DOI: 10.5433/1981-8920.2020v25n1p189

AVALIAÇÃO DO DESEMPENHO DE UM SOFTWARE DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

EVALUATION OF PERFORMANCE OF A SOFTWARE OF AUTOMATIC SUMARIZATION OF TEXTS

Hamilton Rodrigues Tabosa[a]
Osvaldo de Souza[b]
José Carlos dos Santos Cândido[c]
Ana Cristina Azevedo Ursulino Melo[d]
Keila Giullianna Braga Reis[e]

RESUMO
Introdução: Desde 2014 desenvolvemos uma pesquisa com o intuito de produzir um software (protótipo) que seria capaz de elaborar resumos automáticos de textos baseado em técnicas de Processamento de Linguagem Natural e estatísticas de frequência de palavras. Os primeiros testes da ferramenta geraram resultados que indicaram uma significativa redução da dimensionalidade dos textos, com considerável preservação do seu valor semântico. **Objetivo:** Neste artigo, apresentamos os resultados da continuidade do nosso trabalho investigativo, a partir de uma avaliação humana da qualidade desses resumos baseada na realização de testes cegos. **Metodologia:** Um grupo de três bibliotecárias recebeu um bloco misto e não identificado de resumos - produzidos por humanos e os resumos automáticos feitos pelo software - e procedeu a uma avaliação, segundo os critérios de corretude gramatical, preservação das ideias centrais, coerência e legibilidade, extensão do resumo, se houve paráfrase ou cópia de fragmentos e, se houve introdução de ideias não contidas no texto original. **Resultados:** Os resultados mostraram que em quatro, dos cinco critérios de avaliação adotados, houve equivalência qualitativa entre os resumos produzidos por humanos e os produzidos pelo software, o que parece representar um relativo sucesso, uma vez

[a] Doutor em Ciência da Informação pela Universidade Federal da Paraíba (UFPB). Professor do Departamento de Ciências da Informação da Universidade Federal do Ceará (UFC). E-mail: hrtabosa@gmail.com

[b] Doutor em Engenharia de Teleinformática pela Universidade Federal do Ceará (UFC). Professor do Departamento de Ciências da Informação da Univeridade Federal do Ceará (UFC). E-mail: osvsouza@gmail.com

[c] Graduando em Biblioteconomia pela Univeridade Federal do Ceará (UFC). E-mail: jose.z.candido@gmail.com

[d] Mestra em Avaliação de Políticas Públicas e Graduada em Biblioteconomia pela Universidade Federal do Ceará (UFC). E-mail: acrismelo@gmail.com

[e] Graduada em Biblioteconomia pela Universidade Federal do Ceará. E-mail: keillagbreis@gmail.com

que o protótipo poderia substituir uma pessoa na atividade de resumir textos sem deixar a desejar, a não ser no quinto critério de avaliação, referente à dimensão do resumo, em que o texto produzido pelo software foi apontado como extenso além do necessário. **Conclusões:** Apesar dos bons resultados do protótipo, percebemos a necessidade de melhorias em seu desempenho, além de avaliá-lo por métodos mais abrangentes, a partir de amostras mais representativas e por um grupo maior de avaliadores.

Descritores: Sumarização automática de textos. Acesso à informação. Processamento da linguagem natural. Mediação (Prática).

1 INTRODUÇÃO

Conforme apontam os resultados preliminares da pesquisa que resultou na criação do *software* de sumarização automática de textos, o protótipo apresentou resultados satisfatórios quanto à redução da dimensionalidade dos textos (na ordem de até 91% sem perda semântica significativa) e à velocidade de processamento e produção dos resumos (83 segundos), conforme relatam Souza *et al.* (2017).

No entanto, tais resultados técnicos não são suficientes para assegurarmos o valor holístico do protótipo, sem que analisemos se os resumos produzidos são dotados de qualidade textual compatível com uma leitura linear, confortável e inteligível.

Para avaliarmos a qualidade final dos resumos automáticos produzidos pelo protótipo, realizamos testes cegos com base em cinco critérios: corretude gramatical, coerência e legibilidade, introdução de elementos não contidos no texto original, preservação das ideias centrais, extensão do resumo, se houve paráfrase ou cópia de fragmentos.

Esses parâmetros de avaliação foram propostos pela equipe do projeto de pesquisa, a partir do seu entendimento e discussões sobre quais qualidades um resumo deve possuir para ser considerado razoavelmente bom e que podem ser alcançadas tanto por indexadores humanos quanto por máquinas. Dessa forma, acreditamos termos um conjunto de medidas justas para proceder à avaliação do *software*, sem bias e/ou tendenciosidades.

Para cada formulário de avaliação, em cada resumo, havia um campo facultativo para comentários dos avaliadores, onde eles poderiam escrever

considerações gerais sobre os itens, caso desejassem.

Assim, a partir do crivo de três bibliotecárias, que tiveram acesso a 20 textos completos e, para cada um, a três resumos não identificados quanto a sua autoria ou procedência, - tendo sido um produzido pelo *software* e dois elaborados por estudantes do Centro de Humanidades da Universidade Federal do Ceará, (majoritariamente) do Curso de Biblioteconomia-, foi possível qualificar os resumos e chegarmos a algumas conclusões, que apresentamos neste artigo.

2 BREVE INTRODUÇÃO ÀS BASES CONCEITUAIS DA SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

Assim, a partir do crivo de três bibliotecárias, que tiveram acesso a 20 textos completos e, para cada um, a três resumos não identificados quanto a sua autoria ou procedência, - tendo sido um produzido pelo *software* e dois elaborados por estudantes do Centro de Humanidades da Universidade Federal do Ceará, (majoritariamente) do Curso de Biblioteconomia-, foi possível qualificar os resumos e chegarmos a algumas conclusões, que apresentamos neste artigo.

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 de resumos de textos. Tais resumos podem ser produzidos pela simples identificação e extração inalterada das partes relevantes do texto, as quais passariam a compor um produto de menor dimensionalidade. Sparck-Jones (1993) denomina essa categoria de resumo como "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.

Conforme Cabral (2015), as primeiras pesquisas nessa área surgiram em 1958 a partir dos estudos de Luhn, que se propôs a analisar as palavras de um texto estatisticamente com o intuito de medir a relevância das sentenças para criação de resumos.

Outras pesquisas desenvolveram-se desde então, cada uma destacando uma metodologia diferente para a determinação de como avaliar e combinar os trechos mais representativos da essência semântica dos resumos, como por exemplo: as pesquisas de Erkan e Radev (2004), que se propuseram a representar as correlações entre as sentenças por meio de um modelo de grafos, trabalho assemelhado ao apresentado posteriormente por Takamura e Okumura (2009), que avaliaram sentenças de acordo com modelos baseados em *cluster* ou grafos; Wang e Li (2010) que estudaram um algoritmo de agrupamento hierárquico com a intenção de identificar conjuntos de sentenças de conteúdo semelhante e atualizar os resumos ao longo do tempo; e também trabalhando com algoritmos, temos o estudo de Brin e Page (1998), que visa determinar as frases mais relevantes de acordo com a centralidade do autovetor obtida por meio do algoritmo *PageRank*. Também Hartmann *et al.* (2017) realizaram um estudo analisando técnicas de PLN na qual treinaram 31 modelos de *word embedding*, sendo que este tipo de modelo correlaciona uma palavra dentro de uma matriz de correlação para conhecer as interrelações entre grupos de palavras, bem como também indiretamente poder predizer qual palavra possui

maior probabilidade de ser utilizada em uma sequência de palavras. Trabalhos semellhantes foram realizados por Aluísio *et al.* (2003), Baroni, Dinu e Kruszewski (2014) e Bengio *et al.* (2003).

Longe de representar uma lista exaustiva, a lista de estudos acima apenas indica o quanto esse campo de estudos tem se desenvolvido dentro e fora da Ciência da Informação. Acreditamos que esses estudos tendem a se tornar ainda mais numerosos, devido à necessidade da sociedade atual de consumir cada vez mais informação em menos tempo.

Indicamos, para os que desejarem aprofundar seu conhecimento no que diz respeito ao estado da arte sobre sumarização automática de textos, o trabalho de Rino e Pardo (2003), que realizaram um estudo bibliográfico, destacando as principais características das pesquisas nessa área, suas metodologias, bem como as tecnologias empregadas em alguns protótipos e softwares sumarizadores de textos.

Analisando a literatura científica da área de sumarização automática, verificamos que se trata de um campo em crescimento, com muitas abordagens ocorrendo paralelamente e com diversos objetivos, cada uma apresentando tanto méritos quanto dificuldades. As tendências da área apontam para a adoção de metodologias híbridas, priorizando a construção de sistemas voltados para a funcionalidade e com foco no usuário e, além disso, busca-se melhorar os métodos avaliativos, sejam com base computacional ou mesmo humana. Essa busca pode ser vista nos trabalhos de Iriguti e Feltrim (2019) e Costa e Bruno (2015).

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, não compondo um texto de fácil leitura. A solução proposta e desenvolvida na pesquisa e relatada neste artigo procura diminuir a dificuldade supracitada.

O método desenvolvido nesta pesquisa, que resultou no sumarizador por nós desenvolvido, é direcionado para a detecção do assunto mais relevante do texto, e a partir da frase que representa esse tema relevante, construir o resumo

com ele e os demais parágrafos que estejam associados 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 Processamento da Linguagem Natural (PLN) têm procurado soluções para as principais 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 resumo de um documento necessariamente prescinde de algum nível de compreensão do texto. Isso significa que é preciso detectar os maiores valores semânticos nele contidos.

Com essa compreensão, no escopo da pesquisa realizada, trabalhamos 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.

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, conectivos 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 na Frequência do Texto - Frequência inversa do documento (TF-IDF), do inglês *Text Frequency-Inverse Document Frequency*. 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.

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

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.

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

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 e, de fato, aumentam o nível de ruído nas respostas em um sistema de recuperação da informação, o que reflete o pensamento de Sparck-Jones (1972, p. 13, tradução nossa):

> Como os termos que ocorrem com muita frequência são responsáveis pelo ruído na recuperação, um caminho possível é simplesmente removê-los. O fato de isso reduzir o número de termos disponíveis para correspondência conjunta pode ser compensado pelo fato de que menos documentos não relevantes serão recuperados.

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.

3 MÉTODOS E TÉCNICAS EMPREGADOS NA PESQUISA

Na primeira fase da pesquisa, durante a concepção do protótipo, selecionamos um *corpus* de 100 textos pertencentes ao domínio “informação sobre energia limpa”.

A formação do corpus documental sobre o qual a pesquisa se debruçou teve duas etapas: uma manual e outra automatizada. A partir de páginas Web nacionais, foi realizada manualmente uma primeira busca por documentos

majoritariamente textuais, uma vez que o *software* sumariza apenas esse tipo de documento.

A equipe do projeto conseguiu compilar aproximadamente 150 textos, que foram analisados visando à identificação das principais estruturas semânticas presentes em cada documento, parágrafo por parágrafo, frase por frase.

Com base nesse primeiro grupo de documentos, elaboramos uma gramática, a partir do PLN e utilizando estatística, que dotaria o sistema com a capacidade de "entender" a estrutura de sentenças em língua portuguesa. Essa gramática foi utilizada para alimentar um sistema eletrônico capaz de vasculhar a web em busca de outros textos em português, o que nos possibilitou chegar a um segundo corpus, formado por 130.000 (cento e trinta mil) documentos relativos a temas diversos e publicados em jornais e revistas nacionais.

Em seguida, esse segundo córpus foi igualmente analisado e processado com base em PLN e estatística e o resultado desse processamento foi utilizado na identificação da relevância das frases e também na etapa final da construção automática dos resumos.

A terceira fase da pesquisa envolveu a seguinte sucessão de processos: redução da dimensionalidade e condensação semântica dos textos. Esses processos 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.

A redução da dimensionalidade foi alcançada usando-se uma metodologia híbrida baseada em PLN e estatística. A abordagem de PLN, aplicada ao problema da pesquisa, envolveu a remoção de palavras semanticamente menos relevantes no conjunto de documentos. Para a identificação dessas palavras foram utilizadas duas estratégias: uso de um *corpus* composto de um conjunto de palavras descartáveis (s*top words)*; e a remoção das palavras de baixa relevância.

Após a redução de dimensionalidade, o texto em geral apresenta-se

descaracterizado de suas propriedades de legibilidade.

O processo continuou estabelecendo-se a relevância das frases restantes. Para isso, foi necessário calcular o grau de relevância de cada parágrafo. O parágrafo mais relevante foi utilizado como “pivô” (ponto de partida) para a construção automática do resumo.

Para eliminarmos as frases menos relevantes no texto, prescindimos de uma classificação estatística.

Uma vez classificadas as frases (estatisticamente mais relevantes ou menos relevantes), procedemos ao descarte das frases menos relevantes observando sua correlação com a frase-pivô: caso a frase possuísse uma correlação maior do que o valor de descarte, ela não seria incorporada ao resumo.

Como mencionamos anteriormente, o protótipo apresentou resultados satisfatórios quanto à redução da dimensionalidade dos textos (na ordem de até 91%) e à velocidade de processamento e produção dos resumos (83 segundos).

Daí surgiu naturalmente a necessidade de avaliarmos a qualidade semântica e de síntese desses textos resumidos automaticamente.

Rino e Pardo (2003) argumentam que métodos automáticos de avaliação de softwares de sumarização automática de textos apresentam diversos problemas e não são tão satisfatórios como o julgamento humano. Partindo do argumento desses autores, realizamos testes cegos com bibliotecários que leram os textos completos e um conjunto de resumos desses textos, elaborados por humanos e pelo sumarizador automático.

Durante o segundo semestre de 2017, executamos um projeto de extensão, vinculado ao Departamento de Ciências da Informação da Universidade Federal do Ceará, cujo objetivo foi oferecer, aos graduandos do Curso de Biblioteconomia (predominantemente) e demais universitários interessados, oficinas teórico-práticas sobre a elaboração de resumos de textos não acadêmicos/científicos. Nessas oficinas, com duração de 8 horas/aula, ministramos conteúdos referentes aos tipos de resumo, suas características desejáveis, importância, etc., e, na segunda parte da oficina, os estudantes foram convidados a participar de uma atividade prática.

Criamos uma interface (*intranet*) onde cada estudante fez *login* e teve acesso a textos completos com extensão máxima de duas laudas, onde deveriam inserir um resumo para cada um deles, com base no conteúdo teórico explanado na primeira parte da oficina. Optamos por textos de pequena extensão para que os estudantes pudessem realizar a prática dentro do tempo de execução da oficina, o que também favoreceu que conseguissem fazer resumos para mais de um texto. Sem conhecimento dos estudantes, o sistema coletou dados sobre o tempo que cada um precisou para resumir cada texto.

Por meio dessa prática, pudemos reunir, em um banco de dados, 68 textos completos acompanhados de 2 resumos cada, elaborados por sujeitos diferentes.

Após a coleta dos resumos elaborados por humanos, submetemos os mesmos 68 textos à sumarização automática feita pelo *software* por nós desenvolvido, e criamos um banco de dados acessível pelos bibliotecários (voluntários) do Sistema de Bibliotecas da UFC, para avaliação dos resumos.

Por meio dessa ferramenta, cada avaliador pôde ler os textos completos que deram origem aos três resumos (dois humanos e um automático). Do lado esquerdo na interface da *intranet*, eles puderam ver um texto completo e, do lado direito dessa tela, um resumo de cada vez, com os critérios de avaliação logo abaixo. Após a avaliação do primeiro resumo, eles davam um comando e o sistema passava a mostrar o segundo e, assim, igualmente para o terceiro. O tempo que cada avaliador utilizou para analisar cada um dos resumos foi secretamente coletado, para que conhecêssemos a base temporal média despendida em cada avaliação.

Após uma verificação preliminar dos dados, percebemos que algumas avaliações foram realizadas em menos de 30 segundos. Essas avaliações podem ter sido influenciadas pela própria extensão dos textos (havia textos de tamanhos variados - meia lauda, uma lauda - para os quais foram gerados resumos de apenas um parágrafo em alguns casos), além de termos verificado que, à medida que os avaliadores passavam de um resumo para o outro, o tempo gasto na avaliação de cada um diminuiu, sendo que o terceiro resumo sempre levou menos tempo para ser avaliado do que o primeiro e o segundo.

Considerando que a qualidade dessas avaliações feitas em tempo reduzido (em menos de trinta segundos) pode ter sido impactada pela falta da devida atenção ou mesmo pelo possível cansaço por parte dos avaliadores, deliberamos descartar tais tuplas e considerarmos apenas aquelas em que todos os resumos foram analisados em mais de trinta segundos, o que representou um total de 20 textos com 3 resumos avaliados cegamente.

4 RESULTADOS E DISCUSSÕES

Os dados quantitativos (notas) referentes às avaliações dos bibliotecários foram planilhados e organizados graficamente para melhor visualização, seguindo os critérios eleitos como parâmetros de avaliação, nesta ordem: corretude gramatical, coerência e legibilidade, introdução de elementos não contidos no texto original, preservação das ideias centrais, extensão do resumo, se houve paráfrase ou cópia de fragmentos.

Quanto à **corretude gramatical**, a avaliação cega revelou haver uma relação de equivalência entre a qualidade dos resumos elaborados pelos humanos e os resumos produzidos pelo *software*. Laboratorialmente, antes mesmo da realização dos testes cegos aqui relatados, percebemos o quanto a correção gramatical estatística descrita por Souza *et al*. (2017), mostrou-se robusta, trazendo, como resultados, resumos com qualidade gramatical elevada e capazes de se fazerem confundir com resumos elaborados por humanos.

Os testes cegos atestaram esse fato, corroborando nossa impressão preliminar de que, lado a lado, um resumo humano e um produzido pelo protótipo têm aspectos gramaticais isomorfos. Essa é uma característica deveras singular, uma vez que outros sumarizadores automáticos descritos na literatura científica têm enfrentado a dificuldade de, após a fase de remoção de *stopwords* e reconstrução textual, entregar um produto gramaticalmente aceitável.

Analisando o parâmetro **coerência e legibilidade**, chegamos ao resultado mostrado na figura abaixo.

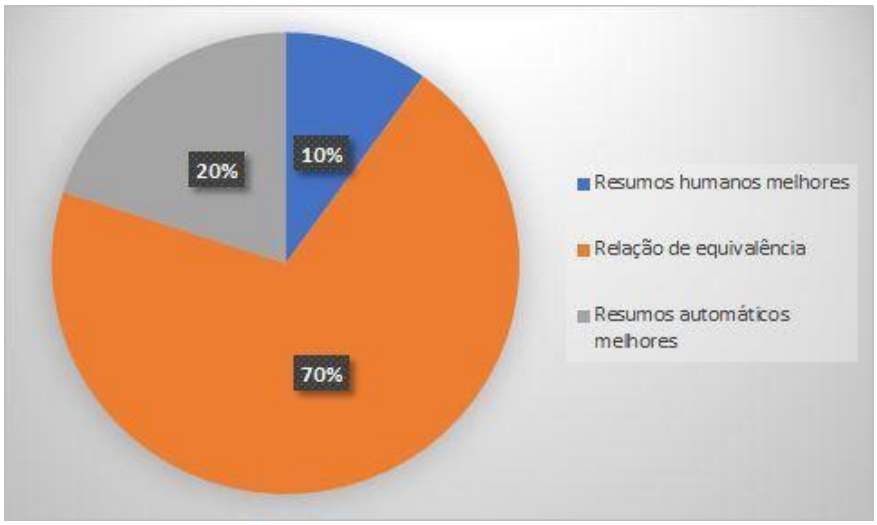
Figura 1 – Coerência e legibilidade

Fonte: Dados da pesquisa.

Analisando se os resumos apresentam coerência e legibilidade, estávamos procurando avaliar sua qualidade quanto à construção textual dotada de lógica, harmonia, clareza, ordem e vocabulário adequado, o que determina a facilidade de leitura e compreensibilidade da mensagem. Quanto a esse quesito, apenas um quarto dos resumos humanos foram avaliados como superiores em relação aos produzidos pelo *software*, sendo que, na maioria das vezes, houve equiparação entre eles.

Quanto à **introdução, nos resumos, de elementos não contidos no texto original**, os dados seguem conforme a figura 2.

Figura 2 – Inserção de informação inexistente no texto original



Fonte: Dados da pesquisa.

Um resumo deve condensar exatamente as ideias principais do texto e nada mais além disso, não devendo apresentar dados e informações não constantes no texto que lhe deu origem, mantendo, para com ele, total fidedignidade. Assim, foram considerados ruins os resumos onde houve algum tipo de inserção de quaisquer elementos não presentes no texto original e, por outro lado, aqueles avaliados como melhores foram os que se não criaram conteúdo.

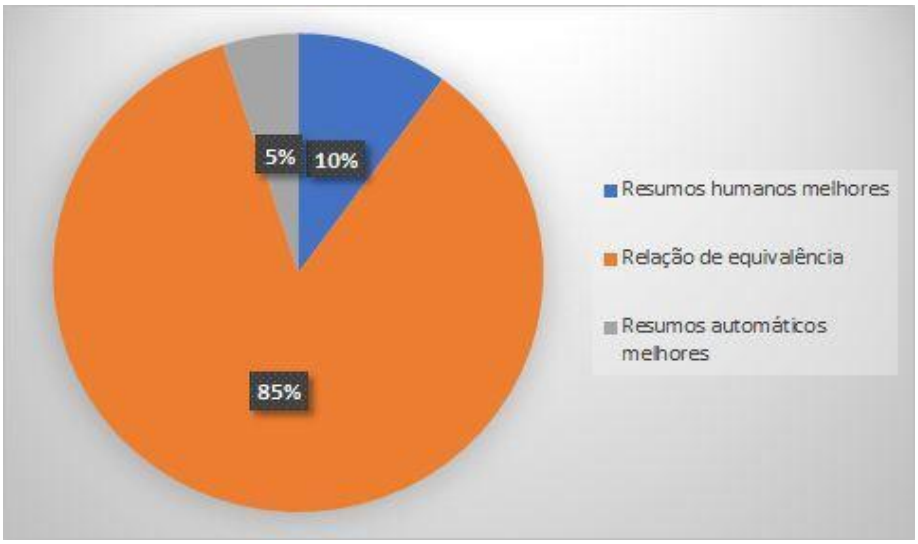
A relação de equivalência (70%) foi calculada quando percebemos que o resumo automático de um determinado texto não foi fidedigno - no sentido aqui colocado - e os resumos humanos também não foram; ou quando tanto o resumo automático quanto os humanos foram igualmente fidedignos.

Dessa forma, na maioria dos casos, ficou determinada a similaridade em número de erros e acertos entre o resumo feito pela máquina e os resumos feitos pelos estudantes.

Se é ruim quando o resumo introduz informações inexistentes no texto original, tanto pior é quando ele é incapaz de **preservar suas ideias centrais**. Isso significa que o sumarizador, seja automático ou humano, não foi capaz de identificar o valor semântico de determinados trechos do texto, ocasionando perda de informação relevante no resumo.

Quanto a esse aspecto, os dados analisados se mostraram conforme demonstrado na figura 3.

Figura 3 – Preservação das ideias centrais do texto



Fonte: Dados da pesquisa.

Obviamente a identificação do que há de mais relevante em um texto é tarefa que o ser humano executa de modo subjetivo, baseado em seu conhecimento de mundo, do assunto em pauta e da sua própria capacidade de interpretação textual.

As técnicas estatístico-matemáticas empregadas na concepção do protótipo aqui avaliado, explicitadas por Souza *et al.* (2017), para tornar o *software* capaz de identificar palavras e frases relevantes, revelaram-se profícuas a ponto de equiparar a capacidade da ferramenta com a capacidade humana de definir o que é mais ou menos relevante em um texto, na ordem de 85%. Nesse ponto, é oportuno enfatizar que a faculdade humana de interpretação de textos, abstração e julgamento só foi suficientemente superior à da máquina em 15% dos casos.

Quanto à **dimensionalidade dos resumos**, os dados apontam o que segue exposto na figura 4.

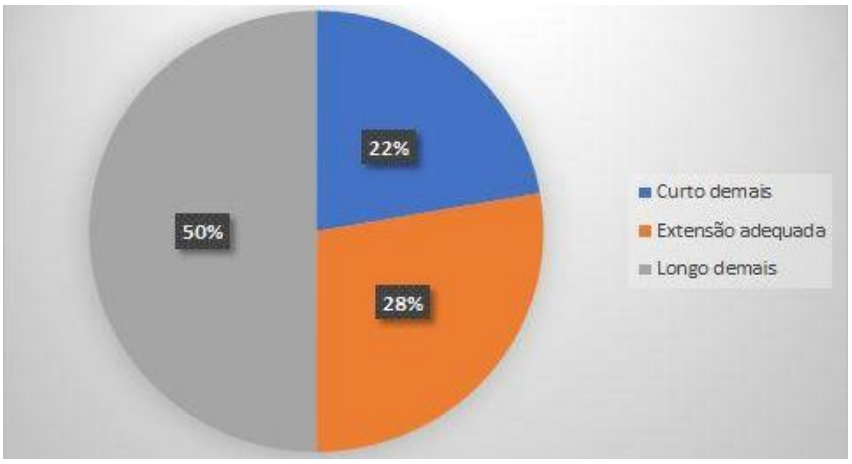
Figura 4 – Extensão dos resumos

Fonte: Dados da pesquisa.

Nesse ponto vemos o critério de avaliação onde os resumos automáticos não foram tão bem quanto gostaríamos. As fases de manutenção dos pontos relevantes e de eliminação de frases menos relevantes descritas por Souza *et al.* (2017) parecem não ter dado bons resultados na avaliação cega aqui discutida. Os resumos automáticos foram melhor avaliados nesse item apenas em 20% dos casos e se mostraram em nível de equivalência com o tamanho dos resumos humanos em 25%.

Como vimos, essa equivalência pode se dar quando o resultado é positivo, negativo ou neutro, ou seja, o resumo feito pelo *software* pode ter sido avaliado como equivalente ao humano ao ser ao longo demais, curto demais ou de extensão adequada. Por isso mostramos, na figura 5, os números aplicáveis apenas ao resultado da avaliação dos resumos automáticos.

Figura 5 – Extensão dos resumos automáticos



Fonte: Dados da pesquisa.

Dessa forma, o protótipo deve melhorar nesse item referente ao tamanho do resumo em relação ao texto correspondente, matéria sobre a qual o grupo de pesquisa deverá empreender esforços futuros para a melhoria da ferramenta.

Durante o processamento da linguagem natural visando à sumarização automática dos textos, diversas atividades/etapas foram realizadas pelo computador conforme relatado por Souza *et al.* (2017). Ao final do processamento, na fase reconstrução do texto condensado para a entrega do resumo, o *software* trabalhou com base no vocabulário existente e na gramática programada para a combinação dos elementos linguísticos que formam as sentenças, frases e orações.

Interessante observar que 25% dos resumos automáticos avaliados foram apontados como havendo neles algum tipo de **paráfrase do texto original**, restando aos outros resumos a característica básica de carregarem trechos transcritos do documento original. Ora, essa característica parece esperada, uma vez que se trata de um documento processado automaticamente e que, estatisticamente e algoritmicamente, reconstitui as partes mais relevantes de

outro documento maior.

No entanto, essa característica nos faz refletir sobre o que é mais adequado a um resumo produzido automaticamente. É melhor um *software* que resume textos transcrevendo seus trechos mais relevantes ou um capaz de fazer resumos parafrásicos? A transcrição/cópia de trechos atribuiria mais exatidão, credibilidade e fidedignidade ao resumo? A paráfrase soaria mais dinâmica e criativa?

O fato é que um sumarizador automático de textos capaz de processar estatisticamente um documento por meio de PNL, retirar o que é irrelevante ou secundário, preservando o teor mais significativo do texto, mantendo coerência e coesão textual e, como se tudo isso fosse simples, ainda ser capaz de condensar o documento utilizando outras palavras que não as ali contidas requer uma capacidade de "inteligência" muito maior e mais complexa, baseada em mais gramáticas e algoritmos associativos de palavras, expressões e até frases com alto grau de sinonímia.

Para melhor visualizarmos os dados obtidos pela pesquisa aqui relatada, sintetizamos no quadro abaixo os resultados a que conseguimos chegar com a avaliação cega realizada.

Quadro 1 – Síntese dos resultados da Avaliação cega

Critérios de avaliação	**Sumarização automática**	**Sumarização humana**
Corretude gramatical	Equivalentes	
Coerência e legibilidade	Equivalentes	
Introdução de elementos não contidos no texto original	Equivalentes	
Preservação das ideias centrais	Equivalentes	
Extensão do resumo	Pior	Melhor
Houve paráfrase ou cópia de fragmentos	Equivalentes	

Fonte: Dados da pesquisa.

Consideramos que todos os itens em que os resumos automáticos foram avaliados como equivalentes aos resumos humanos representam sucesso da ferramenta, uma vez que significa que o *software* é capaz de executar essa tarefa tão bem quanto o ser humano, com a vantagem de fazê-lo em menos tempo.

Esse é um resultado invulgar considerando que o *software* sumarizador de textos está em fase preliminar de construção, sendo um dos resultados de pesquisa em andamento sem financiamento de uma agência de fomento.

5 CONSIDERAÇÕES FINAIS

A partir da consecução dos objetivos deste estudo (priorizar uma avaliação qualitativa dos resumos automáticos gerados pelo nosso protótipo), percebemos a necessidade de melhorias no seu funcionamento, assim como de métodos avaliativos mais abrangentes, a partir de amostragens maiores e por uma população maior de avaliadores, sem perder de vista que, obviamente, qualquer avaliação, cega ou não, nos mesmos moldes da que realizamos, trará resultados parciais e subjetivos, impactados pela competência gramatical maior ou menor dos avaliadores, por exemplo.

Até aqui trabalhamos com apenas três avaliadores, não sendo nenhum deles formado em Letras ou com habilidades linguísticas inquestionáveis. Logo, sua emissão de parecer sobre corretude gramatical ou coerência textual, por exemplo, deve ser considerada com ressalvas, sem falar no nível de empenho pessoal de cada um deles em proceder a uma análise realmente criteriosa de tantos resumos em um tempo reduzido.

Com o desenvolvimento da pesquisa e o gradual aprimoramento da performance desse tipo de software, podemos vislumbrar um futuro onde os sumarizadores de textos poderão substituir o trabalho intelectual lento e cansativo dos bibliotecários em elaborar resumos de documentos e também viabilizar a sumarização automática de quaisquer tipos de texto na web, otimizando a forma como consumimos o grande volume de informação disponível hodiernamente.

Para o futuro, também é possível sonharmos com sumarizadores de conteúdo capazes de operar com outros tipos de documentos, não apenas os textuais, facilitando o acesso à informação relevante, precisa e em menos tempo. Outro capítulo desse sonho que alimentamos, é integrar essa ferramenta aos tradicionais ledores de tela para pessoas cegas ou com baixa visão, de modo que elas possam utilizar os resumos como forma de selecionar qual informação efetivamente querem consumir na íntegra.

REFERÊNCIAS

ALUÍSIO, S. M; PINHEIRO, G. M.; FINGER, M. NUNES, M. G. V; TAGNIN, S. E. **The lacioweb project:** overview and issues in brazilian portuguese corpora creation. [*S. l.*: *s. n.*], 2003. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.523.2121. Acesso em: 10 dez. 2019.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 52., 2014. Baltimore, Maryland. **Anais** [...]. Baltimore, Maryland: Associação de Linguística Computacional, 2014. p.238–247. 2014. Disponível em: https://www.aclweb.org/anthology/P14-1023/. Acesso em: 10 dez. 2019.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C. A neural probabilistic language model. **Journal of machine learning research**, v. 3, p. 1137-1155. 2003. Disponível em: http://jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. Acesso em: 10 dez. 2019.

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. Disponível em: https://repositorio.ufmg.br/bitstream/1843/ECID-7W5JH9/1/dissertacao_graciane_2009.pdf. Acesso em: 13 dez. 2019.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *In*: Conferência Internacional da World Wide Web (WWW 1998), 7., 1998, Brisbane, Austrália. **Anais** [...]. Brisbane, Austrália: Elsevier Science, 1998. p. 107-117. Disponível em: http://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf. Acesso em: 13 dez. 2019.

CABRAL, L. S. **Uma plataforma para sumarização automática de textos independente de idioma**. 2015. 138 f. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Pernambuco. Programa de Pós-Graduação

em Engenharia Elétrica, Recife, 2015. Disponível em: https://www.ufpe.br/documents/39830/745800/54_LucianoCabral/ef123409-aa67-4222-9fd2-4410708ef26d. Acesso em: 13 dez. 2019.

COSTA, M. A. A.; BRUNO, M. Uma comparação sistemática de diferentes abordagens para a sumarização automática extrativa de textos em português. **Linguamática**, v. 7, n. 1, p. 23-40. 2015. Disponível em: https://www.linguamatica.com/index.php/linguamatica/index. Acesso em: 10 dez 2019.

ERKAN, G.; RADEV, D. LexRank: graph-based lexical centrality as salience in text summarization. **J. Artif. Intell. Res. (JAIR)**, v. 22, p. 457-479, 2004. Disponível em: https://arxiv.org/abs/1109.2128. Acesso em: 13 dez. 2019.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. 23., 2003, Campinas. **Anais** [...]. Campinas: Jornada de Mini-Cursos de Inteligência Artificial, 2003.

HARTMANN, N. S.; FONSECA, E.; SHULBY, C.; TREVISO, M. V.; RODRIGUES, J. S.; ALUÍSIO, S. M. **Portuguese word embeddings:** evaluating on word analogies and natural language tasks. Nova Iorque: Universidade Cornell, 2017. Disponível em: https://arxiv.org/pdf/1708.06025.pdf. Acesso em: 14 jan. 2020.

IRIGUTI, A. H.; FELTRIM, V. D. Avaliando atributos para a classificação de estrutura retórica em resumos científicos. **Linguamática**, v. 11, n. 1, p. 41-53, 2019. Disponível em: https://linguamatica.com/index.php/linguamatica/article/view/273/451. Acesso em: 10 dez. 2019.

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.

PEREIRA, S. L. **Processamento de linguagem natural.** [*S. l.*: *s. n.*], 2011. Disponível em: https://www.ime.usp.br/~slago/IA-pln.pdf. Acesso em: 09 fev. 2019.

RINO, L. H. M.; PARDO, T. A. S. A. A sumarização automática de textos: principais características e metodologias. *In*: VIEIRA, R. (org.). **JAIA - Jornada de Atualização em Inteligência Artificial**. Campinas: [*s. n.*], 2003. p. 203-245.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, v. 24, n. 5, p. 513-523, 1988.

SOUZA, O.; TABOSA, H. R.; OLIVEIRA, D. M.; OLIVEIRA, M. H. S. Um método de sumarização automática de textos através de dados estatísticos e Processamento de Linguagem Natural. **Informação & Sociedade:** Estudos, João Pessoa, v. 27, n. 3, p. 307-320, set./dez. 2017. Disponível em: https://www.brapci.inf.br/index.php/article/download/60421. Acesso em: 28 jan. 2019.

SPARCK-JONES, K. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11-21, 1993.

TAKAMURA, H.; OKUMURA, M. Text summarization model based on the budgeted median problem. *In*: ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. 18., 2009, Hong Kong. **Anais** [...]. Nova Iorque: Association for Computing Machinery, 2009. p. 1589-1592. Disponível em: https://dl.acm.org/citation.cfm?id=1646179. Acesso em: 14 dez. 2019.

WANG, D.; LI, T. Document update summarization using incremental hierarchical clustering. *In*: ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. 19. 2010, Toronto. **Anais** [...]. Nova Iorque: Association for Computing Machinery, 2010. p. 279-288. Disponível em: https://dl.acm.org/citation.cfm?id=1871476. Acesso em: 14 dez. 2019.

EVALUATION OF PERFORMANCE OF A SOFTWARE OF AUTOMATIC SUMARIZATION OF TEXTS

ABSTRACT

Intrudoction: Since 2014 we have developed a research to produce a software (prototype) that would be able to elaborate automatic summaries of texts based on techniques of Natural Language Processing and frequency statistics of words. The first empirical tests of the tool generated results that indicated a significant reduction of the dimensionality of the texts, with considerable preservation of their semantic value. **Objective:** In this article, we present the results of the continuity of our investigative work, based on a human evaluation of the quality of these abstracts from blind tests. **Metodology:** A group of three librarians received a mixed and unidentified block of abstracts - produced by humans and the automatic abstracts made by the software - and carried out an evaluation, according to the criteria of grammatical correctness, preservation of central ideas, coherence and readability, extension of abstract, whether there was paraphrase or copy of fragments, and if there was introduction of ideas not contained in the original text. **Results:** The results showed that in four of the five evaluation criteria adopted, there was a qualitative equivalence between the abstracts produced by humans and those produced by the software, which seems to represent a relative success since the prototype could replace a person in the resume activity texts without leaving anything to be desired, except in the fifth evaluation center, referring to the dimension of the abstract, in which the text produced by the software was pointed out as extensive beyond what was necessary. **Conclusions:** Despite the good results of the prototype, we realized the need for improvements in its performance, as well as to evaluate it by more comprehensive methods, from more representative samples and by a larger group of evaluators.

Descriptors: Automatic summarization of texts. Access to information. Natural language processing. Mediation (Practice).

EVALUACIÓN DEL DESEMPEÑO DE UN SOFTWARE DE RESUMEN DE TEXTO AUTOMÁTICO

RESUMEN
Introduccion: Desde 2014 desarrollamos una investigación con el fin de producir un software (prototipo) que sería capaz de elaborar resúmenes automáticos de textos basados en técnicas de Procesamiento de Lenguaje Natural y estadísticas de frecuencia de palabras. Las primeras pruebas empíricas de la herramienta generaron resultados que indicaron una significativa reducción de la dimensionalidad de los textos, con considerable preservación de su valor semántico. **Objetivos**: En este artículo, presentamos los resultados de la continuidad de nuestro trabajo investigativo, a partir de una evaluación humana de la calidad de esos resúmenes a partir de la realización de pruebas ciegos. **Metodología**: Un grupo de tres bibliotecarios recibió un bloque mixto y no identificado de resúmenes - producidos por humanos y los resúmenes automáticos hechos por el software - y procedió a una evaluación, según los criterios de corrección gramatical, preservación de las ideas centrales, coherencia y legibilidad, en resumen, si hubo paráfrasis o copia de fragmentos y, si hubo introducción de ideas no contenidas en el texto original. **Resultados**: Los resultados mostraron que en cuatro de los cinco criterios de evaluación adoptados, hubo equivalencia cualitativa entre los resúmenes producidos por humanos y los producidos por el software, lo que parece representar un relativo éxito, ya que el prototipo podría sustituir a una persona en la actividad de resumir los textos sin dejar a desear, a no ser en el quinto creatorio de evaluación, referente al tamaño del resumen, en que el texto producido por el software fue señalado como extenso más allá de lo necesario. **Conclusiones**: a pesar de los buenos resultados del prototipo, nos dimos cuenta de la necesidad de mejorar su rendimiento, además de evaluarlo con métodos más completos, de muestras más representativas y de un grupo más grande de evaluadores.

Palabras clave: Sumarización automática de textos. Acceso a la información. Procesamiento del lenguaje natural. Mediación (Práctica).

Recebido em: 21/02/2019

Aceito em: 11/10/2019