação, concordamos com Sparck Jones (1972) que nos diz:

> As very frequently occurring terms are responsible for noise in retrieval, one possible course is simply to remove them from requests. The fact that this will reduce the number of terms available for conjoined matching may be offset by the fact that fewer non-relevant documents will be retrieved. (JONES, 1972, p.13)

Utilizando-se esse metadado reduz-se a dimensionalidade do texto, removendo-se dele as palavras e ou frases de menor relevância. O texto restante é considerado de maior relevância, e embora normalmente já seja menor do que o texto original, ainda pode ser reduzido. O RST ou CST são exemplos de técnicas que podem ser aplicadas à seleção final do texto para o sumário.

Com relação à abordagem de uso do PLN na pesquisa realizada, adotou-se uma abordagem híbrida cuja operacionalização encontra-se detalhada adiante.

3 PERCURSO METODOLÓGICO

A pesquisa teve início com um levantamento bibliográfico em periódicos nacionais e internacionais sobre os principais temas presentes neste estudo, a saber: processamento de linguagem natural, sumarização de textos, sumários automáticos de textos e indexação automática.

A partir do reconhecimento do estado da arte puderam-se lançar as bases para a parte empírica da pesquisa, estabelecendo-se categorias de análise e identificando-se as possibilidades de desenvolvimento de contribuições.

Propôs-se então, uma metodologia que combina técnicas baseadas no uso de PLN e métodos estatísticos objetivando a sumarização pela condensação semântica, pela eleição de termos relevantes e descarte dos menos relevantes.

A pesquisa caracteriza-se como experimental quanto aos procedimentos técnicos, e exploratória quanto aos objetivos, desenvolvendo-se em duas fases principais:

1) a construção de protótipos funcionais (*software*) baseados em linguagem Java SE (consiste em uma plataforma para o desenvolvimento de programas de computador baseados na linguagem de programação Java) e Java EE (refere-se à versão da Linguagem Java destinada a aplicações corporativas, compreendidas como aquelas que necessitam de um ambiente de execução, normalmente denominado de servidor de aplicações e um banco de dados relacional).
2) a construção de um *corpora* textual, que compreende:

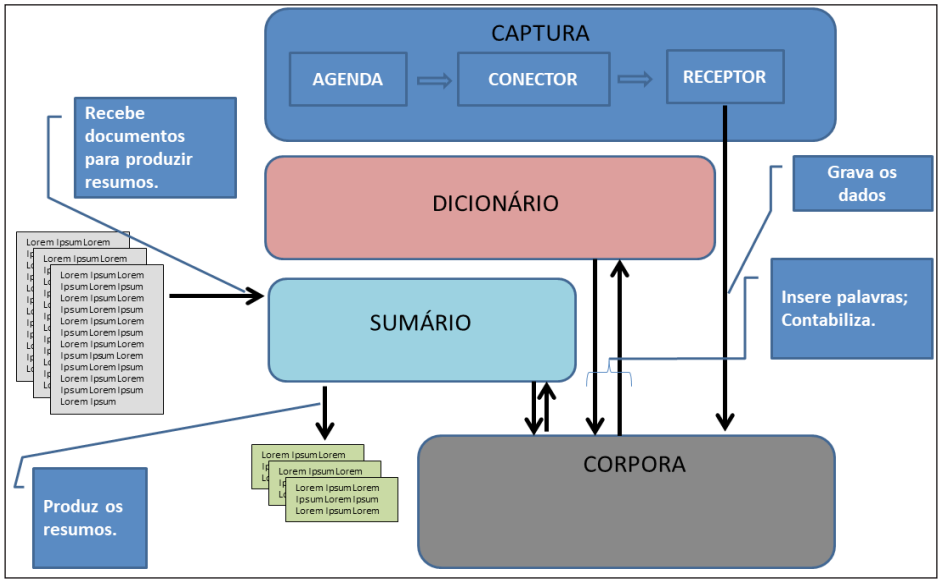
a) um pequeno *corpus* de textos anotados e pertencentes ao domínio "informação sobre energia limpa";

b) um *corpus* abrangente de textos de âmbito geral; e
c) um banco de dados com o cadastro de palavras do português brasileiro.

Quanto aos protótipos funcionais, eles foram construídos de forma modularizada. Cada protótipo foi elaborado como um módulo visando-se diminuir a complexidade de construção da solução de software geral. A Figura 1 apresenta a arquitetura de software utilizada no desenvolvimento do mesmo, a qual retrata também o fluxo de integração entre os módulos.

Pode-se perceber que o módulo CAPTURA é responsável por estabelecer conexão (conector) com as fontes de informação na web para obter os documentos. Ele provê uma funcionalidade (agenda) que suporta o cadastro das fontes a serem usadas e a periodicidade na qual o módulo deve conectar-se à fonte para obter mais documentos. Essa funcionalidade foi bastante útil na obtenção de versões atualizadas de jornais, os quais em geral produzem uma versão nova no começo do dia e ao longo do mesmo vão produzindo pequenas alterações com a inclusão de novas "manchetes".

Figura 1 - O fluxo de integração entre os módulos da solução de software



Fonte: dados da pesquisa

Durante a obtenção dos dados (receptor) nas fontes de informação da web, o CAPTURA faz um processamento inicial no documento, eliminando elementos que não são textuais, ou que representam marcações de layout ou ajustes visuais, normalmente em HTML, bem como elementos que estejam envolvidos na estruturação dos documentos, normalmente em XML. O resultado desse processamento inicial é um texto "limpo", sem anotações de formato de letra (itálico, negrito, etc.).

Após o módulo CAPTURA disponibilizar um novo documento, o módulo DICIONÁRIO começa a processá-lo. Neste processamento as palavras existentes no texto são inseridas uma-a-uma no banco de dados de palavra. Caso uma palavra já conste no banco, ela não é inserida novamente. Quando uma palavra é inserida no banco o "contador de ocorrências" dessa palavra é iniciado em zero. Cada vez que o DICIONÁRIO for inserir uma palavra e a mesma já se encontre inserida, o "contador de ocorrências" é incrementado em 1. O DICIONÁRIO também realiza outra contagem, a qual se refere ao total de palavras que já foram processadas e a cada

palavra processada o contador "total de palavras processadas" é incrementado em 1. Esses dois contadores são então utilizados para o cálculo posterior da frequência de cada palavra no conjunto de todos os textos já processados pelo módulo. Após o documento ter sido totalmente processado, atualiza as frequências TF de ocorrência das palavras nos textos e também à média geométrica das frequências inversas nos documentos, o IDF.

O módulo SUMÁRIO realiza a complexa operação de identificação dos valores semânticos das frases do texto, seleciona os mais relevantes e produz o resumo final utilizando-se de valores estatísticos de combinações das palavras. Esse processo é descrito em detalhes na seção "A Sumarização automática de textos".

Além dos módulos tem-se também o desenvolvimento dos experimentos, cujos resultados são relatados neste artigo. Ambas as fases da pesquisa são detalhadas a seguir.

3.1 A primeira fase da pesquisa

Na primeira fase obtiveram-se as seguintes soluções tecnológicas, caracterizadas como protótipos funcionais de *software*:

1. módulo para a captação de textos na Web, denominada de CAPTURA;
2. módulo de cadastro automático de palavras do português brasileiro, denominado de DICIONÁRIO;
3. módulo para a geração automática de sumário de texto, denominado de SUMÁRIO.

Com a utilização do módulo CAPTURA foi possível selecionar e incluir no *corpus* de textos, 100 (cem) textos pertencentes ao domínio "informação sobre energia limpa". A seleção dos textos sobre esse domínio, a partir de páginas Web nacionais, foi realizada manualmente pela equipe do projeto utilizando-se o módulo. Na seleção direcionou-se a busca a documentos majoritariamente textuais, excluindo-se páginas cujo conteúdo estivesse vinculado a imagens, áudios ou vídeos, uma vez que o *software* sumariza apenas texto escrito em português. O desenvolvimento dessas atividades foi realizado no segundo semestre de 2015.

Esse *corpus* de 100 (cem) documentos foi analisado visando à identificação das principais estruturas semânticas presentes no texto, parágrafo por parágrafo, frase por frase. A anotação realizada sobre esse *corpus* consistiu no registro da quantidade de palavras geral, da quantidade de palavras por frases, da quantidade de frases e da quantidade de frases relevantes.

Utilizando-se o mesmo módulo CAPTURA, foi compilado o segundo *corpus*, para o qual foram obtidos 130.000 (cento e trinta mil) documentos relativos a temas diversos e publicados em jornais e revistas nacionais.

Ambos os *corpus* totalizam um *corpora* de 130.100 (cento e trinta mil e cem) documentos. Esse volume de documentos foi utilizado na construção do banco de dados com o cadastro de palavras da língua portuguesa. Os 130.100 (cento e trinta mil e cem) documentos resultaram em cerca de 1.300.000 (um milhão e trezentas mil) páginas de textos e produziram um cadastro de cerca de 500.000 (quinhentas mil) palavras.

3.2 A segunda fase da pesquisa

A segunda fase da pesquisa teve início a partir do processamento das 1.300.000 (um milhão e trezentas mil) páginas de textos obtidos na primeira fase, as quais foram processadas pelo módulo DICIONÁRIO, que realiza os seguintes processamentos:

1) cadastra uma única vez cada palavra existente nos textos;
2) para cada texto, calcula a frequência de ocorrência das palavras no texto (TF) conforme equação 1;
3) para cada palavra, calcula a média geométrica das frequências inversas nos documentos do *corpus* (IDF) conforme a equação 2;
4) para cada palavra, calcula o TF–IDF conforme a equação 3;
5) para cada palavra, calcula a frequência das palavras que ocorrem a sua esquerda e a sua direita.

Os resultados desse processamento são utilizados na identificação estatística da relevância das frases e também no passo final da construção automática dos sumários, cujos processos envolvidos são apresentados nas seções seguintes.

3.3 A sumarização automática de textos

A sumarização automática de textos objetivada na pesquisa consiste, resumidamente, em uma proposta que envolve a seguinte sucessão de processos: redução de dimensionalidade e condensação semântica. Esses processos são executados pelo módulo SUMÁRIO e foram distribuídos nos seguintes passos: identificação das palavras e frases relevantes no texto, eliminação das frases menos relevantes, arranjo das frases restantes em frases menores e aplicação de uma correção gramatical estatística.

O resultado esperado é um documento significativamente menor em termos de quantidade de palavras, mas com a menor perda possível de conteúdo semântico.

Considerando-se que, idealmente, no sumário espera-se encontrar o mesmo valor semântico do documento integral, ou pelo menos o conteúdo semântico mais relevante, compreende-se que o processo adotado faz uma condensação semântica.

O Quadro 1 apresenta a sucessão de processos e passos evolutivos envolvidos. Após a conclusão do passo evolutivo (a) as frases restantes são classificadas segundo sua relevância e em (b) as frases de baixa relevância são descartadas. Por fim, as frases restantes são processadas em (c), utilizando-se métodos estatísticos. Nesse ponto, um novo texto é reconstruído, o qual passa a ser o sumário criado automaticamente para o texto fornecido.

Quadro 1 - Organização dos Processos

Processo	Passo Evolutivos
1) Redução de Dimensionalidade	(a) Identificar as palavras e frases relevantes
	(b) Eliminar as frases menos relevantes
2) Condensação Semântica	(c) Reescrever as frases restantes em frases menores e aplicar uma correção gramatical estatística

Fonte: dados da pesquisa

Na seção seguinte detalham-se todos os processos e passos evolutivos organizados no Quadro 1.

4 UMA NOVA METODOLOGIA DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

O objetivo desta seção é estabelecer os desdobramentos teóricos pertinentes e resultantes da abordagem combinada entre PLN e os métodos estatísticos utilizados. Objetiva também detalhar e exemplificar a aplicação da teoria no método desenvolvido na pesquisa e apresentado neste artigo. O estabelecimento teórico consiste na formulação matemática pertinente, e na explanação desta formulação. A exemplificação é incluída a fim de facilitar a apreensão do comportamento dos protótipos quanto ao processamento do texto em linguagem natural existente nos documentos.

4.1 A redução da dimensionalidade

A redução da dimensionalidade é alcançada usando-se uma metodologia híbrida baseada em PLN e estatística. A abordagem de PLN, aplicado ao problema da pesquisa, envolve a remoção de palavras semanticamente menos relevantes no conjunto de documentos. Para a identificação dessas palavras são utilizadas duas estratégias:

a) uso de um *corpus* composto de um conjunto de palavras descartáveis (PD) (s*top words)* notadamente apenas estruturais, como por exemplo: a, e, ou, mas, nem, um. No experimento realizado, o conjunto PD utilizado é composto de 258 palavras descartáveis;

b) remoção das palavras de baixa relevância no português brasileiro, utilizando-se como base as estatísticas computadas para o TF existentes no *corpus* produzido na primeira fase do projeto.

Como preparação para a execução das estratégias "a" e "b", o texto é normalizado convertendo-se todo o texto para letras minúsculas. Essa preparação tem por finalidade inibir a ocorrência de processamento indevido

devido a uma mesma palavra possuir grafia diferente.

Após a redução de dimensionalidade, o texto em geral apresenta-se descaracterizado de suas propriedades de legibilidade, como se vê no exemplo contido nos quadros 3, 4 e 5.

A fim de facilitar a compreensão propõe-se um exemplo utilizando-se o texto abaixo:

> A busca por fontes de energia limpa e renováveis é um dos grandes desafios da população mundial, o que faz com que pesquisadores procurem por soluções cada vez mais incomuns e inovadoras.

Pesquisadores da Universidade Brigham Young, de Washington (EUA), criaram uma célula de combustível que retira energia elétrica a partir da glicose e de outros açúcares, também conhecidos como carboidratos.

> Isso mesmo, a fonte de energia preferida do corpo humano pode, em um futuro próximo, alimentar desde o celular até um carro." (CL, 2015).

Observe-se que o texto em questão possui 3 (três) parágrafos (evidenciados intencionalmente por uma linha) e 83 palavras.

4.1.1 Identificar as palavras e frases relevantes

Para identificarem-se as palavras menos relevantes utiliza-se a TF disponível no *corpus*, esse valor, para a primeira frase do texto exemplo pode ser visto no Quadro 2.

Quadro 2 - Relação das palavras, da primeira frase do texto exemplo, com suas frequências

Palavra	TF	Palavra	TF
A *	0.0337239	mundial	0.0001647
busca	0.0002214	o *	0.0323066
por *	0.0053773	que *	0.0212470
fontes	0.0000581	faz	0.0004067
de *	0.0075544	com *	0.0103582
energia	0.0000985	que *	0.0212470
limpa	0.0000211	pesquisadores	0.0000625
e *	0.0263629	procurem	0.0000037
renováveis	0.0000131	por *	0.0053773
é *	0.0078048	soluções	0.0000410
um *	0.0086046	cada *	0.0006471
dos *	0.0048008	vez *	0.0006762
grandes *	0.0002089	mais *	0.0006762
desafios	0.0000421	incomuns	0.0000012
da *	0.0168726	e *	0.0263629
população	0.0003439	inovadoras	0.0000059

Fonte: dados da pesquisa

No Quadro 2 as palavras marcadas com "*" são consideradas *stopwords* e são simplesmente eliminadas. No experimento foram também removidas as palavras com frequência superior a 0.001 (10^{-3}). No Quadro 3 podem-se ver as modificações do texto para cada frase a partir da aplicação parcial do passo evolutivo "a".

Quadro 3 - Exemplo de redução de dimensionalidade

#	Texto original	Texto sem *stopword*	Texto apenas com palavras relevantes
1	A busca por fontes de energia limpa e renováveis é um dos grandes desafios da população mundial, o que faz com que pesquisadores procurem por soluções cada vez mais incomuns e inovadoras.	busca fontes energia limpa renováveis desafio população mundial faz pesquisadores procurem soluções mais incomuns inovadoras	fontes energia limpa renováveis desafios pesquisadores procurem soluções incomuns inovadoras
2	Pesquisadores da Universidade Brigham Young, de Washington (EUA), criaram uma célula de combustível que retira energia elétrica a partir da glicose e de outros açúcares, também conhecidos como carboidratos.	p e s q u i s a d o r e s universidade brigham young, washington (eua) criaram célula combustível retira energia elétrica partir glicose outros açúcares conhecidos carboidratos	brigham young, washington (eua) criaram célula combustível retira energia elétrica glicose conhecidos carboidratos
3	Isso mesmo, a fonte de energia preferida do corpo humano pode, em um futuro próximo, alimentar desde o celular até um carro.	isso mesmo fonte energia preferida corpo humano futuro próximo alimentar desde celular carro	fonte energia preferida humano próximo alimentar celular

Fonte: dados da pesquisa

Após o texto sofrer a primeira transformação, com a remoção de elementos textuais de baixa relevância (*stopwords*) e de acordo com a TF computada para as palavras do português brasileiro presentes nas 1.300.000 (um milhão e trezentas mil) páginas de texto), obtém-se um texto instrumental, o qual é exemplificado no Quadro 3, "Texto apenas com palavras relevantes".

O processo continua estabelecendo-se a relevância das frases restantes. No caso do texto ilustrado no Quadro 3 teremos 3 frases a serem avaliadas.

Para a identificação das frases relevantes no texto específico, é necessário se calcular o grau de relevância de cada parágrafo. O parágrafo mais relevante é utilizado como "pivô" (ponto de partida) para a construção automática do sumário. A partir da frase pivô determinam-se quais demais frases do texto possuem conexão com a frase pivô. Frases com conexão com o pivô são descartadas.

O cálculo da relevância da frase envolve computar o total de frequência de cada frase, e normalizar esse valor pela quantidade de palavras existentes na frase.

O calculo da frequência acumulada normalizada do texto é realizado de acordo com a equação 4.

$$g(p) = \left(\sum_{i=1}^{i \leq tp} y + TFIDF_p \right) \Big/ tp \qquad (4)$$

Na qual $TFIDF_p$ representa o valor de *Text Frequency-Inverse Document Frequency*, calculado pela equação 3, para a palavra p e tp é o total de palavras existentes no texto considerado e y é inicializada em zero.

4.1.2 Eliminar as frases menos relevantes

Para eliminarmos as frases menos relevantes no texto, prescindimos de uma classificação estatística. Essa classificação foi obtida a partir da Equação 4. O restante do processo envolve fixar o processamento (pivotear) a partir da frase de menor frequência normalizada. Em seguida as demais frases são comparadas através de simples correlação entre suas palavras. Essa correlação é realizada conforme a Equação 5.

$$h(f_p, f_2) = \left(\sum_{i=1}^{i \le tp} y + \begin{cases} 1, se\ wp_i \in f2 \\ 0, se\ outro\ caso \end{cases} \right) / tp \qquad (5)$$

Na qual *fp* é a frase pivô e *f*2 é a frase sendo comparada, *tp* é o total de palavras existentes na frase *f*2, e wp_i representa a *i*-ésima palavra da frase pivo, e *y* é inicializada em zero.

Uma vez classificadas as frases procedemos ao descarte das frases com correlação com a frase pivô. Observe-se que a equação necessita de um limiar (*threshold*) para a tomada de decisão quanto ao descarte. Em nossos experimentos adotou-se o limiar *l*=0,2 correspondente a 20%, experimentos futuros considerarão uma busca exaustiva pelo melhor valor a ser adotado como limiar. Caso a frase possua uma correlação maior do que o valor de descarte ela não será incorporada ao sumário.

$$d(f) = \begin{cases} 1, se\ h(f_p, f) > l \\ 0, se\ outro\ caso \end{cases} \qquad (6)$$

Na Equação 3, refere-se a frase pivô e *f* representa a frase para qual se deseja decidir sobre o descarte ou a manutenção da frase no sumário do texto. O valor de *l* corresponde ao valor do limiar adotado.

Perceba-se que a Equação 6 apenas avalia o valor de *h*(*fp*,*f*) estabelecendo dois valores possíveis como resultado, sendo o valor 0 (zero) o indicativo de manutenção da frase. A frase será descartada se *d*(*f*) = 1, pois se considerou nos experimentos que, frases cujo conjunto de palavras possua acima de 20% de semelhança podem ser descartadas.

O quadros 4 e 5 apresentam os resultados das aplicações das equações 4, 5 e 6 nos textos.

Quadro 4 - Valoração semântica a partir da frequência da palavra

Texto resultante (quarta coluna Quadro 3)	**TF-IDF acumulada normalizada do texto**	**Grau de conexão com a frase pivô (0 até 1)**	**Classificação De Relevância**	**Decisão**
fontes limpa renováveis desafios pesquisadores procurem soluções incomuns inovadoras	0,000513	0,0	2ª	mantida
brigham young washington (eua) criaram célula combustível retira energia elétrica glicose conhecidos carboidratos	0,000330	- -	1ª	pivô
fonte energia preferida humano próximo alimentar celular	0,000615	0,14	3ª	mantida

Fonte: dados da pesquisa

Quadro 5 - Frases resultantes do processo de seleção por relevância

Texto resultante
fontes energia limpa renováveis desafios pesquisadores procurem soluções incomuns inovadoras
brigham young washington (eua) criaram célula combustível retira energia elétrica glicose conhecidos carboidratos
fonte energia preferida humano próximo alimentar celular

Fonte: dados da pesquisa

Ao final do processamento e avaliação de cada frase do texto, as frases restantes, conforme exemplo no Quadro 5, são as candidatas à composição do sumário do texto.

O próximo passo na construção do sumário consiste na correção gramatical estatística

4.1.3 Correção gramatical estatística

Uma correção gramatical estatística é utilizada para melhorar o sumário produzido pelo processamento anterior. Essa melhoria consiste em acrescentar artigos, preposições, etc. e demais elementos estruturais necessários ao conforto da leitura.

Para a escolha dos itens a serem acrescidos às frases do sumário, são utilizadas informações estatísticas presentes no *corpus* de palavras do português brasileiro construído na pesquisa. Estas informações estatísticas revelam as relações de co-ocorrência entre grupos de palavras, e a partir destas relações pode-se escolher uma adequada, ou parcialmente adequada. O uso dessas informações estatísticas ocorre de acordo com o operador estatístico definido pela Equação 7.

$$\arg max\,((x_i, y_i, z_i)|t(x, y, z) = \max t(x_i, y_i, z_i)) \quad (7)$$

Na qual é o trio de palavras com maior frequência de associação (max), esse operador estatístico é usado para identificar qual trio de palavra tem a maior correlação.

No Quadro 6 podem-se ver os dados estatísticos para algumas das palavras existentes nos textos (frases) apresentadas no Quadro 5.

Quadro 6 - Exemplos de dados do *corpus* de palavras

Palavras	Palavras de maior frequência de associação						
	Palavras à esquerda				Palavras à direita		
fontes	as	das	de		de	do	e
	0,9%	0,81%	0,64%		0,18%	0,6%	0,02%
energia	de	a	e		a	e	solar
	0,45%	0,09%	0,09		0,02	0,07	0,03

Fonte: Dados da Pesquisa.

Utilizando-se a Equação 4, para as palavras x="fontes", e z="energia", obteremos como resposta os índices que apontam para x="fontes", y="de" e z="energia", pois a direita de "fontes" a palavra de maior uso é "de", e a esquerda de "energia" a palavra mais usada é "de", portanto, na correção gramatical estatística usada, teremos "fontes de energia".

Após esse processo ser realizado para todas as frases do Quadro 5 e reagrupando-se as frases, obtém-se o texto apresentado no Quadro 7.

Quadro 7 - Sumário produzido automaticamente

Sumário produzido automaticamente
As fontes de energia limpa e renováveis são desafios a pesquisadores que procurem soluções incomuns inovadoras. brigham young washington (eua) criaram uma célula de combustível que retira energia elétrica da glicose conhecidos por carboidratos. a fonte de energia preferida do humano é próximo de alimentar um celular

Fonte: dados da pesquisa

O processo descrito e exemplificado nas seções anteriores foi aplicado ao corpus de 100 textos, cujos resultados são apresentados nas próximas seções.

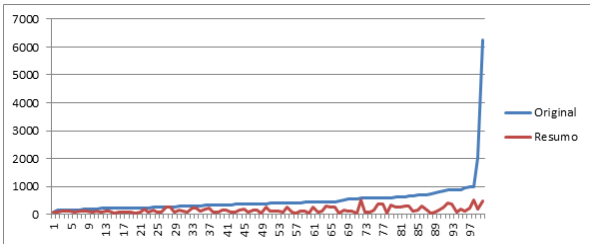
5 RESULTADOS DA EXPERIMENTAÇÃO

No sentido de validar a metodologia apresentada neste documento, foi conduzido um experimento no qual foram produzidos automaticamente sumários para os 100 textos que integram o *corpus* do tema "informação sobre energia limpa". Para esse experimento, obtiveram-se os seguintes resultados quanto à média de diminuição do tamanho do texto:

1) textos com até 500 palavras: 53% de diminuição;
2) textos entre 500 até 1000 palavras: 69% de diminuição;
3) textos entre 1000 até palavras 6000: 91% de diminuição.

A Figura 1 apresenta os resultados individuais de cada uma das produções automáticas de sumários.

Figura 1 - Desempenho na produção automática de sumários – 100 Textos

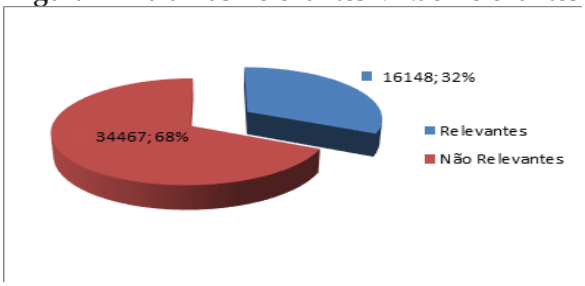


Fonte: dados da pesquisa

Pela Figura 1 pode-se observar que, quanto maior o texto melhor será o desempenho do sumarizador. Essa observação pode ser explicada considerando-se que quanto maior for o texto, menor será a densidade semântica, e, portanto, mais texto poderá ser descartando. A Figura 2 apresenta dados que permitem essa conclusão.

Perceba-se, pela Figura 1, que o tamanho do sumário apresenta forte variação entre os textos 69 até 97. Isso indica que, de fato, há uma grande variação na densidade semântica de um texto para outro, de um autor para outro. Obviamente, autores diferentes podem requerer mais ou menos palavras para expressarem a mesma ideia, e isto é decorrente dos vários fatores que influenciam a formação do acervo cognitivo de cada um de nós.

Figura 2 - Palavras Relevantes x Não Relevantes



Fonte: dados da pesquisa

Observa-se, pela Figura 2, que das 50.615 (cinquenta mil e seiscentas e quinze) palavras usadas nos documentos, 68% são palavras com função mais estrutural do que semântica. Obviamente nossa afirmação refere-se à decisão do que deve constar ou não no sumário.

Os resultados obtidos no experimento conduzido revelam que o uso de um *corpus*, no qual existam informações prévias sobre a distribuição estatística da ocorrência das palavras na expressão escrita, é relevante para o desenvolvimento de soluções de acessibilidade informacional. Quanto ao tempo de processamento, o experimento total, de processamento e produção dos sumários para o *corpus* de 100 textos, foi completado em 83 (oitenta e três) segundos. Portanto, quanto ao tempo de produção automática dos sumários não foi observado nenhum desempenho proibitivo, ao contrário, mesmo o maior dos textos foi processado sendo produzido automaticamente um sumário em poucos segundos. Interpretando-se esse dado, podemos afirmar que o protótipo criado a partir da metodologia proposta, tem capacidade de "ler" textos e "escrever" sumários a uma velocidade mínima de 804 (oitocentos e quatro) palavras por segundo. Obviamente esse desempenho pode ser melhorado adotando-se plataformas de equipamentos superiores às utilizadas no experimento.

Quanto às dificuldades enfrentadas e limitações percebidas, observou-se que quando o texto faz referência a elementos externos, como imagens, por exemplo, a sumarização fica prejudicada. Todavia o mesmo fato é observado quando o sumário é produzido manualmente. Esse é um problema a ser investigado em trabalhos futuros.

6 CONSIDERAÇÕES FINAIS

A redução da dimensionalidade dos textos na Web se torna um tema urgente de pesquisa e inovação, ao passo que cada vez mais se produz e se publica nesse ambiente, de modo que o volume de informação existente se torna inapreensível para o ser humano.

Dispor de uma ferramenta que reduza até 91% da dimensionalidade desses textos, oferecendo uma condensação semântica de informação é não apenas desejável, mas fundamental para pesquisadores e instituições, que estão cada vez mais correndo contra o tempo na busca por produtividade e resultados.

Ficou evidente que é imprescindível o uso de um *corpus* no idioma para o qual se deseja produzir sumários automaticamente, para viabilizar resultados com aplicação que possa ser usada no cotidiano. A metodologia proposta neste documento difere-se fundamentalmente das demais por contar com essa fonte de informações estatísticas. Difere-se também dos trabalhos anteriormente publicados porque neles o TF-IDF é produzido apenas para o conjunto específico para o qual se desejada indexar e ou produzir sumários. Em nossa abordagem, o TF-IDF foi produzido para todas as palavras existentes em um conjunto significativo de 1.300.000 (um milhão e trezentas mil) páginas de textos.

Na continuidade da pesquisa, cujos resultados obtidos até o momento estão relatados neste documento, pretende-se empreender uma avaliação qualitativa dos sumários produzidos. Serão feitos testes "cegos" nos quais voluntários farão a avaliação de dois sumários para um mesmo texto, um produzido pelo computador (com o método aqui relatado) e outro produzido pelo homem. Esperamos que esse teste possa apontar elementos sólidos relativos à avaliação da qualidade da sumarização automática produzida de acordo com a metodologia apresentada neste artigo.

A depender dos resultados do aludido teste, a metodologia será usada na produção de uma ferramenta visando à acessibilidade para deficientes visuais.

Artigo recebido em 25/01/2017 e aceito para publicação em 20/07/2017

A SUMMARIZATION METHOD AUTOMATIC TEXT THROUGH STATISTICAL DATA AND NATURAL LANGUAGE PROCESSING

ABSTRACT *This paper discusses the information mediation in the context of the automatic text summarization, it examines natural language processing techniques (NLP), and analyzes the use of techniques based on statistical methods for word processing of Brazilian Portuguese. It contextualizes the text summarization in the subject of Information Science. It proposes and explains a new method of automatic text summarization based on both NLP and statistical methods. For each of these techniques, it analyzes and exemplifies, and timely presents mathematical equations for such techniques. As results obtained in the research, we highlight an unpublished corpus annotated, composed of approximately half a million words of Brazilian Portuguese, in addition to the average results obtained with the empirical tests of the summarization tool, which indicate a reduction of dimensionality, for texts with up to 500 words, of the order of 53%. The general analysis of the research findings indicates that the results are promising in terms of the ability to reduce and preserve the semantic value of texts.*

Keywords: *Automatic Text Summarization. Information Accessibility. Natural Language Processing. Information mediation. Accessibility.*

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6028**: Informação e documentação - resumo - apresentação. Rio de Janeiro: ABNT, 2003.

BORGES, G. S. B. **Indexação automática de documentos textuais**: proposta de critérios essenciais. 2009. 111 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais, 2009.

MORAES, M.; CARELLI, A. E. A interdisciplinaridade na Ciência da Informação pela perspectiva da análise de citações. **Em Questão**, v. 22. n. 1. p. 137-160. 2016.

GONZALEZ, M.; LIMA, V. L. S. **Recuperação de informação e processamento da linguagem natural.** 2003. Disponível em: <http://www.inf.pucrs.br/~gonzalez/docs/minicurso-jaia2003.pdf>. Acesso em: 9 mar. 2016.

LANCASTER, F. W. **Indexação e sumários:** teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

LUHN, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. **IBM Journal of Research and Development**. N. 1. V. 4. p. 309-317. 1957.

MARTINS, C. B. *et al*. **Introdução à sumarização automática.** Relatório Técnico RT-DC 002/2001. Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP, fev. 2001. 38 f. Disponível em: <http://www.icmc.usp.br/~taspardo/RTDC00201-CMartinsEtAl.pdf>. Acesso em: 06 maio 2016.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. GistSumm: a summarization tool based on a new extractive method. In: WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PTUGUESE LANGUAGE, 6. Faro, 2003. **Proceedings…** Faro: 2003, Portugal. p. 210-218.

PEREIRA, S. L. **Processamento de Linguagem Natural.** 2011. Disponível em: <http://walderson.com/2011-2/IA/07-processamentolinguagemnatural.pdf>. Acesso em: 09 mar. 2016.

SPARCK JONES, K. What might be in a summary? In, **Information Retrieval.** Ed**.** Krause Knorz and Womser-Hacker. n. 93. p. 9-26. Universitatsverlag Konstanz. June. 1993.

SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**. v. 28. n. 1. p. 11-21. 1972.