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

GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos

Thiago Alexandre Salgueiro Pardo

NILC-TR-02-13

SETEMBRO, 2002

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório é apresentado o GistSumm – *GIST SUMMarizer* – um sumarizador automático de textos baseado em sua idéia principal. O GistSumm utiliza um método extrativo inédito que usa uma única sentença – a que melhor representar a idéia principal do texto a sumarizar – para selecionar os fragmentos textuais que comporão o sumário.

ÍNDICE

1. Introdução

A sumarização automática de textos tem se tornado uma área proeminente devido à crescente demanda por informação no menor tempo possível. Com o advento da internet, onde as pessoas se vêem em um mar de informação em constante expansão e atualização, um grande interesse acadêmico, comercial e governamental surgiu por essa área.

Há diversos tipos de sumários e métodos de sumarização. Previsões meteorológicas, sinopses de novelas, chamadas de notícias jornalísticas, resenhas e *abstracts* de livros e teses, por exemplo, podem ser considerados sumários. Estes podem ser classificados como *indicativos*, *informativos* ou *críticos* (Mani and Maybury, 1999): sumários indicativos apenas listam ou indicam o assunto principal dos textos-fonte; os informativos são autocontidos, isto é, possuem toda a informação essencial dos textos-fonte, dispensando a leitura destes; os críticos avaliam ou apenas comentam o conteúdo de suas fontes. Os métodos de sumarização, por sua vez, podem ser divididos em duas grandes abordagens: a superficial e a profunda. A abordagem superficial utiliza dados estatísticos e/ou empíricos para a sumarização, enquanto a profunda procura utilizar teorias formais e modelos lingüísticos[1].

Em vista do alto custo e das dificuldades em se estabelecer e manipular o conhecimento lingüístico e extralingüístico necessário para desenvolver um sumarizador pela abordagem profunda, adota-se, neste trabalho, a abordagem superficial que, apesar de suas limitações, isto é, não utilizar todo o conhecimento em potencial do texto a sumarizar e produzir sumários de baixa qualidade, tem baixo custo e, dependendo da aplicação a que se destina, pode produzir resultados satisfatórios e úteis. Assim, este relatório descreve a implementação de uma abordagem superficial *inédita* de sumarização que resultou no sumarizador automático de textos GistSumm – *GIST SUMMarizer* – o qual procura simular a forma de sumarização humana. Quando uma pessoa sumariza um texto, ela procura identificar a *idéia principal* deste (que, em inglês, é o *gist* do texto) e as informações do texto que a complementem (Mani, 2001; Rino, 1996). O quanto dessa informação complementar é introduzido no sumário depende diretamente do tamanho desejado do sumário[2]. Similarmente, o processo de sumarização no GistSumm é disparado pela idéia principal do texto-fonte (o texto a ser sumarizado): a partir da sentença que melhor representa essa idéia, selecionam-se as sentenças desse texto que comporão seu sumário. O processo principal do GistSumm consiste, portanto, na identificação da sentença principal do texto-fonte e na forma como as outras sentenças são selecionadas em função desta. A sentença principal será referenciada, neste relatório, como "sentença-*gist*" do texto.

Além das características acima, uma questão importante quanto ao GistSumm, que será abordada posteriormente, é que, embora personalizado, no momento, para a língua portuguesa, ele é independente de qualquer língua natural, já que seu método de sumarização é independente de língua. A forma como ele pode ser personalizado para outras línguas é comentada na Seção 5.

A próxima seção apresenta a caracterização do GistSumm, mostrando sua arquitetura e as premissas e hipóteses que guiaram seu desenvolvimento. A Seção 3 descreve o método de sumarização utilizado, enquanto a Seção 4 apresenta a interface para sua manipulação. As considerações finais são realizadas na Seção 5.

---

[1] Para mais detalhes sobre métodos de sumarização, vide Martins et al. (2001).

[2] A informação complementar a ser inserida no sumário depende também do nível de detalhe desejado no sumário. Entretanto, em geral, esse nível de análise não pode ser medido por métodos superficiais.

2. Caracterização do GistSumm

2.1. Premissas e hipóteses

Como a geração do sumário no GistSumm baseia-se na idéia principal do texto-fonte, isto é, o texto a ser sumarizado, considera-se as seguintes premissas:

1) todo texto veicula uma idéia principal;
2) é possível identificar em um texto uma sentença que melhor representa sua idéia principal, isto é, a sentença-gist.

Com base nessas premissas, as seguintes hipóteses para o desenvolvimento do GistSumm foram delineadas:

1) por meio de métodos estatísticos simples, é possível identificar a sentença-gist de um texto-fonte ou, pelo menos, uma sentença que se aproxime significativamente dela;
2) conhecendo-se a sentença-gist, é possível produzir sumários coerentes por meio da justaposição de sentenças do texto-fonte relacionadas à sentença-gist, sendo estas complementares à idéia principal do texto-fonte.

Por meio de uma análise preliminar, a hipótese 1 foi comprovada de forma satisfatória, conforme será relatado na Seção 5. A hipótese 2, por sua vez, deverá ser verificada, futuramente, por uma avaliação sistemática dos sumários gerados automaticamente pelo GistSumm.

A subseção seguinte descreve a arquitetura do GistSumm.

2.2. Arquitetura do sistema

A Figura 1 mostra a arquitetura do GistSumm. Caracterizado como um sumarizador extrativo, ele executa os seguintes passos durante a sumarização:

- primeiramente, as sentenças do texto-fonte são delimitadas fazendo uso dos sinais de pontuação tradicionais (ponto final, de exclamação e de interrogação) presentes no texto;
- as sentenças são então ranqueadas, isto é, ordenadas pelos seus valores/pontos calculados pelo *método de ranqueamento* selecionado pelo usuário, que, no GistSumm, são dois: *Keywords* e *TF-ISF*. Conforme será discutido posteriormente, esses métodos são, de fato, métodos de *determinação da idéia principal* do texto a sumarizar. A sentença com maior pontuação no ranque é considerada a sentença-gist, a partir da qual as outras sentenças que comporão o sumário serão determinadas. Nesta etapa, para aprimorar os resultados do processo de sumarização, conforme será explicado na Seção 3, utiliza-se uma *stoplist* para o português, ou seja, uma lista de palavras muito comuns (e, portanto, normalmente sem relevância para a sumarização) e um léxico simples, contendo somente as palavras do português e suas correspondentes formas canônicas;
- por fim, respeitando a taxa de compressão especificada pelo usuário, selecionam-se as sentenças que formarão o sumário.

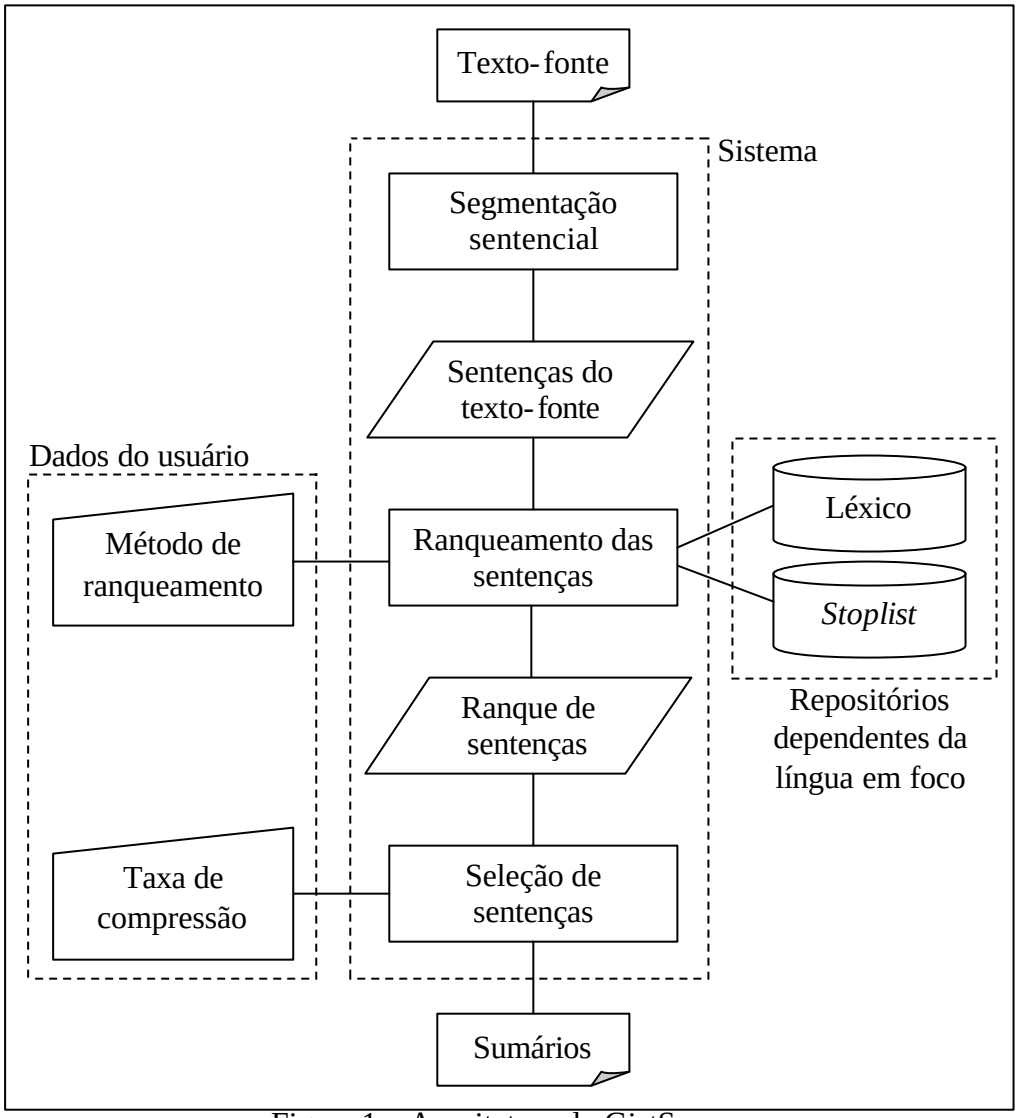


Figura 1 – Arquitetura do GistSumm

A *stoplist* foi montada manualmente e contém 196 palavras, chamadas, neste caso, de *stopwords*. As *stopwords* abrangem artigos, pronomes, preposições, conjunções e interjeições, retiradas, principalmente, de Cunha e Cintra (2001). A listagem completa das *stopwords* encontra-se no Apêndice A.

O léxico simplificado do GistSumm foi extraído automaticamente do léxico do NILC[3] (Núcleo Interinstitucional de Lingüística Computacional), o qual é o maior léxico para a língua portuguesa do Brasil (Nunes et al., 1996). Em casos de palavras com mais de uma categoria gramatical associada, foi selecionada a canônica da classe gramatical mais freqüente[4].

A próxima seção mostra o processo completo de sumarização, descrevendo cada módulo do GistSumm e exemplificando-os.

[3] http://www.nilc.icmc.usp.br/

[4] No léxico do NILC, as várias acepções de uma mesma palavra são organizadas de acordo com sua freqüência de uso.

3. A sumarização no GistSumm

Para ilustrar o funcionamento do sistema, será usado o texto-exemplo abaixo, o qual foi extraído do corpus do NILC (Kuhn et al., 2000).

AMAR FAZ BEM AO CORAÇÃO, AVISAM CARDIOLOGISTAS

Os apaixonados receberam hoje uma boa notícia dos cardiologistas: o amor faz bem ao coração.

Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.

"Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.

A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade – três fatores de risco associados às doenças do coração.

"Uma em cada três mortes no mundo ocorrem devido a problemas no coração e derrame, seis vezes superior do que as mortes associadas à Aids", afirmou o professor Philip Poole-Wilson, cardiologista do Imperial College, em Londres, e presidente da federação.

"É por essa razão que estamos ressaltando a importância de adotar um estilo de vida saudável e o impacto positivo que o amor pode ter para a saúde".

De acordo com a WHF, muitos estudos publicados demonstraram que fatores psicológicos, assim como os físicos, estão envolvidos com a doença cardíaca. Em uma pesquisa de cinco anos, 10 mil homens com risco elevado de desenvolver angina (dor no peito) foram questionados se a mulher com quem estavam demonstrava seu amor por eles. Aqueles que responderam "sim" tinham a metade do risco de apresentar a condição.

3.1. Segmentação sentencial

O processo de segmentação sentencial delimita todas as sentenças do texto-fonte observando os sinais de pontuação tradicionais[5], isto é, ponto final, de exclamação e de interrogação, conforme exemplificado a seguir para o texto-exemplo (com sentenças delimitadas por colchetes e numeradas para referência):

[Os apaixonados receberam hoje uma boa notícia dos cardiologistas: o amor faz bem ao coração.]$_1$

[Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês)

[5] É importante notar que o método de segmentação sentencial do GistSumm não é válido para línguas que não seguem a forma de pontuação mencionada.

divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.]$_2$

["Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.]$_3$

[A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade – três fatores de risco associados às doenças do coração.]$_4$

["Uma em cada três mortes no mundo ocorrem devido a problemas no coração e derrame, seis vezes superior do que as mortes associadas à Aids", afirmou o professor Philip Poole-Wilson, cardiologista do Imperial College, em Londres, e presidente da federação.]$_5$

["É por essa razão que estamos ressaltando a importância de adotar um estilo de vida saudável e o impacto positivo que o amor pode ter para a saúde".]$_6$

[De acordo com a WHF, muitos estudos publicados demonstraram que fatores psicológicos, assim como os físicos, estão envolvidos com a doença cardíaca.]$_7$ [Em uma pesquisa de cinco anos, 10 mil homens com risco elevado de desenvolver angina (dor no peito) foram questionados se a mulher com quem estavam demonstrava seu amor por eles.]$_8$ [Aqueles que responderam "sim" tinham a metade do risco de apresentar a condição.]$_9$

Problemas com a segmentação ocorrem quando o GistSumm se defronta com situações em que a pontuação não é usada na forma que o sistema é capaz de reconhecer. Por exemplo, quando se reproduz uma fala em um texto usando aspas, o ponto final normalmente ocorre antes da aspa final. Neste caso, o GistSumm não consegue reconhecer se aquele ponto realmente é o ponto final da sentença ou somente o ponto final da fala reproduzida. De forma padronizada, o GistSumm assume que o ponto seguido pela aspa não indica o final de sentença. Essa diferença entre ser o ponto final da sentença ou da fala reproduzida introduz diferenças no momento de ranquear as sentenças durante o processo de *ranqueamento de sentenças* do GistSumm, podendo produzir sumários diferentes no que diz respeito às sentenças selecionadas para formar o sumário ou à própria idéia principal do texto-fonte. Por exemplo, considerando a sentença 6 na forma como ela se apresenta no texto-exemplo (ponto após aspa), o seguinte sumário é produzido pelo GistSumm:

Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.

"Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.

Caso o ponto viesse antes da aspa, o seguinte sumário seria produzido:

"Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.

"É por essa razão que estamos ressaltando a importância de adotar um estilo de vida saudável e o impacto positivo que o amor pode ter para a saúde." De acordo com a WHF, muitos estudos publicados demonstraram que fatores psicológicos, assim como os físicos, estão envolvidos com a doença cardíaca.

No primeiro caso, as sentenças 2 e 3 foram selecionadas para formar o sumário. No seguinte, as sentenças 3, 6 e 7 foram selecionadas. Neste último caso, com o ponto antes da aspa, as sentenças 6 e 7 foram consideradas como sendo uma única sentença pelo GistSumm, recebendo uma pontuação diferente durante o processo de *ranqueamento de sentenças*, alterando, portanto, o *ranque de sentenças* e o próprio sumário produzido.

A subseção seguinte descreve o processo de ranqueamento de sentenças, no qual fica evidente que a segmentação errada de um texto pode produzir resultados diferentes.

3.2. Ranqueamento de sentenças

O processo de ranqueamento de sentenças, no Gistsumm, consiste em pontuar as sentenças delimitadas pelo segmentador sentencial. O processo de pontuação de cada sentença ocorre em várias etapas, conforme mostrado abaixo.

CRIAÇÃO DE VETORES DAS SENTENÇAS

Inicialmente, as sentenças são convertidas em vetores de palavras. Por exemplo, a seguir é mostrada a sentença 9 do texto-exemplo e seu respectivo vetor.

Aqueles que responderam "sim" tinham a metade do risco de apresentar a condição.

Aqueles	que	responderam	sim	tinham	a	metade	do	risco	de	apresentar	a	condição

Neste vetor, cada palavra da sentença é armazenada em uma posição diferente do vetor, mas mantêm a ordem em que estão na sentença.

CASE FOLDING, TROCA POR CANÔNICAS E REMOÇÃO DE *STOPWORDS*

A seguir, para aprimorar os resultados do processo de sumarização, o GistSumm adota os processos delineados por Witten et al. (1994), a saber: *case folding*, *stemming* e remoção de *stopwords*. O processo de *case folding* transforma todas as letras das palavras em letras minúsculas, para uniformização; o processo de *stemming*, que consiste na determinação da raiz de uma palavra, é substituído no GistSumm pelo uso de canônicas (já armazenadas no léxico simplificado); por fim, a remoção de *stopwords* permite um cálculo mais preciso da relevância das sentenças, pois remove palavras consideradas irrelevantes para a sumarização. Cada um desses processos é exemplificado a seguir.

Case Folding: a palavra “Aqueles” é trocada por “aqueles”

aqueles	que	responderam	sim	tinham	a	metade	do	risco	de	apresentar	a	condição

Troca por Canônicas: primeiramente, as palavras são substituídas por suas canônicas; a seguir, canônicas iguais são unificadas em uma única posição do vetor (caso da palavra

"a", com canônica "o"); como resultado, valores são incluídos em cada posição do vetor para indicar o número de vezes que cada canônica ocorreu na sentença

aquele	que	responder	sim	ter	o	metade	do	risco	de	apresentar	condição
1	1	1	1	1	2	1	1	1	1	1	1

Remoção de *Stopwords*: as palavras que são *stopwords* são desconsideradas no vetor anulando-se seu número de ocorrências na sentença (caso das palavras "aquele", "que", "o", "do" e "de")

aquele	que	responder	sim	ter	o	metade	do	risco	de	apresentar	condição
0	0	1	1	1	0	1	0	1	0	1	1

PONTUAÇÃO DAS SENTENÇAS

A pontuação das sentenças pode ocorrer pelo uso de um de dois métodos: palavras-chave (Black and Johnson, 1988) ou TF-ISF (*Term Frequency-Inverse Sentence Frequency*) (Larocca Neto et al., 2000). O método das palavras-chave parte do pressuposto de que a idéia principal de um texto pode ser expressa por algumas palavras-chave. O método TF-ISF[6], por sua vez, determina a importância das sentenças de um texto, de forma que a sentença mais importante será aquela que melhor o representa.

O método das palavras-chave foi adotado por ser um método clássico de sumarização; o método TF-ISF, por sua vez, foi adotado por representar uma abordagem recente e interessante para a sumarização. No GistSumm, os sumários produzidos por esses métodos, em geral, são diferentes, pois eles pontuam de forma diversa as sentenças. Pelo método das palavras-chave, cada vetor recebe como pontuação a soma do número de ocorrências de cada uma de suas palavras no texto inteiro (ou seja, em todos os vetores). No vetor acima, como o número de palavras com canônica *responder* é 1 no texto todo, *sim* é 1, *ter* é 4, *metade* é 1, *risco* é 3, *apresentar* é 1 e *condição* é 1, a pontuação da sentença do exemplo é de 1+1+4+1+3+1+1 = 12 pontos. Pelo método TF-ISF, a pontuação do vetor corresponde à média da pontuação de cada uma de suas palavras. A pontuação de cada palavra, por sua vez, se dá da seguinte forma:

Pontuação da palavra W = (nro. vezes que W ocorreu na sentença) x log((nro. palavras da sentença) / (nro. sentenças em que W ocorreu))

Para a sentença do exemplo, a pontuação obtida é 0,549.

Por qualquer um dos métodos, palavras-chave ou TF-ISF, a sentença com maior pontuação é considerada como sendo a sentença-gist do texto-fonte. Por isso, no GistSumm, os métodos de ranqueamento são, na realidade, métodos de *determinação da idéia principal*, sendo somente esta sua função no processo de sumarização.

A partir da sentença-gist, o GistSumm seleciona as sentenças que formarão o sumário, conforme relatado na subseção seguinte.

[6] O método TF-ISF é uma variação do método TF-IDF (*Text Frequency-Inverse Document Frequency*) (Salton, 1988) usado na área de Recuperação da Informação.

3.3. Seleção de sentenças

No processo de seleção de sentenças do texto-fonte para formar o sumário, o GistSumm executa os seguintes passos:

1) calcula a média da pontuação das sentenças do texto-fonte e assume essa média como sendo a *baseline*[7] para corte das possíveis sentenças que formarão o sumário;
2) para formar o sumário, o GistSumm seleciona, juntamente com a sentença-gist, todas as sentenças do texto-fonte que:
   - contenham pelo menos uma palavra que tenha uma das canônicas da sentença-gist;
   - possuam uma pontuação maior que a *baseline* calculada no passo 1.

Como será mostrado na próxima seção, um fator também determinante para as sentenças que comporão o sumário é a taxa de compressão[8] selecionada pelo usuário. Com uma taxa de compressão irrestrita, o passo 2 é realizado da forma como foi delineado acima; com uma taxa de compressão especificada, o passo 2 incorpora a restrição de selecionar tantas sentenças quanto possível com as maiores pontuações, desde que o tamanho desejado do sumário seja respeitado.

Para o texto-exemplo, com o método das palavras-chave para ranqueamento das sentenças e taxa de compressão irrestrita, a Tabela 1 mostra a pontuação das sentenças, com a *baseline* calculada em 34 ((21+50+57+49+37+30+26+28+12)/9=34).

Tabela 1 – Pontuação das sentenças

Sentença	Pontuação
1	21
2	50
3	57
4	49
5	37
6	30
7	26
8	28
9	12

Desta forma, a sentença-gist é a sentença 3:

> ["Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.] $_3$

Em seguida, todas as sentenças que possuam pelo menos uma palavra com canônica igual a uma das canônicas da sentença 3 e que tenham pontuação maior que 34 serão selecionadas para formar o sumário. Assim, são escolhidas as sentenças 2, 4 e 5:

> [Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em

[7] Da forma como é calculada, a *baseline* elimina sentenças candidatas para formar o sumário que não sejam muito importantes ou representativas no texto.

[8] A taxa de compressão é calculada como 1 – (tamanho do sumário / tamanho do texto-fonte).

muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.]$_2$

[A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade – três fatores de risco associados às doenças do coração.]$_4$

["Uma em cada três mortes no mundo ocorrem devido a problemas no coração e derrame, seis vezes superior do que as mortes associadas à Aids", afirmou o professor Philip Poole-Wilson, cardiologista do Imperial College, em Londres, e presidente da federação.]$_5$

Portanto, o sumário será formado pelas sentenças 2, 3, 4 e 5, nesta ordem, conforme mostrado abaixo:

Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.

"Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.

A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade - três fatores de risco associados às doenças do coração.

"Uma em cada três mortes no mundo ocorrem devido a problemas no coração e derrame, seis vezes superior do que as mortes associadas à Aids", afirmou o professor Philip Poole-Wilson, cardiologista do Imperial College, em Londres, e presidente da federação.

"É por essa razão que estamos ressaltando a importância de adotar um estilo de vida saudável e o impacto positivo que o amor pode ter para a saúde".

O sumário acima possui uma taxa de compressão de 39%. Em média, selecionando-se a taxa de compressão *irrestrita* no GistSumm, o sumário resultante possui uma taxa de compressão de 40%.

A próxima seção descreve a interface para operação do GistSumm.

4. Interface do GistSumm

Ao ser executado, o GistSumm exibe uma tela com duas janelas filhas: uma para o texto-fonte (*Source text*) e outra para o sumário gerado automaticamente (*Summary*), conforme mostra a Figura 2.

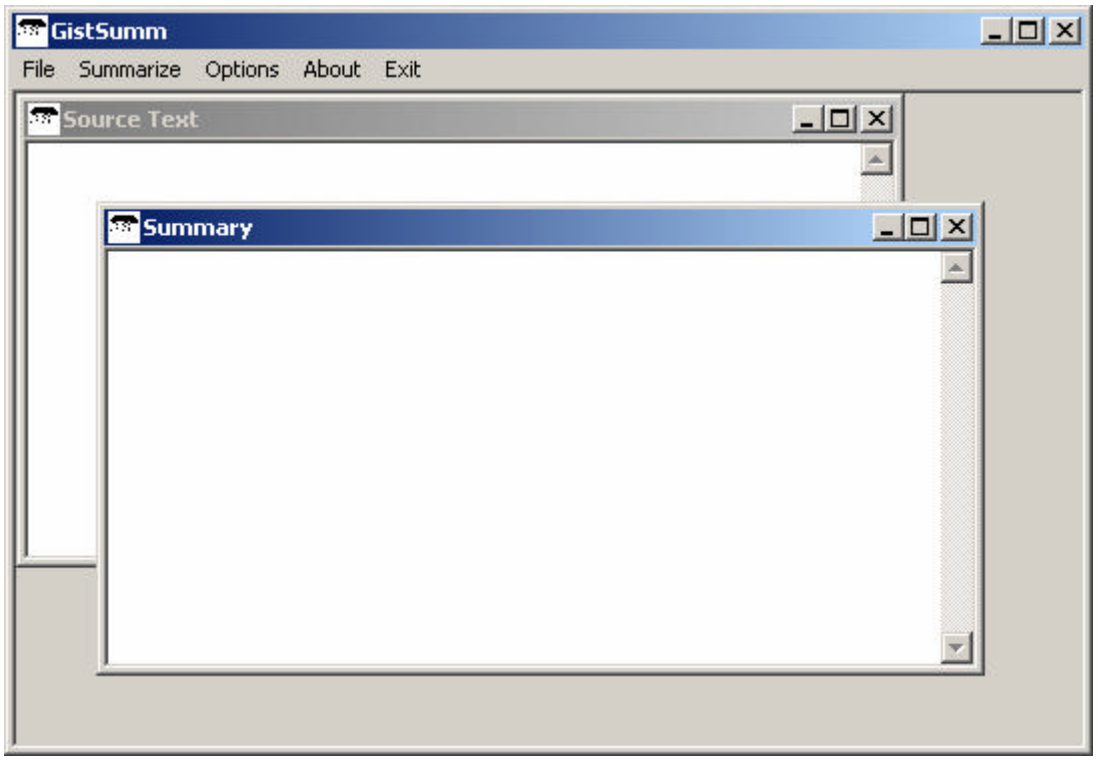


Figura 2 – Tela inicial do GistSumm

Para sumarizar um texto, o usuário deve, primeiramente, selecionar o menu *File>Open source text*, como mostra a Figura 3. Ao fazer isso, uma janela de diálogo irá aparecer para que o arquivo com o texto a ser sumarizado seja selecionado. Esse arquivo deve conter o texto-fonte não formatado, isto é, deve ser um arquivo com extensão *txt* (no sistema operacional Microsoft Windows). A Figura 4 mostra o GistSumm com o texto-exemplo ilustrado anteriormente, selecionado para sumarização automática. Pode-se notar que a legenda da janela do texto-fonte indica o caminho e o nome do arquivo carregado.

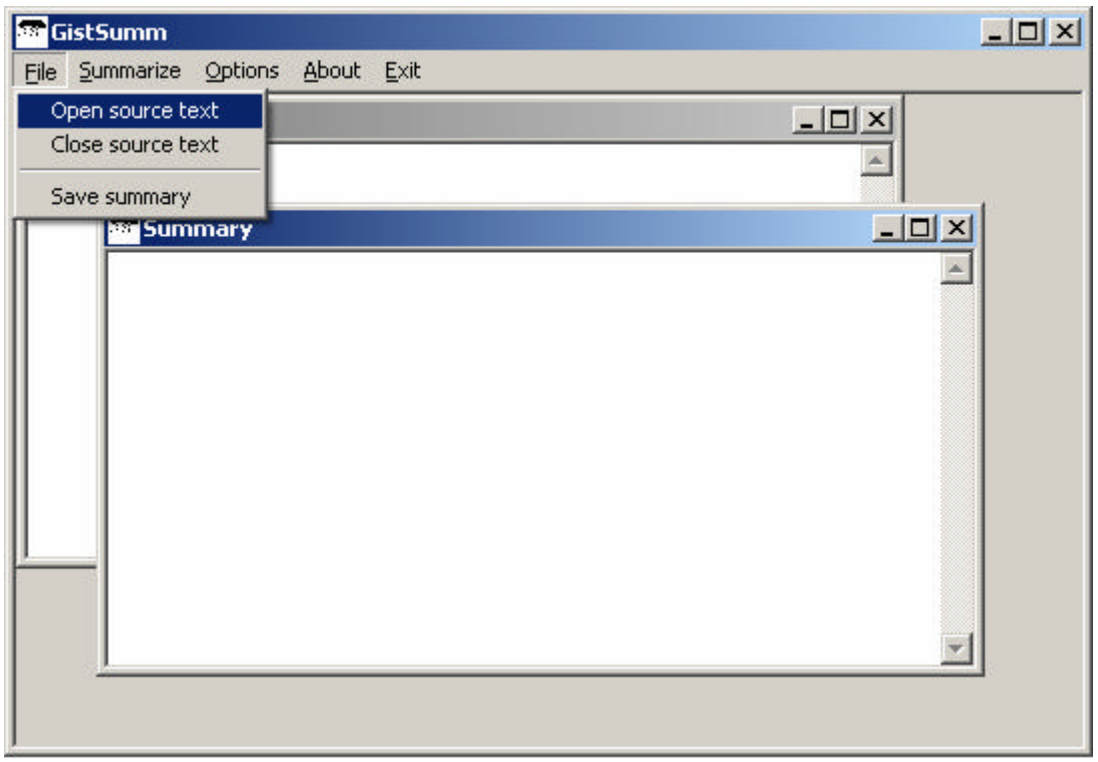


Figura 3 – Abertura do texto-fonte

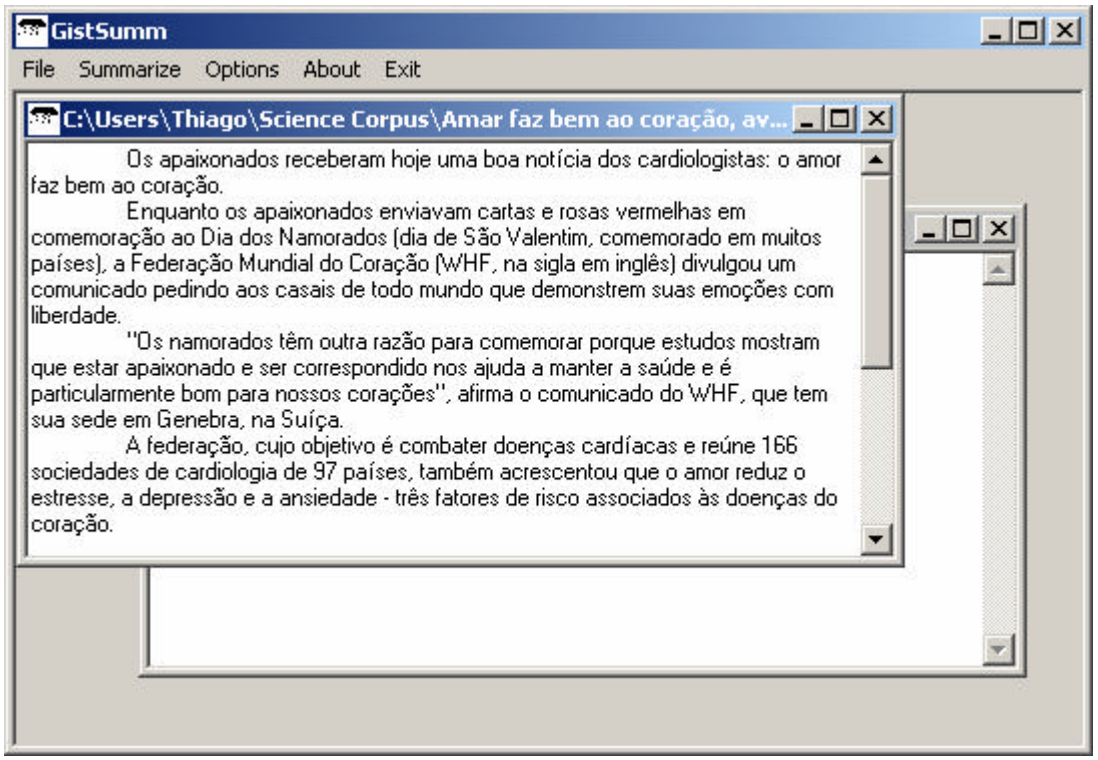


Figura 4 – Texto-fonte a sumarizar

Após carregar o texto-fonte, o usuário poderá selecionar o menu *Options* e escolher o método de ranqueamento, chamado, na interface, de método de determinação da idéia principal (*Gist determination method*) (isto é, ou o *Keywords* ou o TF-ISF), e a taxa de compressão (que pode ser irrestrita – *Best* – ou uma taxa especificada pelo usuário). As Figuras 5 e 6 mostram, respectivamente, essas duas etapas de escolha do usuário para a

sumarização. Caso o usuário não personalize as opções de sumarização, opções padrões serão selecionadas: o método de ranqueamento *Keywords* e a taxa de compressão *Best*.

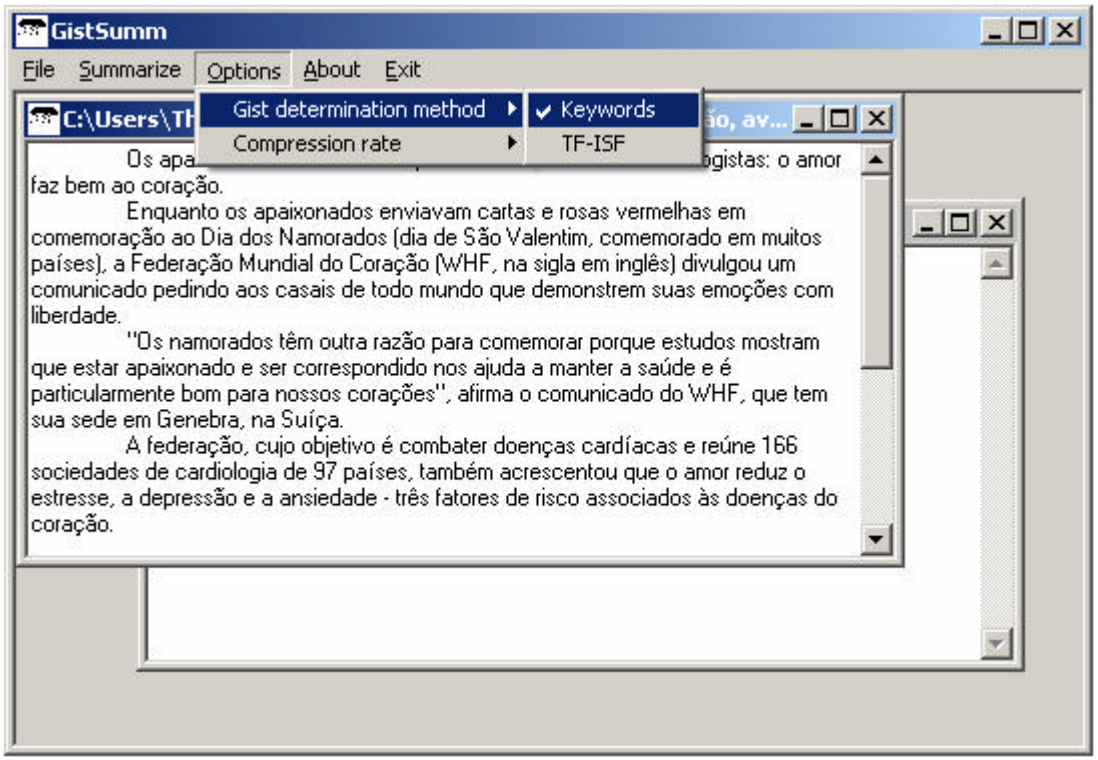


Figura 5 – Seleção do método de determinação da idéia principal

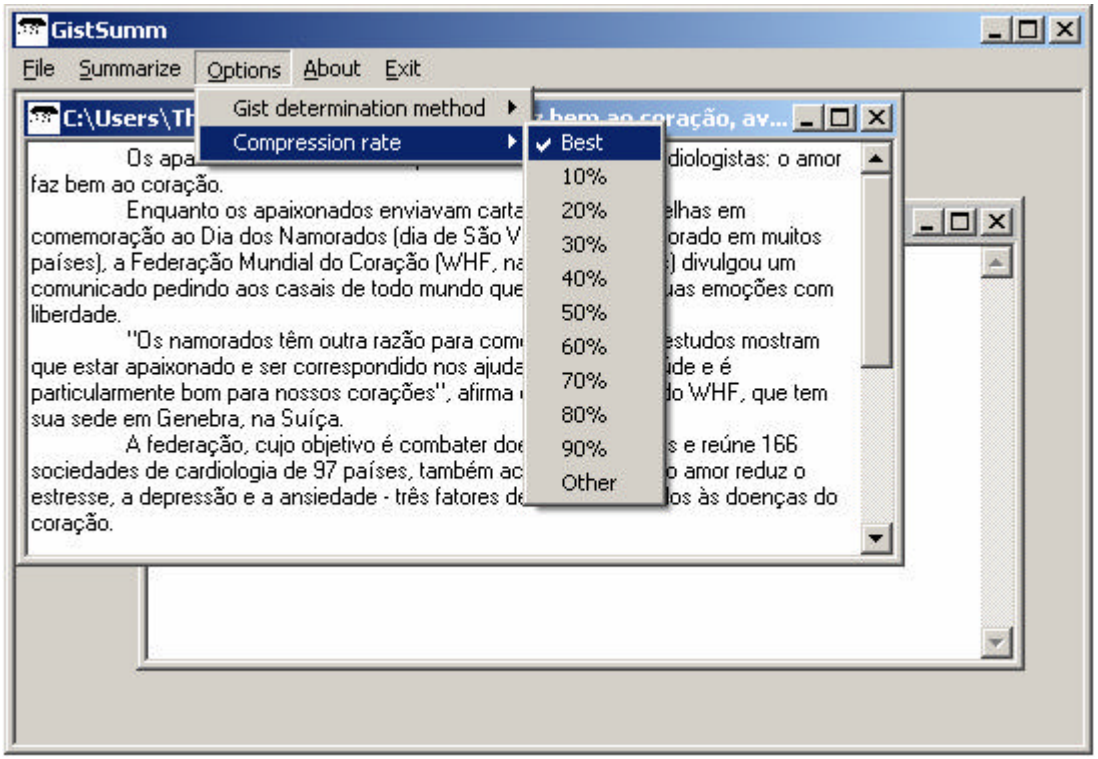


Figura 6 – Seleção da taxa de compressão

Por meio de testes, verificou-se que a opção *Best* produz, normalmente, sumários com taxa de compressão de 40%, ou seja, o sumário é condensado em 40% em

relação ao texto-fonte, ficando, portanto, com o tamanho equivalente a 60% do texto-fonte.

Para sumarizar o texto-fonte, basta agora clicar no menu *Summarize*, conforme mostrado na Figura 7. O sumário gerado aparecerá, então, na janela *Summary*, como é mostrado na Figura 8. Depois de gerado, ele pode ser salvo selecionando-se o menu *File>Save source text*, como mostra a Figura 9.

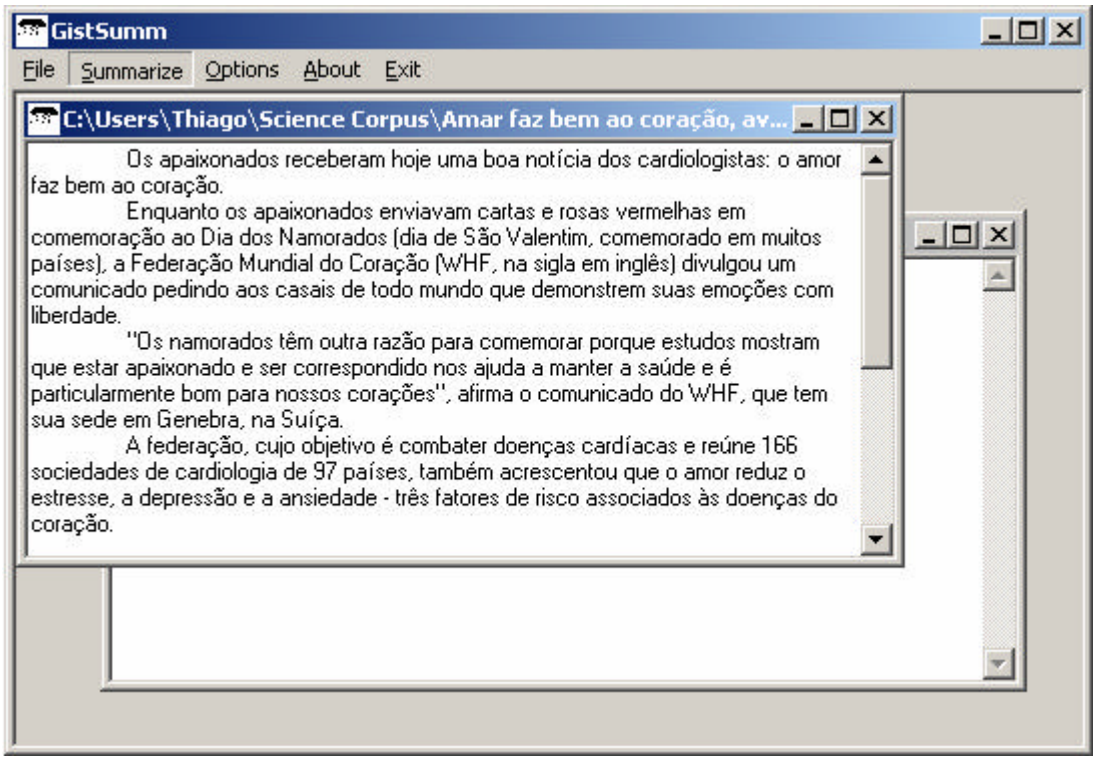


Figura 7 – Sumarizando o texto

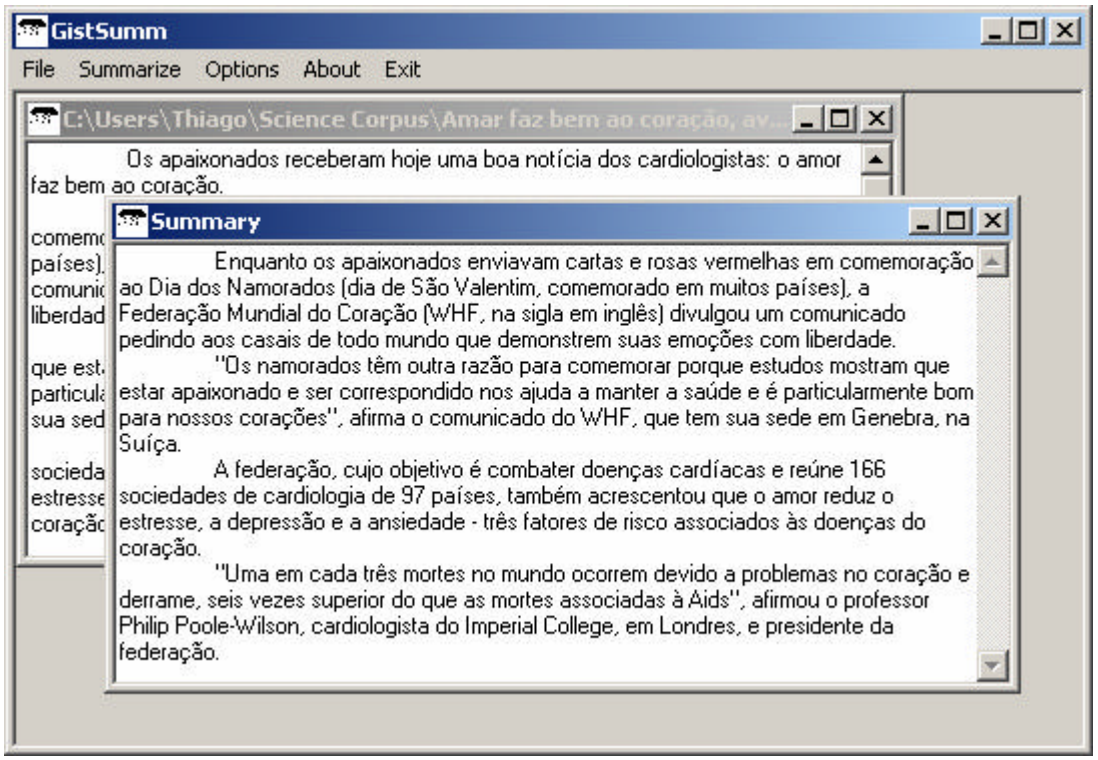


Figura 8 – Sumário gerado com método de ranqueamento *Keywords* e taxa de compressão *Best*

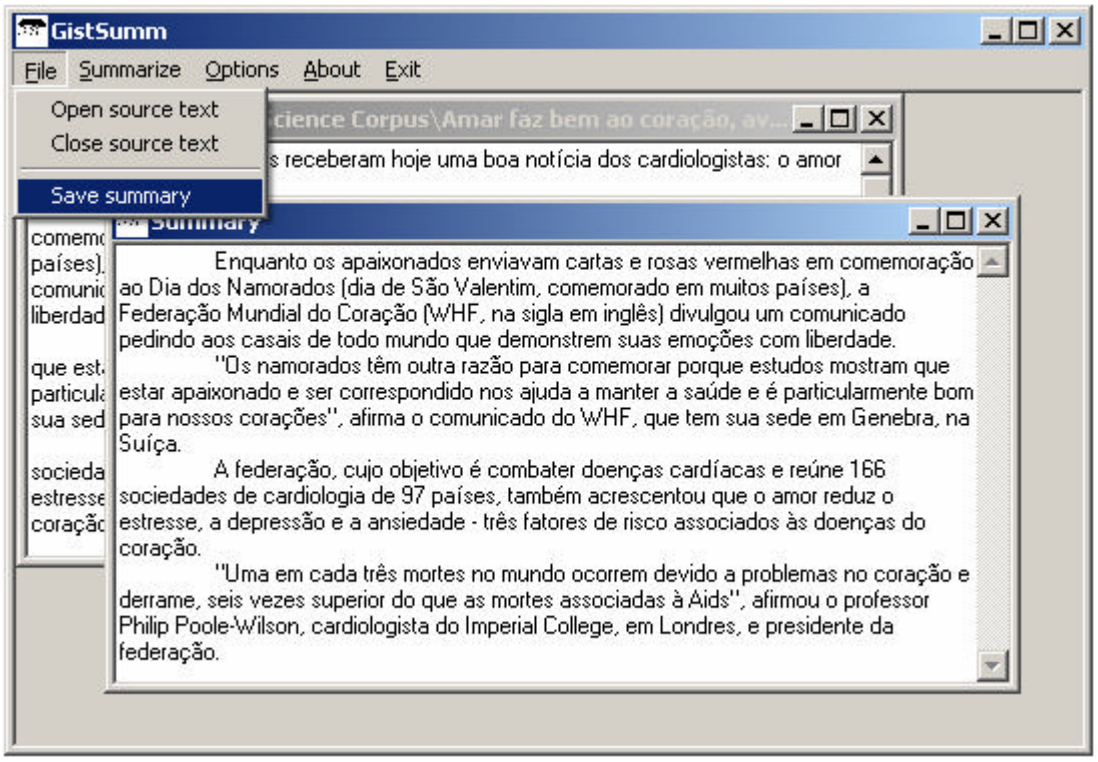


Figura 9 – Salvando o sumário gerado automaticamente

Vários sumários para o mesmo texto podem ser gerados, bastando que o usuário indique configurações distintas. Neste caso, basta repetir os passos anteriores. Desse modo, será possível, depois, recuperar dos arquivos todos os sumários de um mesmo

texto, para compará-los e, por exemplo, avaliar o desempenho do GistSumm ou escolher o melhor deles para outro uso qualquer.

As Figuras 10, 11 e 12 mostram outros sumários gerados para o texto ilustrado com as opções (1) método de ranqueamento *Keywords* e taxa de compressão 60%, (2) método de ranqueamento TF-ISF e taxa de compressão *Best* e (3) método de ranqueamento TF-ISF e taxa de compressão 60%. Pode-se notar que todos os sumários são informativos e inteligíveis. Entretanto, pode-se perceber também que os dois últimos (gerados pelo método TF-ISF) são iguais, apesar das taxas de compressão diferenciadas, e não mantiveram a idéia principal do texto-fonte, enfocando outras informações que não são tão relevantes.

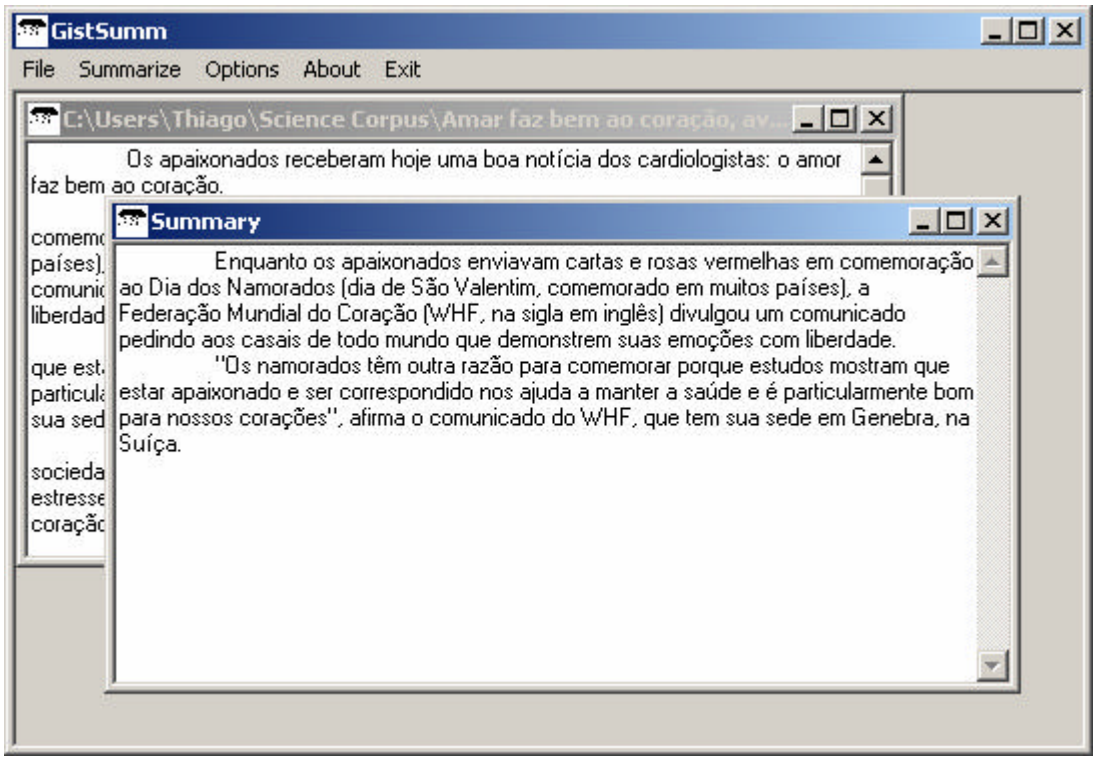


Figura 10 – Sumário gerado com método de ranqueamento *Keywords* e taxa de compressão de 60%

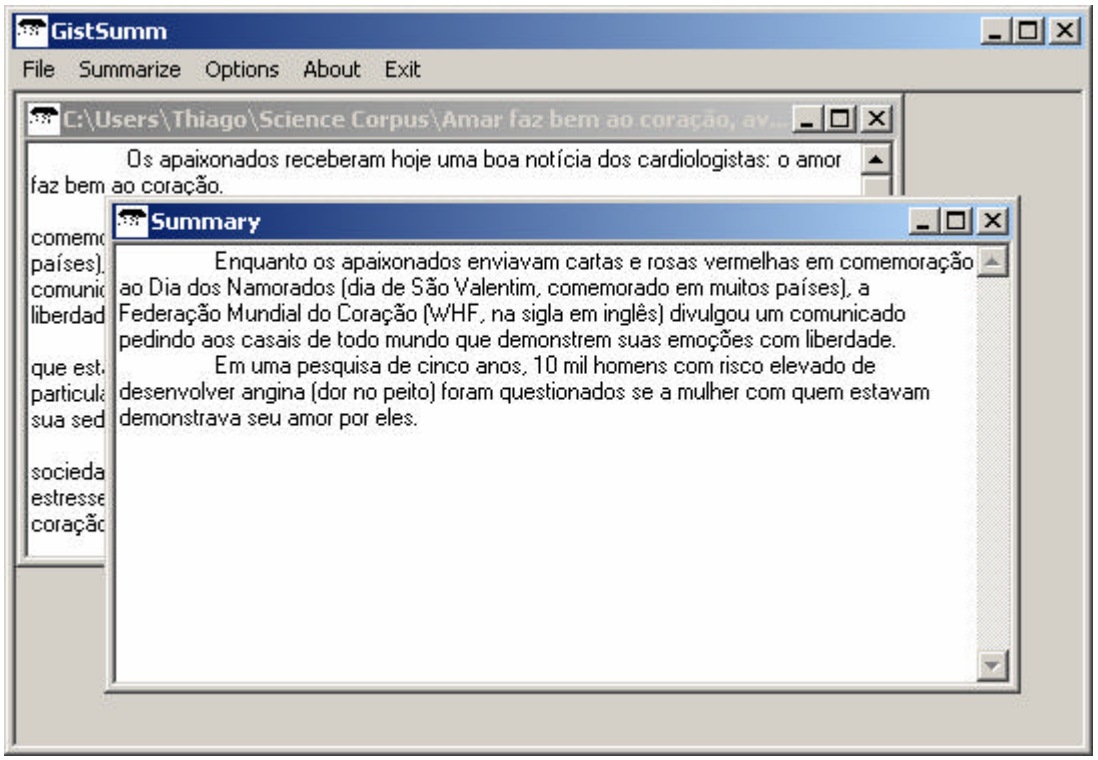


Figura 11 – Sumário gerado com método de ranqueamento TF-ISF e taxa de compressão *Best*

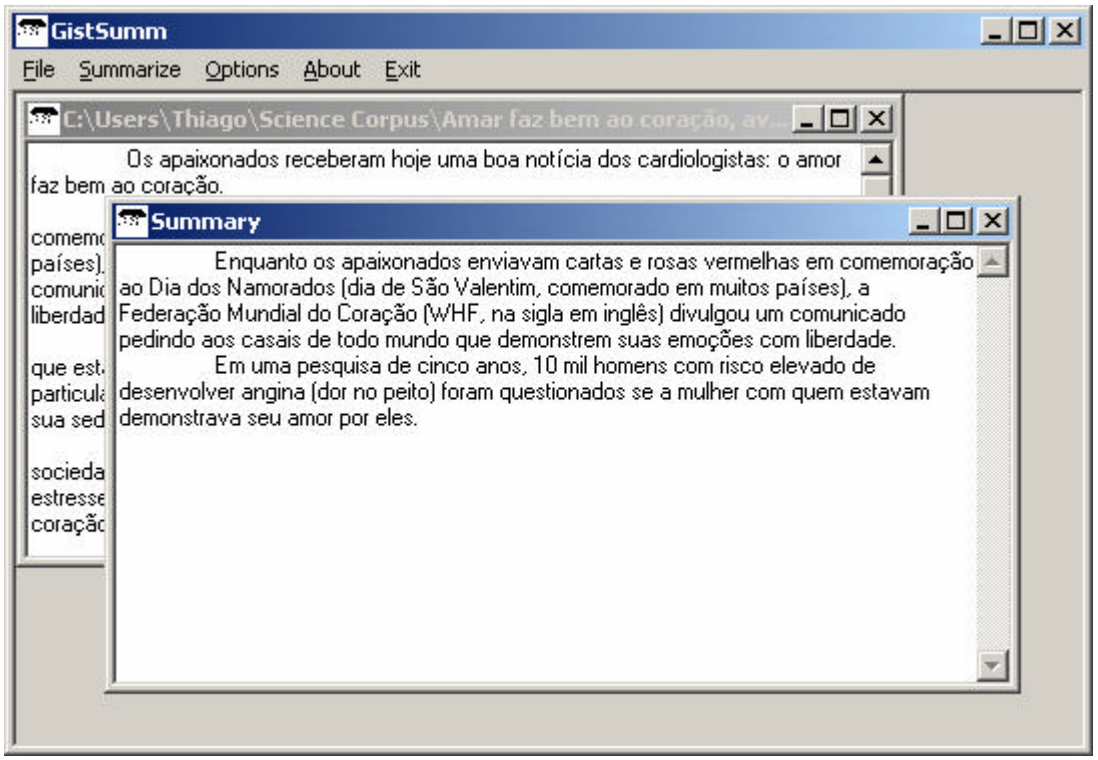


Figura 12 – Sumário gerado com método de ranqueamento TF-ISF e taxa de compressão de 60%

Durante a geração do sumário, a legenda *GistSumm* da janela principal da interface é substituída por *Wait...* para indicar que o sumário está em processo de

geração. Ao final da geração, uma mensagem *Done!* é exibida para o usuário, indicando que a geração do sumário terminou.

Para sair da interface, basta clicar no menu *Exit*.

A próxima seção apresenta algumas considerações finais para futuros aprimoramentos do GistSumm.

5. Considerações finais

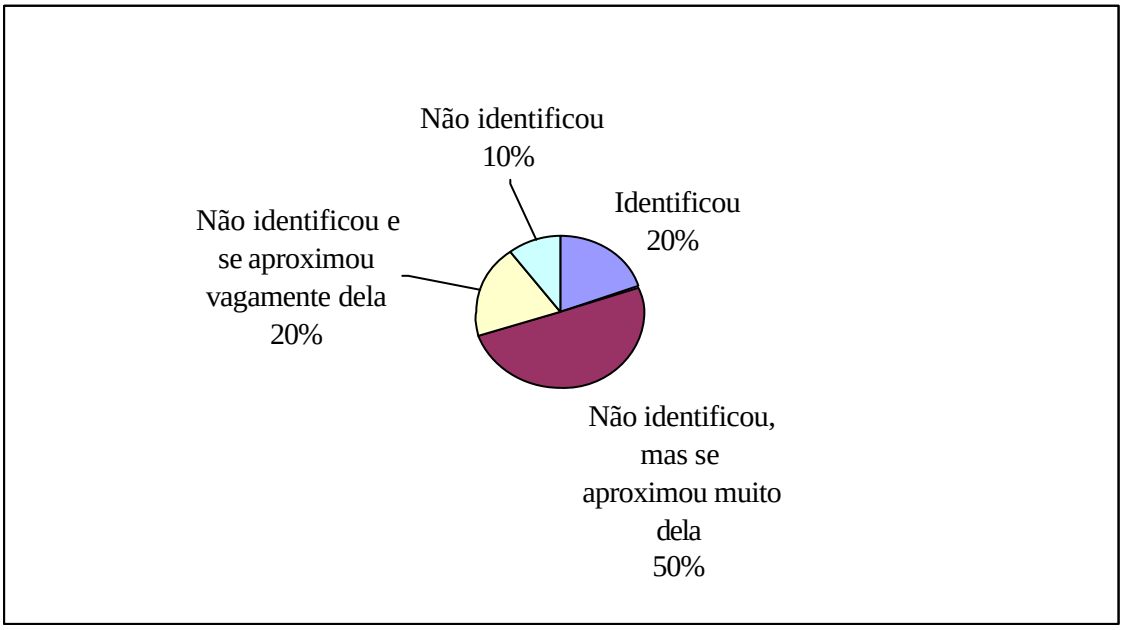
O GistSumm constitui uma aproximação computacional para a forma como as pessoas sumarizam textos. Primeiramente, seleciona-se a sentença que melhor representa a idéia principal de um texto, a sentença-gist, para, a seguir, selecionar outras sentenças cuja informação complementa a idéia principal do texto.

Para identificar a sentença-gist, dois métodos são usados: *Keywords* e TF-ISF. As sentenças são pontuadas e a sentença com maior pontuação é considerada como sendo a sentença-gist. A partir dessa sentença, busca-se outras sentenças do texto-fonte para formar o sumário, sendo que estas outras sentenças devem ter algumas restrições verificadas para assegurar sua relevância textual para formar o sumário.

Foi realizado um experimento para verificar se o GistSumm captura de forma efetiva a idéia principal dos textos-fonte nos sumários que produz. Para tanto, utilizou-se o *Theses Corpus*, o qual é um recorte do *CorpusDT* (Feltrim et al., 2000) anotado com as sentenças-gist por juízes humanos. O *Theses Corpus* contém 10 introduções de textos científicos com uma média de 530 palavras cada.

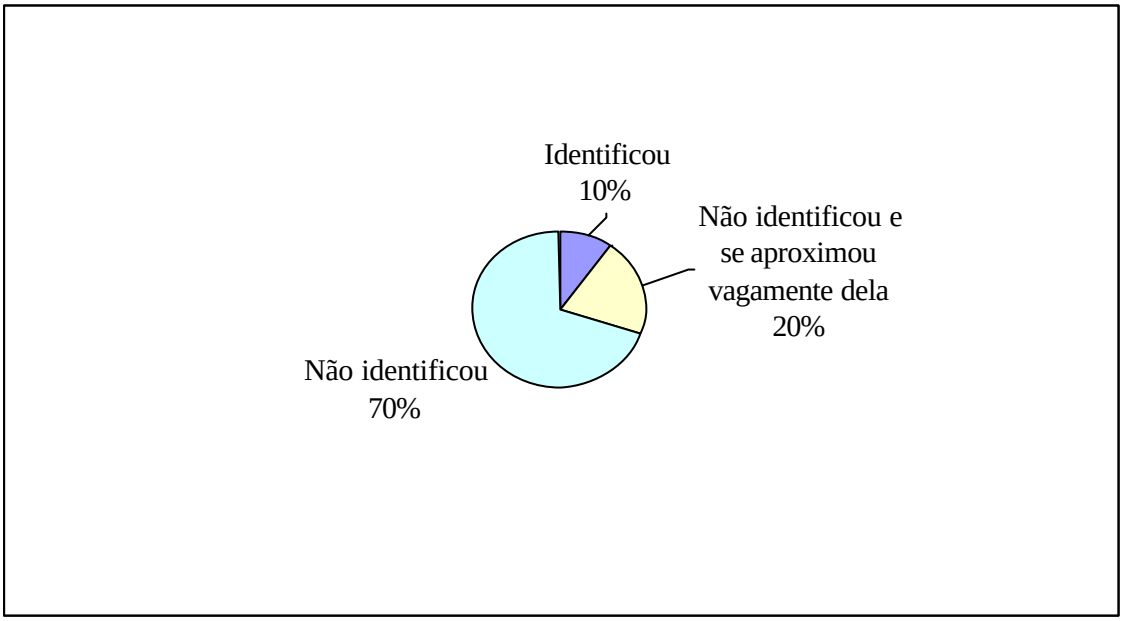
O Gráfico 1 mostra o quão efetivo o GistSumm foi em capturar a idéia principal dos textos-fonte nos sumários utilizando o método das palavras-chave, com taxa de compressão *Best*. Em 20% dos casos, a sentença-gist foi identificada como tal; em 50% dos casos, ela ficou muito próxima de ser a sentença com maior pontuação e, assim, foi selecionada para o sumário; em 20% dos casos, ela ficou distante da sentença com maior pontuação, mas foi selecionada para o sumário; em 10% dos casos, ela ficou muito distante da sentença com maior pontuação e não foi selecionada para o sumário.

Gráfico 1 – Identificação da idéia principal usando o método das palavras-chave



O Gráfico 2 mostra o quão efetivo o GistSumm foi em capturar a idéia principal dos textos-fonte nos sumários utilizando, agora, o método TF-ISF, com taxa de compressão *Best*. Em 10% dos casos, a sentença-gist foi identificada como tal; em 20% dos casos, ela ficou distante da sentença com maior pontuação, mas foi selecionada para o sumário; em 70% dos casos, ela ficou muito distante da sentença com maior pontuação e não foi selecionada para o sumário.

Gráfico 2 – Identificação da idéia principal usando o método TF-ISF



É importante notar que quando a sentença-gist (aquela indicada pelos juízes humanos para os textos do corpus usado) não é a sentença à qual se associa a maior pontuação no GistSumm, ela pode ou não ser inserida no sumário, dependendo da taxa de compressão.

O Gráfico 3 sintetiza os dois gráficos anteriores, permitindo uma melhor comparação entre os resultados.

Gráfico 3 – Síntese dos experimentos

Método	Identificou sentença-gist		Proximidade com sentença-gist		
	Sim	Não	Não	Vaga	Muito
Keywords	20%				
		80%	10%	20%	50%
TF-ISF	10%				
		90%	70%	20%	0

Pela análise dos resultados, pode-se notar que o método das palavras-chave é uma melhor aproximação para identificar a idéia principal do que o método TF-ISF. Além disso, em grande parte dos testes, o método TF-ISF produziu sumários muito curtos e sem coerência. Entretanto, uma avaliação mais sistemática ainda precisa ser realizada.

Há vários trabalhos sobre a identificação da idéia principal de um texto e a sumarização automática deste. Particularmente, na linha superficial, além dos já mencionados, destacam-se os seguintes:

- Luhn (1958), Edmundson (1969) e Lin e Hovy (1997): dependendo da posição de uma sentença, ela pode carregar informação importante do texto;
- Baxendale (1958), Paice (1981) e Black e Johnson (1988): sentenças que contenham palavras/termos importantes podem indicar que seu conteúdo é relevante.

O GistSumm, como mostrado neste relatório, faz uso dos métodos de Black e Johson e de Larocca Neto et al. para identificar a sentença-gist. Outros métodos, entretanto, poderiam ser usados.

A seleção das sentenças da forma como é feita no GistSumm, usando uma *baseline* para diminuir o número de sentenças candidatas para formar o sumário, equivale ao que é feito por Larocca Neto em seu trabalho, sendo que a *baseline* do GistSumm corresponde ao que ele chama de *threshold*. Outros critérios também poderiam ser adotados para selecionar as sentenças candidatas ao sumário, por exemplo (Kupiec et al., 1995):

- tamanho da sentença: sentenças curtas tendem a não ser incluídas no sumário;
- termos indicativos: por exemplo, em textos científicos, sentenças que não contenham termos como "este relatório", "em conclusão" e "resultados" podem não ser tão relevantes para estarem no sumário;
- nomes próprios: dar alta prioridade para sentenças que contenham nomes próprios, pois estes são normalmente importantes.

Dentre os aprimoramentos possíveis para o GistSumm, destacam-se:

- o acoplamento de um *tagger* ao sistema, de forma que seja possível determinar de forma mais precisa as canônicas corretas das palavras do texto;
- o desenvolvimento de um *stemmer* para o português do Brasil, permitindo, assim, um cálculo mais apurado para o ranqueamento das sentenças;
- o desenvolvimento de um segmentador textual mais preciso e inteligente que possa não só delimitar de forma mais precisa as sentenças de um texto, mas também as orações intra-sentenciais;
- o desenvolvimento de um processo de resolução anafórica que permita resolver as dependências contextuais das sentenças do sumário para melhorar o nível de coerência dos sumários produzidos.

Ainda, como extensão do sistema, é importante ressaltar a possibilidade de personalização do GistSumm para outras línguas naturais, o que pode ser feito pelo simples acoplamento de novos repositórios de dados da língua desejada, isto é, de um léxico e de uma *stoplist* específicos. Caso o GistSumm venha a ser usado em domínios específicos, a *stoplist* também pode ser substituída por uma outra específica do domínio.

Quanto às limitações do sistema, a principal delas é a adoção de uma única sentença como representante da idéia principal de um texto. Isso pode ser um problema para o processo de sumarização na forma como ele é feito no GistSumm, problema este já apontado por Pardo (2002) e Pardo e Rino (2002) em suas avaliações do sumarizador DMSumm (*Discourse Modeling Summarizer*), o qual também assume tal restrição em seu modelo. Na avaliação sistemática a ser realizada, tal fator deve ser investigado, determinando-se o quanto isso influencia na informatividade e qualidade dos sumários.

Apêndice A – Stoplist

A *stoplist* completa utilizada pelo GistSumm é mostrada na Tabela 1, sendo composta por artigos, pronomes, preposições, conjunções e interjeições do português do Brasil, contendo, no total, 196 palavras retiradas, principalmente, de Cunha e Cintra (2001).

Tabela 1 – *Stoplist*

-	cujo	me	os	sua
a	cujos	mesmos	ou	suas
à	da	meu	outra	tanta
ah	das	meus	outras	tantas
ai	de	mim	outrem	tanto
algo	dela	minha	outro	tantos
alguém	delas	minhas	outros	te
algum	dele	muita	para	teu
alguma	deles	muitas	per	teus
algumas	desde	muito	perante	ti
alguns	do	muitos	pois	toda
alô	dos	na	por	todas
ambos	e	nada	porém	todo
ante	eia	nas	porque	todos
ao	ela	nela	portanto	trás
após	elas	nelas	pouca	tu
aquela	ele	nele	poucas	tua
aquelas	eles	neles	pouco	tuas
aquele	em	nem	poucos	tudo
aqueles	embora	nenhum	próprios	ué
aquilo	enquanto	nenhuma	psit	uh
as	entre	nenhumas	psiu	ui
até	essa	nenhuns	quais	um
bis	essas	ninguém	quaisquer	uma
cada	esse	no	qual	umas
certa	esses	nos	qualquer	uns
certas	esta	nós	quando	vária
certo	estas	nossa	quanta	várias
certos	este	nossas	quantas	vário
chi	estes	nosso	quanto	vários
com	eu	nossos	quantos	você
comigo	hem	o	que	vós
conforme	hum	ó	quem	vossa
conosco	ih	ô	se	vossas
consigo	isso	oba	sem	vosso
contigo	isto	oh	seu	vossos
contra	lhe	olá	seus	
convosco	lhes	onde	si	
cuja	logo	opa	sob	
cujas	mas	ora	sobre	

Referências

Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, Vol. 2, pp. 354-361.

Black, W.J. and Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, No. 3. Department of Computation. University of Manchester Institute of Science and Technology.

Cunha, C. e Cintra, L.F.L. (2001). *Nova Gramática do Português Contemporâneo*. 3ª. edição. Editora Nova Fronteira.

Edmundson, H.P. (1969). New methods in automatic extracting. *Journal of the ACM*, Vol. 16, pp. 264-285.

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC. NILC-TR-01-4.

Kuhn, D.; Abarca, E.; Nunes, M.G.V. (2000). *Corpus NILC - Situação em Maio/2000*. (NILC-TR-00-7).

Kupiec, J.; Pedersen, J.; Chen, F. (1995). A trainable document summarizer. *ACM SIGIR*, pp. 68-73.

Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. In the *Proceedings of the International Joint Conference IBERAMIA/SBIA*, Atibaia, SP.

Larocca Neto, J. (2002). *Contribuição ao Estudo de Técnicas para Sumarização Automática de Textos.* Dissertação de Mestrado. Pontifícia Universidade Católica do Paraná, Brasil.

Lin, C. and Hovy, E.H. (1997). Identifying Topics by Position. In the *Proceedings of the Applied Natural Language Processing Conference*, pp. 283-290.

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

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Mani, I. and Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

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

Nunes, M.G.V.; Vieira, F.M.C.; Zavaglia, C.; Sossolote, C.R.; Hernandez, J. (1996). *A construção de um léxico para suporte à revisão automática do português do Brasil.* Relatório Técnico do ICMC-USP, No. 42. São Carlos.

Paice, C.D. (1981). The automatic generation of literature abstracts: na approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co. (Publishers).

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos – SP.

Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

Rino, L.H.M. (1996). Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos. Tese de Doutorado. IFSC-USP. São Carlos – SP.

Salton, G. (1988). *Automatic Text Processing*. Reading, MA: Addison-Wesley.

Witten, I.H.; Moffat, A.; Bell, T.C. (1994). Managing Gigabytes. *Van Nostrand Reinhold*. New York.