Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Utilização de Métodos Extrativos na Sumarização Automática de Textos

Alice Picon Espina
Lucia Helena Machado Rino

NILC-TR-02-06

Março, 2002

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional

NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil


DC
Desenvolvimento para o Futuro
Departamento de Computação


ufscar

Sumário

Neste relatório, dois métodos extrativos de Sumarização Automática são descritos: o da palavra-chave e o de *Text Mining* aplicado à Sumarização Automática. Na construção de extratos para diferentes textos-fonte em português, o primeiro método foi explorado manualmente, com o auxílio do *software* de processamento estatístico *WordSmith Tools*, o qual forneceu a distribuição de freqüência dos componentes textuais. O segundo foi explorado com o auxílio de um protótipo, ao qual denominamos *TMSummarizer*, ou TMSumm. Os extratos foram produzidos manualmente na SA pelo método da palavra-chave e automaticamente pelo TMSumm e foram avaliados por juízes humanos. Além disso, eles foram comparados aos extratos produzidos de forma análoga pelo AutoResumo, do *Microsoft Word*.

Índice

1. Introdução

Métodos extrativos de Sumarização Automática (SA) consistem em a) identificar unidades de conteúdo que possam indicar a relevância dos segmentos textuais para a composição de um sumário; b) extrair do texto-fonte os segmentos textuais mínimos que incluam tais unidades; c) justapor cada um dos segmentos, resultando no sumário final. Em geral, os segmentos textuais relevantes são indicados por palavras ou elementos-chave mais complexos, cuja identificação é feita com base na análise da distribuição de freqüência das palavras do texto-fonte, considerando-se somente aquelas que possuem significado próprio, ou seja, as palavras de classe aberta (adjetivos, substantivos, advérbios e verbos). A etapa (a) é a mais importante de um método extrativo, pois dela depende, em grande parte, a qualidade dos extratos finais. Entretanto, a etapa (c) pode exigir, p.ex., modificações de pequena monta, para adequação do português.

Para ilustrar a SA extrativa, apresentam-se dois métodos, aplicados a textos em português: o método da palavra-chave (Luhn, 1958) e um método derivado daquele proposto por Larocca Neto et al. (2000), baseado em técnicas de *Text Mining* (TM). Um protótipo para sumarizar textos em português foi implementado a partir dessa proposta, ao qual deu-se o nome de *TMSummarizer* (*Text Mining Summarizer*), ou TMSumm. Ambos os métodos, ainda que de maneira distinta, baseiam-se no estudo da distribuição da freqüência de diferentes componentes textuais para a seleção de informações relevantes à construção de um extrato.

A denominação de *extrato*, aqui, se deve exclusivamente à metodologia *extrativa*, de identificação, seleção e *extração* de informações dos textos-fonte, para compor seus correspondentes sumários. Assim, um *extrato* é uma especificação terminológica de sumário, para a SA extrativa.

Apresentam-se, a seguir, os métodos da palavra-chave (Seção 2) e de TM (Seção 3), ilustrando sua utilização e discutindo seus resultados. Na Seção 4, são comparados os resultados gerados com base nos dois métodos e, adicionalmente, no AutoResumo. Finalmente, na Seção 5, é apresentada uma breve discussão dos procedimentos relatados.

2. O método da palavra-chave

2.1. Descrição

O método da palavra-chave (Luhn, 1958) consiste na extração de sentenças de um texto cujas palavras-chave sejam representativas de seu conteúdo ou idéia principal. A base deste método é a de que os termos-chave de um texto refletem o desenvolvimento das idéias do autor (Black and Johnson, 1988). Assim, sua reincidência e, logo, sua maior freqüência de ocorrência, é importante para determinar a importância que o autor atribui a suas idéias. Por essa razão as palavras em foco são as de classe aberta, já que são estas que melhor expressam o conteúdo textual.

Sua aplicação consiste, simplesmente, em se produzir a distribuição de freqüência dos componentes textuais e, a seguir, buscar as sentenças que os contenham. Em geral, utiliza-se um limite mínimo (*threshold*): os termos acima

dele indicam as sentenças mais relevantes. A seleção dos segmentos textuais para compor os extratos é feita da seguinte forma:

1º passo: gera-se a distribuição de freqüência de todas as palavras do texto-fonte;
2º passo: seleciona-se, dessa lista, a(s) palavra(s)-chave de classe aberta de mais alta freqüência (acima do *threshold*);
3º passo: produz-se o extrato correspondente, pela simples justaposição das sentenças que contêm as palavras-chave.

No trabalho relatado aqui, a geração da distribuição de freqüência foi feita com o auxilio do software *WordSmith Tools*[1], largamente utilizado para manipulação de corpora. Suas principais ferramentas, para o método extrativo em foco, são *WORDLIST, CONCORD* e *KEYWORDS*. As características dessas ferramentas são descritas abaixo:

WORDLIST: Gera listas de palavras baseadas em um ou mais textos, juntamente com a freqüência de cada uma delas nos textos processados e com sua freqüência em relação ao total de palavras, dentre outras informações estatísticas.
Essa ferramenta foi utilizada para gerar a lista de palavras à qual nos referimos no 1º passo.

KEYWORDS: Identifica as palavras-chave de um ou mais textos selecionados, a partir da *wordlist* gerada pela ferramenta *WORDLIST*. A fim de extraí-las dos textos, o *KEYWORDS* compara cada texto com outro texto ou com um conjunto de textos, chamado *corpus de referência*. Os domínios do texto sob análise e do corpus de referência devem ser o mesmo, para melhores resultados. Assim, baseando-se na freqüência que uma determinada palavra apresentou no texto estudado e comparando-a com a freqüência obtida no corpus de referência, o programa determina se essa palavra deve ser uma palavra-chave: palavras-chave são aquelas com alta freqüência no texto em foco e baixa freqüência no corpus de referência. Segundo esse critério, o programa pode não encontrar nenhuma palavra-chave ou pode indicar várias delas.
Essa ferramenta foi utilizada no 2º passo acima. Nos casos em que o *KEYWORDS* indicou mais de uma palavra-chave para o texto do corpus de teste, optou-se por selecionar apenas uma – a de maior freqüência – para produzir um extrato do texto-fonte em foco.

CONCORD: Busca em um ou mais textos o contexto de ocorrência de uma palavra específica. As entradas necessárias ao *CONCORD* são: a palavra de interesse e o(s) texto(s) sob investigação. Após a busca, os diversos contextos de ocorrência da palavra são apresentados, como ilustra a Figura 1, para a palavra 'Globo'.
Esta ferramenta foi utilizada para buscar, nos textos-fonte, as sentenças que continham a palavra-chave indicada no 2º passo. O 3º passo descrito, i.e., a produção do extrato, propriamente dito, foi realizado manualmente,

[1] Disponível em http://www.liv.ac.uk/~ms2928/wordsmit.html; notas em (Sardinha, 1996).

a partir dos resultados fornecidos pelo *CONCORD*. Vale notar, no entanto, que este processo poderia ser automatizado.

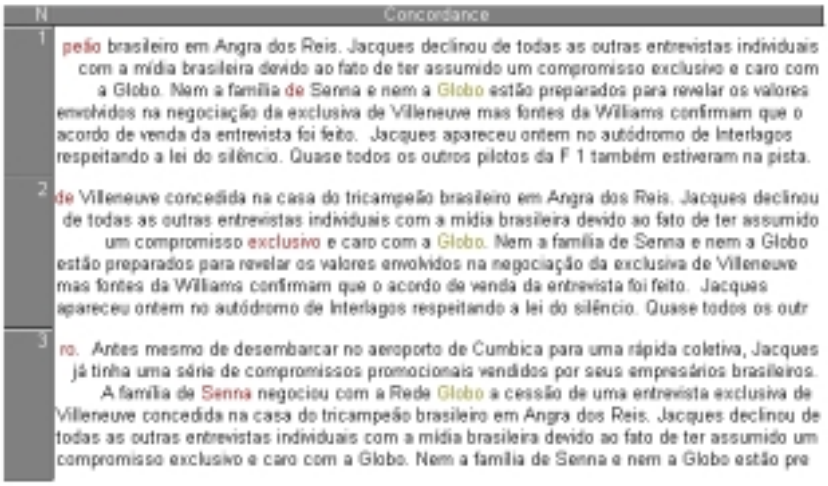

Figura 1: Contextos de ocorrência da palavra 'Globo'

2.2. Ilustração do método da palavra-chave

Como exemplo, esse método foi aplicado ao Corpus Esportes, composto por vinte crônicas esportivas, retiradas do corpus de textos corrigidos do NILC[2]. Um extrato para cada texto do corpus foi produzido. Para o Texto 1 abaixo[3], p.ex., o método de determinação das sentenças relevantes para compor o Extrato 1-PC é ilustrado. O texto-fonte é segmentado em parágrafos, para facilitar sua análise posterior. Outros exemplos de utilização do método encontram-se em (Martins et al., 2001).

Texto 1: O lucrativo Villeneuve

1. O novo herói da Fórmula 1, Jacques Villeneuve, viabilizará uma significativa injeção de recursos aos cofres da família e da fundação Ayrton Senna. O canadense, que teve sua entrada na F1 negociada pelo escritório de Senna e até hoje tem sua carreira administrada pelo grupo Senna, foi recebido no Brasil como uma galinha dos ovos de ouro.
2. Antes mesmo de desembarcar no aeroporto de Cumbica para uma rápida coletiva, Jacques já tinha uma série de compromissos promocionais vendidos por seus empresários brasileiros. A família de Senna negociou com a Rede Globo a cessão de uma entrevista exclusiva de Villeneuve concedida na casa do tricampeão brasileiro em Angra dos Reis. Jacques declinou de todas as outras entrevistas individuais com a mídia brasileira devido ao fato de ter assumido um compromisso exclusivo e caro com a Globo. Nem a família de Senna e nem a Globo estão preparados para revelar os valores envolvidos na negociação da exclusiva de Villeneuve mas fontes da Williams confirmam que o acordo de venda da entrevista foi feito.
3. Jacques apareceu ontem no autódromo de Interlagos respeitando a lei do silêncio. Quase todos os outros pilotos da F1 também estiveram na pista. Falta só o bicampeão mundial Michael Schumacher que tem chegada prevista para hoje. Rubens Barrichello também esteve ontem na pista.

[2] Núcleo Interinstitucional de Lingüística Computacional (www.nilc.icmc.sc.usp.br).
[3] Na reprodução, possíveis inadequações lingüísticas são mantidas como no original.

Passo 1: Construção da lista de freqüência de palavras do Texto 1 (*wordlist*)

A Tabela 1[4] abaixo ilustra somente parte da *wordlist* gerada pela ferramenta *WORDLIST*. Dessa tabela, são consideradas somente as palavras de classe aberta e, logo, as sombreadas.

Tabela 1: Freqüência de alguns componentes do Texto 1

N	Palavra	Freqüência	%
1	DE	15	6,85
2	A	8	3,65
3	DA	7	3,20
4	NA	5	2,28
5	O	5	2,28
6	SENNA	5	2,28
7	UMA	5	2,28
8	E	4	1,83
9	JACQUES	4	1,83
10	VILLENEUVE	4	1,83
11	COM	3	1,37
12	FAMÍLIA	3	1,37
13	**GLOBO**	**3**	**1,37**
14	NO	3	1,37
15	PARA	3	1,37
16	QUE	3	1,37
17	DO	2	0,91
18	DOS	2	0,91
19	ENTREVISTA	2	0,91
20	EXCLUSIVA	2	0,91

Passo 2: Seleção da palavra-chave para o Texto 1

Neste passo, a ferramenta *KEYWORDS* faz um cruzamento das palavras indicadas pela *wordlist* (Tabela 1) com as palavras do corpus de referência, para indicar as palavras-chave do texto em foco. Para o Texto 1, o corpus de referência também é o Corpus Esportes, agora composto por 34 textos. Para esse texto, o *KEYWORDS* só gerou a palavra-chave 'Globo', muito embora outras palavras-chave de igual ou maior freqüência de ocorrência tenham sido indicadas pela ferramenta *WORDLIST*. A geração das informações pela *KEYWORDS* é ilustrada abaixo[5]:

[4] N: ordem de freqüência de ocorrência da palavra; %: porcentagem de ocorrência da palavra em relação ao total de palavras do texto.

[5] N: ordem de freqüência de ocorrência da palavra; Freq: número de ocorrências no texto sob análise; Texto1.TXT %: porcentagem de ocorrência da palavra no texto-fonte; Freq: número de ocorrências no corpus de referência; Esportes.LST %: porcentagem de ocorrência da palavra no corpus de referência.

N	WORD	FREQ.	Texto1.TXT %	FREQ.	Esportes.LST %
1	GLOBO	3	1,37	0	28,5

Como se pode notar, a freqüência de 'Globo' no corpus de referência é nula, já que o Texto 1 foi retirado do Corpus Esportes e não havia outro texto em que tal palavra ocorria. É por esse motivo que essa é considerada uma boa palavra-chave do Texto 1.

Conforme os critérios descritos na seção anterior, pode haver textos para os quais o utilitário não seja capaz de sugerir palavras-chave. Isso ocorre quando eles contêm apenas palavras comuns aos textos do corpus de referência. Neste caso, adotou-se diretamente uma ou mais palavras-chave da lista de freqüência comum, *wordlist*, como a ilustrada na Tabela 1.

Neste trabalho, considerou-se apenas uma palavra-chave para cada texto-fonte, na geração de um extrato. No entanto, seria interessante construir extratos de modo diverso, p.ex., ou extraindo todas as sentenças cujas palavras-chave estão acima de um limite de representatividade (*threshold*) para compor exatamente um extrato, ou considerando todas essas palavras-chave, mas produzindo extratos isoladamente, i.e., um para cada uma delas, ou, ainda, simplesmente combinando diferentes palavras-chave.

Uma vez identificadas as sentenças de interesse, sua extração e justaposição para a produção do extrato, propriamente dito (3º passo), foram feitas manualmente, pois, apesar de tal tarefa ser totalmente automatizável, não havia um programa disponível para tal fim. Mais tarde, foi disponibilizado um protótipo que inclui tanto a produção da distribuição de freqüência (Pereira et al., 2002) quanto o método de SA extrativa descrito (Souza e Nunes, 2001).

O Extrato 1-PC[6] é o resultado da aplicação do método para o Texto 1, em função da palavra-chave 'Globo'. Entre parênteses, são dados o número total de palavras do extrato, sua taxa de compressão em relação ao texto-fonte e uma lista indicando os segmentos textuais que o compõem (parágrafo e posição da sentença no parágrafo).

> **Extrato 1-PC** (89 palavras; 59% de compressão; [par.2,sent.2; par.2,sent.3; par.2,sent.4])
>
> A família de Senna negociou com a Rede Globo a cessão de uma entrevista exclusiva de Villeneuve concedida na casa do tricampeão brasileiro em Angra dos Reis. Jacques declinou de todas as outras entrevistas individuais com a mídia brasileira devido ao fato de ter assumido um compromisso exclusivo e caro com a Globo. Nem a família de Senna e nem a Globo estão preparados para revelar os valores envolvidos na negociação da exclusiva de Villeneuve mas fontes da Williams confirmam que o acordo de venda da entrevista foi feito.

[6] A nomenclatura Extrato N-PC indica o N-ésimo extrato ilustrado para o método da palavra-chave (PC). Analogamente, 'Extrato N-AR' se referirá ao AutoResumo e 'Extrato N-TM', ao TMSumm.

A Tabela 2 apresenta as informações gerais para os 20 textos do Corpus Esportes, assim como a taxa de compressão para cada extrato gerado e sua correspondente avaliação por um juiz humano (o Texto 1 ilustrado acima corresponde ao texto-fonte # 20 na tabela)[7]. Os índices de avaliação seguem as métricas apontadas na Tabela 3 (os parâmetros de avaliação são detalhados na seção seguinte).

[7] Os textos indicados por '*' tiveram as correspondentes palavras-chave selecionadas com base nas freqüências da *wordlist*, já que o *KEYWORDS* não indicou nenhuma palavra-chave para os mesmos.

Tabela 2: Características dos textos e extratos correspondentes

Textos-fonte	# palavras-chave	# palavras texto	# palavras extrato	Taxa compressão (%)	Avaliação
1	7	356	193	45.8	1
2	1	392	171	56.4	1
3	1	386	200	48.2	2
4	1	380	82	78.5	4
5	1	358	130	63.7	1
6	3	401	78	80.6	3
7	3	431	108	75	1
8	3	387	66	83	3
9	4	363	201	44.7	1
10*	2	586	172	70.6	2
11	3	258	131	49.3	1
12*	1	158	46	70.89	3
13*	1	317	91	71.3	2
14*	1	418	177	57.7	2
15	1	291	184	36.77	1
16*	1	348	107	69.26	1
17	1	463	231	50.1	2
18	2	267	105	60.7	3
19	5	451	132	70.7	1
20	1	216	89	58.8	1

Tabela 3: Métricas para avaliação dos extratos

Qualidade do sumário	Descrição	Índices
Bom	Extrato é coerente & preserva idéia central	1
Razoável	Extrato NÃO é coerente, mas preserva idéia central	2
Ruim	Extrato é coerente, mas NÃO preserva idéia central	3
Péssimo	Extrato NÃO é coerente & NÃO preserva idéia central	4

2.3. Análise de desempenho do método da palavra-chave

Como se pode notar pela Tabela 3, consideram-se dois parâmetros para avaliação dos extratos gerados: a coerência e a preservação da idéia central do texto-fonte. A coerência é assegurada por construções lingüísticas coesivas, como, p.ex., a escolha de construções gramaticais que garantam a progressão temática do texto ou o uso de marcadores de discurso que auxiliem sua compreensão, permitindo a extração do seu significado. Este conceito, por sua vez, é inerente ao conceito de coerência textual: um texto incoerente é aquele ao qual não se consegue atribuir um significado. Denomina-se *textualidade* a medida que reflete a propriedade de um texto ser coerente e coeso ao mesmo tempo.

Assim, na avaliação dos extratos construídos a partir da distribuição de freqüência das palavras-chave, verificou-se 1) se a palavra-chave poderia ser usada para indicar os segmentos textuais suficientemente relevantes para transmitir a idéia principal do texto-fonte; 2) se o sumário era bom, segundo os parâmetros de textualidade.

Na avaliação subjetiva dos extratos usaram-se as métricas apontadas na Tabela 3: considera-se que o juiz humano é, simplesmente, um leitor do texto condensado (nativo do português), que deve atribuir um daqueles índices em sua avaliação.

Assim, na comparação do Extrato 1-PC com o Texto 1, verificou-se a correspondência das posições das sentenças extraídas com as suas posições originais, a fim de analisar sua coerência. Nota-se, assim, que ele é composto somente pelas três últimas sentenças do segundo parágrafo do Texto 1. Coincidentemente, foram justapostas as sentenças que contêm a palavra-chave 'Globo' que já estão justapostas originalmente. Considerando-se que o texto-fonte é coeso (hipótese verificável, aliás), a coesão se mantém naturalmente no extrato. Em relação à sua coerência, ao se fazer sua leitura corrida, nota-se que é possível apreender satisfatoriamente sua mensagem, a qual se aproxima da idéia central do texto-fonte. De fato, esta é expressa pelo título do texto: 'O lucrativo Villeneuve'. Vale ressaltar, ainda, que esse título deixa implícito, também, que o lucro é da Rede Globo, informação esta somente recuperável do texto exatamente a partir das três sentenças que compõem o extrato. Assim, além de coerente e coeso, o Extrato 1-PC também preserva a idéia central de seu texto-fonte.

A não seleção de sentenças de quaisquer outros parágrafos do Texto 1 também é justificável: por tratarem de informação de *background* (i.e., que pode ser previamente compartilhada com o leitor), as sentenças do parágrafo 1 podem ser excluídas; por se referirem a outro assunto, incluindo a lei do silêncio e a presença de outros pilotos na pista, as sentenças do parágrafo 3 também podem ser consideradas irrelevantes. Essas justificativas estão de acordo com a distribuição de freqüência dos componentes textuais (Tabela 1), assim como com o fato de a palavra-chave escolhida, 'Globo', estar ausente de tais parágrafos.

O Extrato 1-PC observa, portanto, as métricas de textualidade e preservação da idéia central do texto-fonte.

2.4. Síntese da análise para o Corpus Esportes

A síntese da análise para os 20 extratos produzidos a partir do método da palavra-chave é indicada pelas Figuras 2 e 3, que mostram, respectivamente, as taxas de compressão de cada extrato e a concentração dos mesmos por índice de avaliação (cf. Tabela 3).

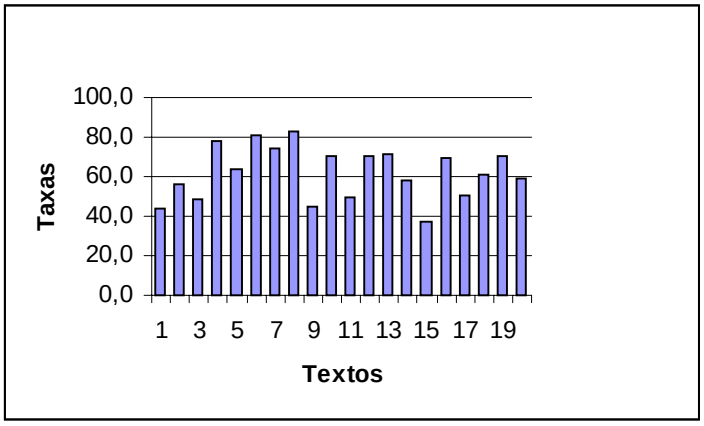

Figura 2: Taxa de compressão dos extratos

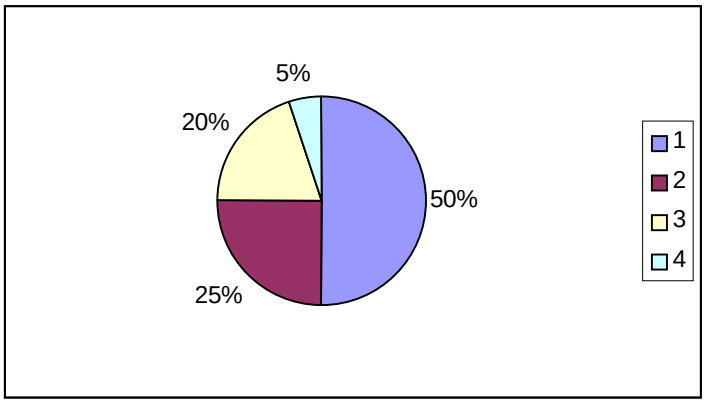

Figura 3: Concentração de sumários por avaliação

A Figura 3 mostra que 75% dos extratos gerados preservam a idéia central dos textos-fonte correspondentes, embora 25% desses não sejam coerentes, segundo a avaliação humana (vide Tabela 3). Ainda, as piores avaliações correspondem, em geral, às maiores taxas de compressão (cf. Tabela 2 & Figura 2), i.e., quanto mais curto o sumário, menor é a probabilidade de ele ser bom, segundo a avaliação dos juízes humanos. Apesar de somente 5% dos extratos haverem sido considerados péssimos, há espaço para melhora, considerando-se que o objetivo da SA extrativa, para nós, é produzir resultados *textuais*, i.e., coerentes e coesos.

2.5. Crítica geral ao método da palavra-chave

Os problemas apontados acima são comuns aos métodos de SA extrativa, principalmente no que diz respeito aos dois parâmetros adotados em sua avaliação: a textualidade e preservação da idéia central dos textos-fonte correspondentes. Além da taxa de compressão, outros fatores responsáveis por esses resultados dependem do gênero e domínio dos corpora escolhidos para a aplicação do método. Textos sobre esportes, p.ex., geralmente são escritos e destinados a pessoas com nível médio de proficiência na língua e possuem vocabulário bastante restrito, influindo, como conseqüência, também na estrutura e qualidade dos extratos correspondentes.

Problemas de textualidade, em geral, podem ser contornados se textos mais bem estruturados forem usados para a SA. Deve-se também observar que o método da palavra-chave, por ser "cego", não permite considerar outras características lingüísticas, tais como o posicionamento na sentença, a proximidade semântica dos componentes textuais, etc. Ao contrário, ao considerar as sentenças do texto-fonte cujos componentes sejam os de maior freqüência no texto-fonte, a representatividade ou organização dos demais componentes são completamente ignoradas e isto pode ser um problema, tanto para a preservação da idéia central do texto-fonte, quanto para a garantia de textualidade dos resultados automáticos.

Para explorar maior informatividade, outro método de SA, que utiliza técnicas de *Text Mining* para identificar as informações mais relevantes para compor os sumários, foi investigado, conforme segue.

3. Método de *Text Mining*

A diferença deste método em relação ao anterior é que ele usa outro nível informacional para a determinação do que é relevante para compor um sumário. Assim, embora ele também use a distribuição da freqüência dos componentes textuais, seus critérios adicionais buscam melhorar o desempenho dos clássicos sumarizadores extrativos. Abaixo seguem a descrição do método (3.1), acompanhada de uma ilustração (3.2) e da discussão de seu desempenho (3.3).

3.1. Descrição

O campo de *Text Mining* (TM), ou *Text Data Mining* (TDM – Hearst, 1999) é relativamente novo e consiste em se buscar as relações existentes entre componentes textuais. A sua principal diferença em relação ao campo de *Data Mining* (DM) é que este trabalha com dados estruturados, enquanto TM trabalha com dados não estruturados. Originalmente proposta para a Recuperação de Informação (RI), a técnica de TM explorada aqui é estendida para a SA por Larocca Neto et al. (2000), para a extração das sentenças mais relevantes de um texto.

O método parte da idéia de que, se uma palavra pode ser significativa em um texto-fonte por sua freqüência, ela também pode indicar a importância relativa das sentenças que a contêm. Trabalha-se com a noção da freqüência inversa de termos sentenciais (em geral, palavras), utilizando a medida chamada *Term Frequency–Inverse Sentence Frequency* ou, simplesmente, TF-ISF[8], dada pela fórmula:

$$\text{TF-ISF}(p,s) = \text{TF}(p,s)*\text{ISF}(p)$$

onde TF(p,s) é o número de vezes em que a palavra **p** ocorre na sentença **s** e ISF(p) é a significância de **p** em **s**, em relação à sua ocorrência no restante das sentenças do texto-fonte. ISF(p) é definida pela fórmula

$$\text{ISF}(p) = \log(|S|/\text{SF}(p))$$

onde $|S|$ é o número total de sentenças do texto e $\text{SF}(p)$ o número de sentenças em que a palavra **p** ocorre.

A partir da distribuição de relevância de p em cada sentença s, dada por $\text{TF-ISF}(p,s)$, pode-se obter o índice de relevância de cada sentença no texto-fonte, calculando-se a média TF-ISF de s.

Visando a SA, Larocca Neto et al. propõem que, quanto maior for a média TF-ISF de uma sentença, mais representativa ela será do conteúdo textual. Assim, extratos poderiam ser construídos estabelecendo-se um limite (*threshold*) mínimo para essa média: as sentenças acima do *threshold* comporiam o extrato. Esse valor pode variar, dependendo do grau de compressão desejado pelo usuário.

Os seguintes passos são indicados para a aplicação do método, que é parcialmente dependente da língua natural em foco. Larocca Neto et al. trabalham com a língua inglesa. Como este trabalho contempla o português, foi

[8] Essa métrica é análoga à TF-IDF (*Term Frequency–Inverse Document Frequency*) de Salton and Buckley (1988), aplicada para a determinação de termos-chave de documentos, i.e., com função análoga à do KEYWORDS.

preciso modificar os módulos dependentes da língua e simplificar a proposta original, devido a limitações de recursos computacionais para o português. O resultado dessas modificações é o sistema TMSumm (*Text Mining Summarizer*). A descrição dos passos envolvidos na SA extrativa, aqui, é feita, portanto, em função do TMSumm.

1º passo: Pré-processamento do texto de entrada, consistindo das seguintes etapas (as marcadas com '*' são dependentes da língua natural em foco):

- *Case Folding*: converte todos os caracteres de um documento a um só formato: maiúsculo ou minúsculo.. Assim: casa, Casa, cAsa, etc, podem ser igualmente convertidas para a *token* 'casa'.
- *Stemming**: Determinação do radical (*stem*) de cada uma das palavras. De forma geral, esta etapa consiste em eliminar sufixos das palavras.
- *Remoção de Stop Words**: Chamam-se *stop words* as palavras de classe fechada (i.e., preposições, conjunções, artigos, numerais, etc.) ou outras, que, no algoritmo de *Text Mining*, não tenham função para a identificação da relevância de segmentos textuais. Assim, considera-se que elas podem ser removidas do texto-fonte antes de este ser processado para a SA, propriamente dita. Assim, uma lista de *stop words*, ou *stoplist*, constitui um conjunto variável de palavras da língua natural em foco. No inglês, por exemplo, *can, will, do* e *does* poderiam ser elementos de uma *stoplist*. Para o português, a *stoplist* considerada é a seguinte: {a, e, o, as, da, de, do, na, no, os, ou, em, um, das, dos, mas, nas, nos, por, que, quê, seu, uma, como, essa, esse, mais, nada, pela, seus, porque, porquê, dessa, dessas, esses, desse, desses, desta, deste, nosso, nossos, aquilo, aquele, aquela, aqueles, aquelas, ninguém, ainda, através, menos, porém, contudo, todavia, entretanto, entanto, se, não, é}. Sua utilização será ilustrada na Seção 3.2.

Para o pré-processamento de textos-fonte em inglês, essas três etapas são executadas nessa ordem, sendo que as duas últimas podem, ainda, ser combinadas. Já para o correspondente processo no português, não há algoritmos de *stemming* apropriados. Neste caso, um processamento baseado em *n-grams* poderia substituí-lo. Um *n-gram* é uma parte de uma palavra consistindo de n caracteres. Por exemplo, a palavra DATA pode ser representada pelos trigramas _DA, DAT, ATA, TA_ ou pelos bigramas _D, DA, TA, entre outros.

Stems e *n-grams*, aqui, servem ao mesmo propósito: o de uniformizar a representação dos termos textuais, para a busca de termos relevantes para a SA. A representação em *n-grams* é particularmente importante para línguas de origem latina, para as quais nem sempre é possível estabelecer algoritmos de *stemming*. No TMSumm, não foi utilizada nenhuma dessas alternativas, sendo o pré-processamento constituído somente do *case folding* e da remoção de *stop words*.

2º passo: Delimitação de segmentos textuais

Em geral, a unidade textual mínima considerada é a sentença e, portanto, utilizam-se os símbolos usuais da própria língua natural, para delimitá-la (por exemplo, '.', '! ', '? '), seguidos ou não por um espaço em branco ou por um caractere que indique a quebra da linha sob análise.

Para a sentença S1 abaixo, por exemplo, os passos descritos acima resultam no fragmento S1'.

S1: Diamantino arremessou a bola.

S1': Diamantino arremessou bola

3º passo: Atribuição de pesos a cada palavra do segmento textual

A esta altura, o texto já fragmentado está disponível para o cálculo da freqüência de seus componentes e, portanto, para a identificação de seus segmentos relevantes. Para tanto, produz-se um vetor em que cada coordenada é dada por um par no formato [palavra,(TF,SF)], que armazena os valores individuais de freqüência de cada componente sentencial (freqüência do termo e freqüência na sentença). O vetor correspondente a S1', por exemplo, é dado por V1:

V1=[[Diamantino, (1,8)], [arremessou, (1,1)], [bola, (1,3)]]

4º passo: Cálculo do peso TF-ISF de cada palavra da sentença, segundo a fórmula anterior:

$$TF\text{-}ISF(p,s) = TF(p,s) * ISF(p)$$

Os pesos de cada componente sentencial são armazenados no vetor V1' (ilustrado abaixo), substituindo-se o par (TF,SF) de V1. Para S1', o vetor V1' atualizado resultaria, assim, em (valores fictícios):

V1'=[[Diamantino,0.60], [arremessou, 0.35], [bola, 0.40]]

5º passo: Cálculo da média TF-ISF para cada uma das sentenças

A média TF-ISF de cada sentença do texto é calculada com base na freqüência inversa de seus componentes, para se determinar sua freqüência relativa no texto. Assim, a média de S1 é dada por:

$$\text{Média S1} = \frac{\text{peso [Diamantino]} + \text{peso [arremessou]} + \text{peso [bola]}}{\text{nº total de palavras da sentença}}$$

Para os valores ilustrados em V1', essa média seria, portanto,

Média S1= 0.45

Após o cálculo da média de relevância de cada sentença no seu contexto textual, parte-se para a produção automática do extrato, descrita nos próximos passos.

6º passo: Cálculo da média MAX

A média MAX é utilizada para calcular o limitante inferior para a seleção das sentenças do extrato em construção (vide próximo passo). MAX é obtido a partir da maior média TF-ISF (sempre no intervalo 0-1), de todas as sentenças

do texto. Assim, para o texto exemplo composto por cinco sentenças (S1, S2, S3, S4, S5) abaixo e suas respectivas médias, MAX=0,6 (de S3).

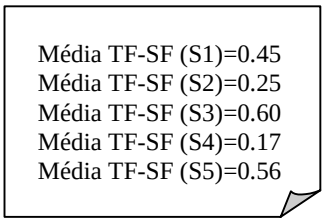

Figura 4: Texto Exemplo

7º passo: Cálculo do limitante inferior para seleção de sentenças

Este constitui o cálculo numérico final para aplicação do método e pode ser feito com a intervenção do usuário. Ao sugerir o comprimento desejado para seu extrato, por exemplo, C, esse cálculo será baseado em uma aproximação de C que seja viável para se determinar o número de sentenças do extrato. Ou seja, C indica um limitante para a compressão do texto-fonte: quanto mais próximo de zero, maior será o extrato e, logo, menor o índice de compressão; quanto mais próximo de um, maior será o índice de compressão do texto-fonte.

Para o Texto Exemplo da Figura 4, o limitante é calculado pela fórmula

Limitante = C x MAX

Supondo que o usuário queira um comprimento aproximado de seu extrato de C=0,7, o limitante do texto ilustrado será

Limitante = 0,7 x 0,6 = 0,42

8º passo: Produção do extrato

A partir do limitante, basta extrair as sentenças do texto-fonte, em função de seu valor médio TF-ISF: para cada sentença S, se média(S) >= Limitante, S é selecionada para compor o extrato. Assim, para o exemplo acima, as sentenças selecionadas serão S1, S3 e S5, levando ao extrato da Figura 5, composto por sua simples justaposição na ordem em que aparecem no texto-fonte.

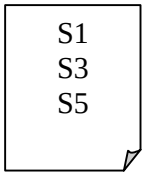

Figura 5: Extrato Exemplo

Diferentemente do método da palavra-chave, que pode ser aplicado com o auxílio de uma ferramenta comercial, como o *WordSmith Tools* ilustrado neste relatório, o método de *Text Mining* para a sumarização extrativapara o português foi integralmente implementado no DC/UFSCar[9], a partir da proposta de Larocca Neto et al. (2000). As adaptações dessa proposta se referem, sobretudo, ao módulo de pré-processamento (1º passo), devido à sua dependência da língua natural em foco. Os demais passos são independentes da língua e, assim, foram reproduzidos como se apresentam na proposta original a partir do 2º passo.

As particularidades do TMSumm são, assim, as seguintes:

a) Na fase de pré-processamento, somente foram removidas as *stop words* a partir da lista para o português já ilustrada. Não foram utilizados, aqui, nenhuns algoritmos de *stemming* ou de processamento de *n*-grams, já que estes não se aplicavam diretamente do inglês e, na ocasião do desenvolvimento deste projeto, não havia nenhum outro recurso lingüístico disponível para o português.
b) O TMSumm, foi implementado em Prolog e permite que o usuário determine o grau de compressão que ele deseja em seu sumário. A taxa de compressão, no TMSumm, não é linear, i.e., não é obtida, simplesmente, pela proporção entre o tamanho do extrato e seu texto-fonte correspondente. Isto se justifica pelo nível de granularidade dos extratos, que é sentencial. Assim, apesar de o usuário indicar uma porcentagem preferida para seu extrato, esta é aproximada, ao se delimitar somente sentenças completas para compô-lo.
c) Ao executar o TMSumm, é possível, também, obter outras informações de processamento, tais como: a porcentagem de compressão real, i.e., a calculada em função do extrato automaticamente, o número de palavras (ou tamanho) do texto-fonte, o número de palavras (ou tamanho) do extrato produzido, o número de *stopwords* contidas no texto-fonte, o *threshold* utilizado para a seleção das sentenças do extrato e o índice de compressão do texto-fonte. Essas informações são utilizadas na avaliação dos extratos, conforme veremos na Seção 4.

3.2. Ilustração do funcionamento do TMSumm

Nesta seção a geração de extratos de vários textos em português pelo TMSumm é ilustrada, avaliando-se os resultados e comparando-os com os resultados correspondentes gerados pelo AutoResumo do Word.

3.2.1. Seleção do corpus de teste

[9] Trabalho cooperativo de Camilla Brandel Martins (mestranda em Ciência da Computação) e Alice Picon Espina (graduanda em Ciência da Computação).

A sumarização extrativa no protótipo TMSumm explorou um corpus de 7 textos-fonte selecionados do Corpus Teses[10], o qual, por sua vez, é uma coletânea de textos do Corpus de Dissertações e Teses do NILC. Esse corpus, de gênero científico no domínio de Computação, consiste de introduções de artigos científicos e de monografias de mestrado e de conclusão de curso. Esse domínio foi escolhido por sua facilidade de compreensão, i.e., ele coincide com o próprio domínio deste trabalho. Além disso, os textos selecionados possuem seus próprios sumários autênticos (i.e., construídos pelos próprios autores dos mesmos) e são utilizados em outros projetos de SA, permitindo a perspectiva de, no futuro, comparar diversas metodologias de SA.

[10] Este corpus é utilizado em várias propostas de SA no Projeto EXPLOSA (FAPESP, Proc. Nro. 01/08849-8).

3.2.2. Delimitação do TMSumm

Nos exemplos ilustrados aqui, o TMSumm foi aplicado variando-se o valor limitante, para se gerar vários extratos. Foram consideradas as taxas de compressão de 90%, 75% e 50% dos respectivos textos-fonte, lembrando que o protótipo não trabalha com medidas exatas, mas com aproximações desse limitante. Os extratos gerados automaticamente correspondem, efetivamente, a aproximadamente 10%, 25% e 40% dos textos-fonte, tamanhos esses considerados na maioria dos trabalhos de SA que têm a língua inglesa como foco.

3.2.3. Conjunto de extratos produzidos automaticamente

Para cada um dos textos-fonte foram construídos três extratos, um para cada uma das taxas de compressão, resultando em 21 extratos. Uma amostra da aplicação do método segue abaixo, para o texto-fonte denominado Texto 2.

Texto 2: Uma interface amigável para o gerenciamento de configuração de software

1. Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. Todos esses itens, relativos ao software que está sendo construído ou modificado, constituem a "Configuração de Software". A configuração de software é a base para o entendimento de software desenvolvido e modificado por outras pessoas, e, portanto deve estar sempre atualizada. Para isso, os itens de informação precisam ser "controlados" durante todo o ciclo de vida do software.
2. O "Gerenciamento de Configuração de Software" é uma disciplina que ajuda a solucionar esse problema, pois envolve o desenvolvimento de padrões e procedimentos para administrar um software evolutivo. No entanto, devido à complexidade de suas tarefas, o gerenciamento de configuração de software necessita de ferramentas automatizadas que possam torná-las um pouco mais simples. O RCS (Revision Control System) é uma dessas ferramentas [TIC85].
3. O RCS é uma ferramenta de domínio público, desenvolvida para ambientes de Workstations e PC. O RCS para UNIX foi alvo de um estudo detalhado e constatou-se que o RCS é um sistema difícil de ser usado, principalmente devido à sua interface, constituída apenas de comandos de linha. Os usuários, principalmente aqueles não familiarizados com o sistema, encontram extrema dificuldade em utilizar e explorar toda a sua funcionalidade, pois apesar de haver poucos comandos, cada comando possui várias opções associadas, que também podem ser combinadas para obter resultados diferentes. Outro problema do sistema é que os resultados das operações ficam "soltos" na tela, dificultando o seu entendimento.
4. Assim o objetivo deste artigo é apresentar uma interface amigável para o sistema RCS, seguindo as noções básicas de Projeto de Interface. Possibilitando que o sistema seja utilizado com mais freqüência e interesse, através de uma interface mais "simpática" e fácil de utilizar, e além disso divulgar os importantes conceitos de gerenciamento de configuração de software, os quais se tornarão mais "visíveis" através da interface amigável.
5. Na seção seguinte, apresenta-se uma visão geral do RCS. Na seção 3, é feita uma breve descrição das vantagens de uma interface amigável para o RCS.

Seguem abaixo alguns extratos do Texto 2, gerados pelo TMSumm, cujas informações entre parênteses indicam suas propriedades mínimas (número de palavras e índice de compressão do texto-fonte) e as sentenças do texto-fonte que os compõem (parágrafo correspondente e posição da sentença no mesmo).

Extrato 1-TM (41 palavras; 88% de compressão; [par. 3, sent. 3])

Os usuários, principalmente aqueles não familiarizados com o sistema, encontram extrema dificuldade em utilizar e explorar toda a sua funcionalidade, pois apesar de haver poucos comandos, cada comando possui várias opções associadas, que também podem ser combinadas para obter resultados diferentes.

Extrato 2-TM (94 palavras; 72% de compressão; [par.1,sent.1; par.3,sent.3; par.4,sent.1;par.5,sent.2.])

Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. Os usuários, principalmente aqueles não familiarizados com o sistema, encontram extrema dificuldade em utilizar e explorar toda a sua funcionalidade, pois apesar de haver poucos comandos, cada comando possui várias opções associadas, que também podem ser combinadas para obter resultados diferentes. Assim o objetivo deste artigo é apresentar uma interface amigável para o sistema RCS, seguindo as noções básicas de projeto de interface. Na seção 3, é feita uma breve descrição das vantagens de uma interface amigável para o RCS.

Extrato 3-TM (178 palavras; 47 % de compressão; [par.1,sent.1; par.1,sent.3; par.2,sent.2; par.3,sent.2; par.3,sent. 3; par.4,sent.1; par.5,sent.2.])

Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. A configuração de software é a base para o entendimento de software desenvolvido e modificado por outras pessoas, e portanto deve estar sempre atualizada. No entanto, devido à complexidade de suas tarefas, o gerenciamento de configuração de software necessita de ferramentas automatizadas que possam torná-las um pouco mais simples. O RCS para Unix foi alvo de um estudo detalhado e constatou - se que o RCS é um sistema difícil de ser usado, principalmente devido à sua interface, constituída apenas de comandos de linha. Os usuários, principalmente aqueles não familiarizados com o sistema, encontram extrema dificuldade em utilizar e explorar toda a sua funcionalidade, pois apesar de haver poucos comandos, cada comando possui várias opções associadas, que também podem ser combinadas para obter resultados diferentes. Assim o objetivo deste artigo é apresentar uma interface amigável para o sistema RCS, seguindo as noções básicas de projeto de interface. Na seção 3, é feita uma breve descrição das vantagens de uma interface amigável para o RCS.

3.3. Crítica dos resultados do TMSumm

Como se pode notar, os extratos ilustrados não possuem boa qualidade. O Extrato 1-TM é totalmente descontextualizado e, logo, não permite recuperar a mensagem original, apesar de ser textual. Os demais, apesar de perderem um pouco de sua textualidade, são mais coerentes e, mesmo que implicitamente, como é o caso do Extrato 2-TM, sugerem que o grande volume de informações dificulta a exploração de um sistema por seus usuários, o que justifica o objetivo relatado.

Muitas vezes, os extratos produzidos levam a questionar, também, a própria qualidade dos textos-fonte escolhidos. Isto é particularmente importante na SA extrativa, cujos componentes escolhidos não sofrem qualquer alteração (ou sofrem mínimas alterações, quando muito). Entretanto, neste trabalho não foi feito esse tipo de análise.

A Tabela 5 apresenta uma síntese do desempenho do TMSumm, para as três taxas de compressão, segundo avaliação humana a partir dos mesmos índices indicados na Tabela 3 (Seção 2). Nessa avaliação, a verificação de preservação da idéia central foi feita objetivamente, considerando-se que, antes,

juízes humanos já haviam indicado a idéia central de cada texto-fonte a ser sumarizado. Ao contrário dos resultados avaliados no método da palavra-chave, portanto, aqui se tomou como base o fato de cada extrato ter que preservar a já apontada idéia central correspondente. A tarefa de identificação da idéia central constituiu um experimento adicional, que será relatado na Seção 4.2.

Tabela 5: Desempenho do TMSumm para o Corpus Teses

Textos-fonte	**TMSumm**		
	90%	75%	50%
1	1	1	1
2	2	1	1
3	4	2	1
4	4	1	2
5	4	4	4
6	4	1	1
7	4	2	1
Médias	3.3	1.7	1.6

A Tabela 6 apresenta uma síntese da avaliação geral do TMSumm, pela distribuição dos extratos em cada nível de compressão.

Tabela 6: Avaliação geral dos extratos produzidos pelo TMSumm

Qualidade do Extrato	**90%**	**75%**	**50%**
1	14,3%	57,2%	71,4%
2	14,3%	28,5%	14,3%
3	0	0	0
4	71,4%	14,3%	14,3%

Conforme se pode verificar, quanto à preservação da idéia central, com a porcentagem de compressão de 50% os resultados são praticamente iguais àqueles gerados com compressão de 75%, ou seja, abaixo de 75% de compressão do texto-fonte, os extratos se tornam menos condensados e apresentam melhor avaliação. Pode-se deduzir, daqui, que, por apresentarem mais informação, têm maior probabilidade de preservar a idéia central dos textos-fonte correspondentes e, ainda, ser mais coerentes que os mais condensados. Isto é compreensível, ao se considerar que, no método extrativo, quanto mais sentenças forem escolhidas para cada extrato, maior a probabilidade de não haver segmentos não coesos. Esta foi, também, a conclusão a que se chegou ao analisar os extratos produzidos pelo método da palavra-chave.

Os resultados gerados pelo TMSumm, cuja avaliação é apresentada nesta seção, foram ainda comparados aos resultados gerados pela ferramenta AutoResumo, do Word, conforme descrição a seguir.

4. A comparação do TMSumm com o AutoResumo

Em primeiro lugar, uma breve descrição do experimento com juizes humanos, que indicaram a idéia central de cada texto-fonte e, assim, a base para a comparação dos resultados automáticos, é apresentada. Também é fornecida uma breve descrição da ferramenta AutoResumo, cujos extratos para o mesmo Corpus Teses foram usados para a comparação com o TMSumm.

4.1. Descrição do experimento para indicação da idéia central de textos-fonte

Participaram desse experimento nove avaliadores humanos, falantes nativos do português, de diferentes graus de instrução, porém, com nível superior mínimo. Com exceção de um juiz, graduado em Letras, o restante pertencia à área de Computação. O experimento consistiu em uma tarefa de leitura e indicação da idéia central dos textos lidos (sete) e teve duração média de 3 horas. Os textos continham cada sentença numerada e, assim, essa indicação consistiu em apontar as sentenças que melhor indicavam a idéia central por seus números correspondentes. Partiu-se da hipótese, aqui, que a idéia central pode estar "difusa" no texto-fonte e, logo, não haveria somente uma sentença mais significativa, mas um conjunto delas.

Além dessa tarefa, solicitou-se aos juizes que indicassem se tinham ou não familiaridade com os textos lidos. A Tabela 7 indica que 61% deles era familiar com o assunto (S=familiar, N= não familiar). Apesar de esse índice, na média, ser baixo, assumiu-se sua indicação da idéia central de cada texto como *baseline* para as avaliações dos extratos automáticos.

Tabela 7: Familiaridade dos juizes humanos com o assunto do corpus de teste

Juiz	Texto 1	Texto 2	Texto 3	Texto 4	Texto 5	Texto 6	Texto 7
1	S	N	S	S	S	N	N
2	S	S	S	S	S	S	S
3	S	N	N	S	S	S	N
4	S	N	N	N	N	N	N
5	S	N	N	N	S	S	S
6	S	N	N	N	N	N	N
7	S	S	S	S	S	S	S
8	N	N	S	S	S	S	S
9	S	N	S	S	N	S	S
Médias (%)	89	22	56	67	67	67	56

4.2. Ferramenta AutoResumo

O AutoResumo é uma ferramenta incorporada ao Microsoft Word (Janela 'Ferramentas', Opção 'AutoResumo'), para sumarizar textos ou documentos. Sua utilização é bastante simplificada. Basta acessá-lo e dizer qual a compactação desejada.

A ferramenta foi aplicada ao Texto 2, resultando em extratos de, aproximadamente, 10%, 25% e 50% dos textos originais, conforme ilustrações abaixo.

Extrato 1-AR (26 palavras - 90% de compressão)

> Todos esses itens, relativos ao software que está sendo construído ou modificado, constituem a "Configuração de Software".
>
> Na seção seguinte, apresenta-se uma visão geral do RCS.

Extrato 2-AR (95 palavras - 75% de compressão)

> Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. Todos esses itens, relativos ao software que está sendo construído ou modificado, constituem a "Configuração de Software".
>
> O "Gerenciamento de Configuração de Software" é uma disciplina que ajuda a solucionar esse problema, pois envolve o desenvolvimento de padrões e procedimentos para administrar um software evolutivo. O RCS (Revision Control System) é uma dessas ferramentas [TIC85].
>
> Na seção seguinte, apresenta-se uma visão geral do RCS. Na seção 3, é feita uma breve descrição das vantagens de uma interface amigável para o RCS.

Extrato 3-AR (152 palavras - 50% de compressão)

> Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. Todos esses itens, relativos ao software que está sendo construído ou modificado, constituem a "Configuração de Software". A configuração de software é a base para o entendimento de software desenvolvido e modificado por outras pessoas, e portanto deve estar sempre atualizada.
>
> O "Gerenciamento de Configuração de Software" é uma disciplina que ajuda a solucionar esse problema, pois envolve o desenvolvimento de padrões e procedimentos para administrar um software evolutivo. O RCS (Revision Control System) é uma dessas ferramentas [TIC85].
> O RCS para UNIX foi alvo de um estudo detalhado e constatou-se que o RCS é um sistema difícil de ser usado, principalmente devido à sua interface, constituída apenas de comandos de linha.
> Na seção seguinte, apresenta-se uma visão geral do RCS. Na seção 3, é feita uma breve descrição das vantagens de uma interface amigável para o RCS.

Neste caso, nota-se que o extrato menos problemático é o Extrato 3-AR, ou seja, o que apresenta taxa de compressão de somente 50%.

4.3. Comparação dos extratos gerados automaticamente pelo AutoResumo e TMSumm

Vale lembrar que, na comparação descrita aqui, foram utilizados os 7 textos-fonte do Corpus Teses, para produzir três sumários com as taxas de compressão de 90, 75 e 50%, respectivamente. Assim, foram avaliados 42 extratos, segundo o seguinte:

1) A comparação dos extratos produzidos por ambos os aplicativos (AutoResumo e TMSumm) se deu considerando que ambos comprimem de modo similar os textos-fonte. Entretanto, vale lembrar que, no TMSumm, a taxa de compressão é aproximada, para delimitar a unidade textual a compor o extrato ao nível sentencial. Aparentemente, medida similar é adotada no AutoResumo.
2) A verificação da preservação da idéia central nos extratos se baseou no experimento com juizes humanos, relatado anteriormente.

A Tabela 8 indica a ocorrência da idéia central de cada texto-fonte nos extratos correspondentes, para cada sumarizador em foco (S = o extrato a preserva; N = o extrato não a preserva). Como se pode verificar, o TMSumm apresentou melhores resultados do que o Auto Resumo, em relação à preservação da idéia central nos extratos. O AutoResumo não produziu nenhum extrato cuja idéia central fosse preservada, segundo o julgamento humano. Já o TMSumm apresenta essa condição para os extratos cujo índice de compressão não é máximo, confirmando os resultados anteriormente apresentados (particularmente, os extratos que representam somente 10% dos textos-fonte foram os que pior preservaram a idéia central).

Tabela 8: Preservação da idéia central nos extratos de diversos tamanhos

Textos-fonte	**AutoResumo**			**TMSumm**		
	90%	75%	50%	90%	75%	50%
1	S	S	S	S	S	S
2	N	N	S	S	S	S
3	N	S	S	N	S	S
4	N	N	N	N	S	S
5	N	N	N	N	N	N
6	N	N	N	N	S	S
7	N	N	N	N	S	S
Médias	0.14	0.29	0.43	0.29	0.86	0.86

A Tabela 9 mostra o número de palavras dos extratos produzidos por ambos os aplicativos, para evidenciar que, embora estejamos tratando com taxas de compressão supostamente iguais, há uma pequena variação, devido à adaptação do TMSumm ao cálculo da porcentagem.

Tabela 9: Tamanho dos textos-fonte e correspondentes extratos

Texto-fonte	**nº palavras do texto-fonte**	**AutoResumo**			**TMSumm***		
		nº palavras extrato/taxa compressão			**nº palavras extrato/taxa compressão**		
		90%	**75%**	**50%**	**90%**	**75%**	**50%**
1	586	51	161	283	69	194	303
2	681	64	174	327	68	187	369
3	943	86	213	482	126	272	497
4	712	75	182	366	69	199	365
5	726	65	153	371	93	218	366
6	660	58	157	326	75	191	348
7	334	26	95	152	41	94	180

5. Conclusões

De um modo geral, os extratos gerados manual (no caso do método da palavra-chave) ou automaticamente (no caso do TMSumm) foram satisfatórios, se consideradas as métricas simples de preservação da idéia central dos textos-fonte e garantia de textualidade e, ainda, os gêneros de textos em foco na exploração de cada um dos métodos propostos.

Como já foi observado, o gênero de crônicas esportivas é voltado a um leitor de formação média em sua própria língua nativa e, por essa razão, os textos apresentam vocabulário e estruturas simples. O gênero científico, particularmente o do Corpus Teses, também não apresenta maiores problemas, embora seu vocabulário seja mais técnico do que o outro. Em termos estruturais, as frases desses textos são bem marcadas, razão pela qual os juizes humanos não tiveram dificuldade em apontar suas idéias principais (experimento relatado na Seção 4.1).

Essas características interferem, de certa forma, também na qualidade dos extratos. Entretanto, é importante observar que, nas avaliações relatadas neste relatório, a qualidade melhora sensivelmente quando se consideram extratos maiores, pois eles incluem mais informações dos textos-fonte correspondentes, tendendo, assim, a minimizar os problemas de falta de coesão ou coerência textuais.

Em relação à idéia central, os métodos apontados não privilegiam qualquer estratégia para seu reconhecimento, a não ser o fato de se supor, no caso de ambos os métodos, que a maior incidência de um termo implica sua maior relevância no texto. Entretanto, não existem estudos conclusivos sobre a relação entre maior incidência, relevância e idéia central. Muitos autores discutem a singularidade informacional de um texto como sendo seu fato inovador e, muito provavelmente, por isso mesmo sua idéia central. Nesse caso, é evidente que uma informação singular não será palavra-chave, para a aplicação do método da palavra-chave. Entretanto, ela pode ser o foco do método de *text mining*, já que este trabalha com a freqüência inversa do texto-fonte.

Quando comparados para o Corpus Esporte, o método que apresentou melhores resultados, para índices de compressão compatíveis, foi o do TMSumm; este também foi considerado melhor para o Corpus Teses, agora em relação ao AutoResumo. No entanto, o volume de textos processados foi muito pequeno, para que se possa afirmar que a SA com base em TM é superior à baseada em palavras-chave. Larocca Neto (2002), p.ex., explorou-o para sumarizar textos jornalísticos de gênero político em inglês, obtendo resultados bastante insignificantes, comparáveis ao pior caso da SA extrativa, que é o de geração de extratos pela escolha de sentenças aleatórias.

Outras propostas de SA extrativa têm sido sugeridas mais recentemente, envolvendo processamento superficial baseado em características lingüísticas clássicas dos textos-fonte a sumarizar, como, p.ex., o cômputo das cadeias lexicais mais fortes que, ao contrário das palavras-chave mais representativas, indicariam o *contexto* mais relevante. De um modo geral, tais propostas se inclinam a modelar de modo cada vez mais informativo o processamento superficial de textos, buscando extratos de melhor qualidade.

Referências Bibliográficas

Black, W.J. and Johnson, F.C. (1988). A Practical Evaluation of Two Rule-based Automatic Abstracting Techniques. Expert Systems for Information and Management, Vol.1, N.3, pp. 159-177. Manchester.

Hearst, M. A. (1999). Untangling Text Data Mining. In: *Proceedings of ACL´99, the 37th Annual Meeting of the ACL*, University of Maryland, USA.

Larocca Neto, J. (2002). *Contribuição ao Estudo de Técnicas para Sumarização Automática de Textos*. Dissertação de Mestrado. PUC-PR. Curitiba, PR.

Larocca Neto, J.; Santos, A. D.; Kaestner, C. A. A. & Freitas, A. A. (2000). Document clustering and text summarization. *In Proceedings of the 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining (Padd-2000)*, 41-55. London: The Practical Application Company.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, pp. 159-165.

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à Sumarização Automática*. Rel. Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos. Abril, 2001. 38p.

Pereira, M.B.; Souza, C.F.R. e Nunes, M.G.V. (2002) Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. Submetido à Revista Eletrônica de Iniciação Científica, SBC.

Salton, G. and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* 24, pp. 513-523.

Sardinha, A. P. Berber (1996). *Review: WordSmith Tools*. Disponível no site http://info.ox.ac.uk/ctitext/publish/comtxt/ct12/sardinha.html

Souza, C.F.R. e Nunes, M.G.V. (2001). Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português. Relatórios Técnicos do ICMC-USP (NILC-TR-01-9), Nro. 153, Outubro, 14p.