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

TeMário: Um Corpus para Sumarização Automática de Textos

Thiago Alexandre Salgueiro Pardo
Lucia Helena Machado Rino

NILC-TR-03-09

Outubro 2003

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

Resumo

Este relatório descreve o TeMário, um corpus voltado para a Sumarização Automática de Textos. Desenvolvido para vários fins de sumarização, como análise lingüística, treinamento de sumarizadores automáticos e sua avaliação posterior, o TeMário é composto, basicamente, por textos jornalísticos e seus sumários em português. Os sumários foram construídos por um sumarizador profissional, professor e consultor de editoração de textos em português. O TeMário será usado, primeiramente, para investigações específicas de métodos de Sumarização Automática no Projeto EXPLOSA[1].

[1] Desenvolvido com apoio da FAPESP (Proc. Nro. 01/08849-8).

ÍNDICE

1. Introdução

Este relatório descreve o TeMário (sigla de '**TE**xtos com su**MÁRIO**s'), um corpus construído com vistas à Sumarização Automática de Textos, no âmbito do Projeto EXPLOSA[2] (**EXPLO**ração de métodos diversos para a **S**umarização **A**utomática). Esse corpus é composto, basicamente, de textos jornalísticos e de seus respectivos sumários manuais, construídos por um professor e consultor de editoração de textos em português[3]. Além de servir a diversos fins de sumarização automática, como, por exemplo, à análise lingüística de textos e sumários, à construção e treinamento de sumarizadores automáticos e à avaliação desses sistemas, ele também servirá a outras tarefas relacionadas, cujas áreas atuais de interesse envolvem a Recuperação de Informação e a Detecção de Tópicos.

No Projeto EXPLOSA, em particular, há vários sistemas que podem se beneficiar desse corpus para treinamento ou avaliação, como o GistSumm (Pardo et al., 2003a), o NeuralSumm (Pardo et al., 2003b), o DMSumm[4] (Pardo, 2002), o SuPor (Módolo, 2003) e o UNLSumm (Martins, 2002). Além desses sistemas, cujas informações genéricas podem ser encontradas no site do NILC (http://www.nilc.icmc.usp.br/), outras atividades podem ser elaboradas, que façam uso do TeMário. Por exemplo, estudos sobre a forma como o sumarizador profissional reconhece as informações relevantes de um texto, para compor seus sumários, ou a identificação de parâmetros de indicação dos critérios de sumarização, para a modelagem de sistemas computacionais. Detalhes sobre essas tarefas e sua relação com a Sumarização Automática podem ser encontrados, inicialmente, em Rino e Pardo (2003) e Martins et al. (2001).

Além das tarefas diretamente relacionadas à Sumarização Automática, há, atualmente, outros projetos no NILC que podem fazer uso do TeMário, como o Projeto LACIO-WEB, de construção de recursos para investigações variadas, incluindo a própria Recuperação de Informação, assim como a etiquetagem de textos e/ou informações em português. Em contexto mais amplo, o TeMário será parte da Linguateca[5], grande repositório internacional de recursos, dados e informações para o processamento automático da língua portuguesa.

---

[2] http://www.dc.ufscar.br/~lucia/PROJECTS/EXPLOSA.htm (FAPESP, Proc. Nro. 01/08849-8).

[3] Sumários manuais são usados, aqui, para indicar os sumários construídos por um profissional em escrita. No inglês, a denominação de sumários profissionais ou sumários humanos também é utilizada por alguns autores.

[4] Todos esses disponíveis para download em http://www.nilc.icmc.usp.br/~thiago/.

[5] http://www.linguateca.pt/

2. O TeMário

2.1. Características gerais

O nome "TeMário" para o corpus em foco foi escolhido por duas razões: por remeter aos objetos que o compõem – textos e sumários – e por sugerir a palavra *tema* em seu nome, cujo reconhecimento é imprescindível na tarefa de sumarização.

Para construir o TeMário, foram coletados 100 textos jornalísticos, totalizando 61.412 palavras. 60 textos constam do jornal on-line Folha de São Paulo (doravante, identificada pela sigla **FSP**) e estão distribuídos igualmente nas seções Especial, Mundo e Opinião; os 40 textos restantes foram publicados no Jornal do Brasil (doravante, identificado pela sigla **JB**), também on-line, e estão também uniformemente distribuídos nas seções Internacional e Política. A Tabela 1 sintetiza esses dados, mostrando também o número de palavras por seção e o número médio de palavras por texto de cada seção. Segundo a tabela, a média de palavras por seção é de 12.282 palavras e a média de palavras por texto é de 613 palavras, sendo esta correspondente a textos variando de 1 a 2 ½ páginas.

Tabela 1 – Características do corpus de textos-fonte

Jornais	**Seções**	**Número de textos**	**Número de palavras**	**Média de palavras/texto**
Folha de São Paulo	Especial	20	12.340	617
	Mundo	20	13.739	686
	Opinião	20	10.438	521
Jornal do Brasil	Internacional	20	12.098	604
	Política	20	12.797	639
	Total	**100**	**61.412**	
	Média		**12.282**	**613**

Foram escolhidos textos jornalísticos para compor o corpus pelo fato de apresentarem uma linguagem voltada a uma grande audiência de leitores abrangente e, assim, um português de abrangência média, quer em termos de vocabulário, quer em termos de construções gramaticais. Desse modo, foram excluídos automaticamente da seleção os suplementos como o Mais ou o Jornal de Resenhas, da FSP, por exemplo, que se destinam a um público mais literato. Essa limitação visa, sobretudo, a facilitar as tarefas relacionadas à Sumarização Automática: é usual recorrer-se a mão-de-obra especializada para elaborar avaliações dos resultados automáticos. Um estilo mais rebuscado imporia maior dificuldade de leitura, compreensão e avaliação, levando a resultados duvidosos sobre o foco real da tarefa.

Essa razão se evidencia também pelo fato de o gênero jornalístico ser, atualmente, o mais utilizado em avaliações em larga escala, na Sumarização Automática: os concursos internacionais de avaliação de sumarizadores automáticos, como a SUMMAC[6] (*text **SUMMA**rization evaluation **C**onference*) e a DUC[7] (***D**ocument **U**nderstanding **C**onference*), têm utilizado textos jornalísticos que remontam a grandes volumes de dados. A última DUC, por exemplo, disponibilizou 900 textos jornalísticos em inglês, de diversas fontes, para tarefas de avaliação que envolveram grandes comitês de juizes humanos.

[6] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/
[7] http://www-nlpir.nist.gov/projects/duc/

Para a produção do TeMário, uma vez coletados os textos jornalísticos procedeu-se à construção dos sumários correspondentes, razão pela qual os textos são denominados *textos-fonte.*

2.2. Construção dos sumários

Os textos coletados foram enviados ao professor e consultor de editoração de textos em português para a execução de duas tarefas: a construção dos sumários correspondentes (Tarefa 1, principal) e a indicação, para cada texto-fonte, de sua idéia principal (Tarefa 2). Desse modo, na Tarefa 1 esse professor assumiu a posição de sumarizador profissional, devendo produzir sumários informativos. Na Tarefa 2, ele assumiu a posição de mero leitor dos textos, apreendendo o que eles apresentam de mais importante. Neste caso, foi solicitado que ele simplesmente grifasse as sentenças (nos sumários) que lhe indicassem a idéia principal.

Relacionando ambas as tarefas, a apreensão da idéia principal de um texto constitui, certamente, condição essencial para a produção de bons sumários informativos. Ou seja, ao identificar sentenças que remetem à idéia principal, estas serão as que direcionarão o sumarizador em sua construção dos sumários correspondentes, já que estes devem apresentar toda (ou parte significativa) informação principal do texto-fonte, podendo, inclusive, substituí-lo (condição principal dos sumários informativos). Considera-se, aqui, a alternância entre as funções do sumarizador, de leitor a escritor, na tarefa comumente conhecida como tarefa de *reescrita* do texto-fonte, em forma condensada (Mani, 2001).

Além da necessidade de produzir sumários informativos, o sumarizador tinha uma restrição adicional: o tamanho de cada sumário deveria ser de, aproximadamente, 25-30% do tamanho de seu texto-fonte. Do ponto de vista da Sumarização Automática, isso é equivalente a fixarem-se as taxas de compressão dos textos-fonte ao intervalo de 70-75%, ou seja, 70 ou 75% do conteúdo desses textos devem ser *desconsiderados*, ao se elaborarem os sumários.

As instruções para o sumarizador profissional realizar ambas as tarefas, de sumarização e marcação de sentenças, encontram-se no Apêndice A.

2.3. Complementação do corpus

O TeMário, assim composto de 100 textos-fonte e respectivos sumários manuais, certamente constitui um repositório significativo (embora relativamente pequeno) de dados textuais para várias tarefas de Sumarização Automática, como o já citado treinamento de sistemas automáticos, personalizando-os para a sumarização de textos reais de mesmo gênero e domínio. No entanto, para tarefas de avaliação ele nem sempre supre a devida demanda: considerando que sumários manuais são resultado de um processo de reescrita do conteúdo do texto-fonte que o escritor achou mais relevante, utilizar sumários manuais como sumários "ideais", a fim de compará-los com os sumários gerados automaticamente, não é tarefa fácil: dificilmente eles terão essa correspondência explícita. Por esse motivo, avaliações entre sumários manuais e automáticos, em geral, dependem do juízo humano e, assim, de experimentos de avaliação caros e complexos, em sua implementação. Para minimizar esse problema, é comum utilizarem-se *extratos* "ideais", em vez de sumários ideais, para a comparação com resultados automáticos, sobretudo quando se estiver tratando da Sumarização Automática extrativa.

Neste caso, a própria terminologia indica que tanto extratos ideais quanto extratos resultantes do processo de sumarização, propriamente dito, são derivados de uma metodologia extrativa, qual seja, a pura transposição de segmentos textuais selecionados para o texto condensado, cuja principal característica é reproduzir literalmente partes do texto-fonte. Assim, torna-se possível considerar o simples casamento de padrões entre os extratos ideais e seus correspondentes extratos sob avaliação, para determinar se estes são bons representantes da idéia principal do texto-fonte, assim como o são os extratos ideais. Claramente, essa etapa pode ser realizada de forma automática, na maioria das vezes, diferentemente da forma de avaliação anterior, com ganhos consideráveis em termos de custo e complexidade. Por essa razão, complementou-se o TeMário com extratos ideais produzidos com base nos sumários manuais, por um gerador de extratos ideais[8].

O gerador de extratos ideais identifica e justapõe as sentenças dos textos-fonte que apresentam o mesmo conteúdo das sentenças dos sumários manuais correspondentes. Para isso, utiliza a medida do co-seno de Salton (1989), segundo a metodologia descrita por Rino e Pardo (2003). É importante dizer que os extratos ideais podem não ser, de fato, ideais no sentido de espelharem de forma completa e totalmente satisfatória o conteúdo relevante do texto-fonte a ser sumarizado, como o faria um escritor humano: a medida do co-seno, por se basear puramente na co-ocorrência de palavras entre o sumário manual e o texto-fonte, pode produzir extratos com sentenças inapropriadas. Porém, esses extratos serão considerados ideais por serem os melhores possíveis, do ponto de vista de custo/benefício da produção automática.

A Tabela 2 correlaciona os tamanhos dos sumários manuais e extratos ideais. Pode-se notar que o número médio de palavras dos sumários manuais é significantemente menor do que o número médio de palavras dos extratos ideais. Essa diferença pode se dever ao fato de o sumarizador humano ser capaz de condensar o conteúdo que deseja da melhor forma possível, para satisfazer restrições de condensação, usando o processo de reescrita. No caso de extratos ideais, satisfazer essas restrições nem sempre é trivial, pois fixa-se previamente a unidade mínima a extrair dos textos-fonte – em geral, as sentenças são extraídas integralmente para compor os sumários. Por essa razão, é mais comum terem-se extratos maiores do que os sumários manuais.

Tabela 2 – Características dos sumários manuais e extratos ideais

		Sumários manuais		**Extratos ideais**	
Jornais	**Seções**	**Número de palavras**	**Média de palavras/seção**	**Número de palavras**	**Média de palavras/seção**
Folha de São Paulo	Especial	4.313	215	4.450	222
	Mundo	4.234	211	4.706	235
	Opinião	3.373	168	3.980	199
Jornal do Brasil	Internacional	3.734	186	5.676	283
	Política	3.791	189	4.451	222
	Total	**19.445**		**23.263**	
	Médias gerais	**3.889**	**193**	**4.652**	**232**

[8] Disponível para download em http://www.nilc.icmc.usp.br/~thiago.

2.4. Organização do TeMário

Considerando um ambiente hierárquico em que arquivos podem ser armazenados pelo uso do Microsoft Windows, o TeMário está organizado em uma única pasta, com duas subpastas que agregam, respectivamente, os textos-fonte e os sumários, como mostra a Figura 1.

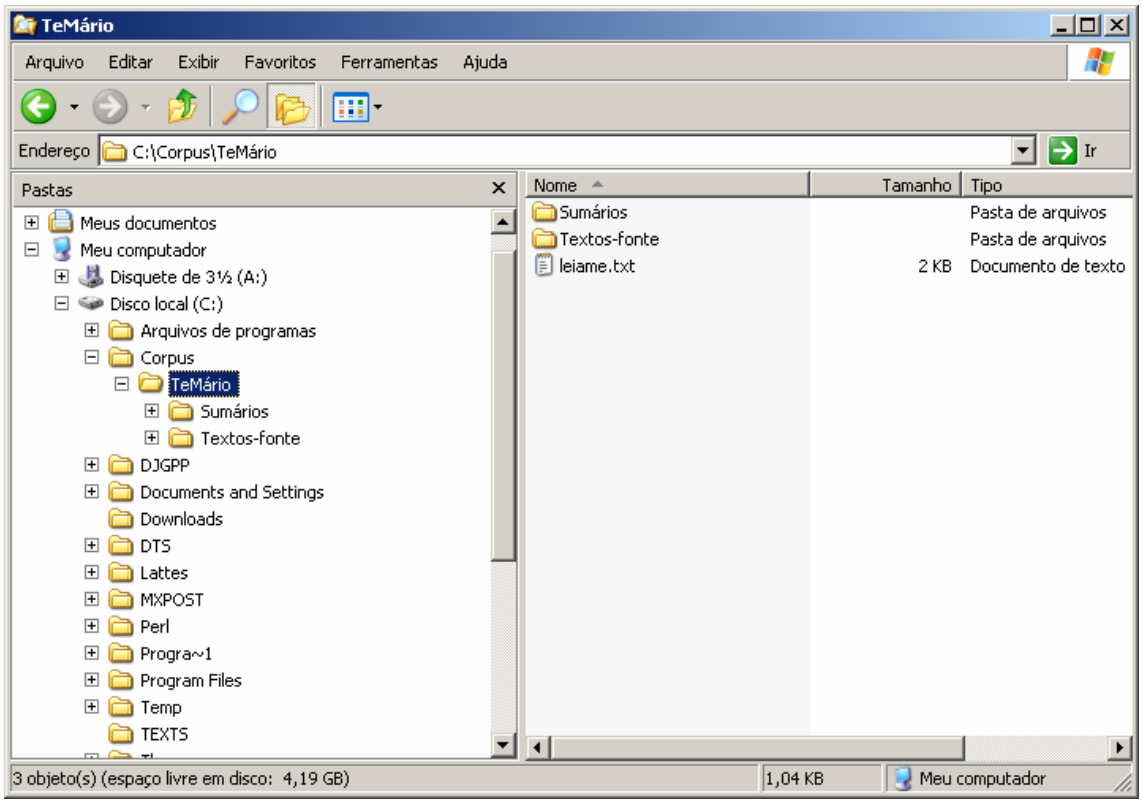

Figura 1 – Diretório "TeMário"

Na pasta de textos-fonte, há três pastas assim organizadas (Figura 2):

a) A primeira delas contém os textos-fonte com seus títulos, organizados por suas origens, ou seja, os textos estão agrupados por jornal (FSP ou JB, como mostra a Figura 3) e seção (Especial, Mundo e Opinião, para os textos da FSP, e Internacional e Política, para o JB), totalizando 60 textos da FSP e 40 do JB;
b) A segunda contém todos os textos-fonte com seus títulos, sem discriminação de origem;
c) A terceira contém os textos-fonte sem quaisquer informações de origem ou título.

Os arquivos textuais estão todos em formato txt, já adequado para o processamento automático. Com exceção de seus prefixos, todos os nomes de arquivo incluem o ano (NN), mês (AA) e dia de publicação (de 1 a 31)[9]. Os prefixos indicam as seções dos jornais correspondentes, como seguem:

- Textos-fonte da seção Especial da FSP têm o prefixo "ce" (de *Caderno Especial*);
- Textos-fonte da seção Mundo da FSP têm o prefixo "mu";

[9] NN para dois dígitos numéricos e AA para as duas primeiras letras do mês correspondente.

- Textos-fonte da seção Opinião da FSP têm o prefixo “op”;
- Textos-fonte da seção Internacional do JB têm o prefixo “in”;
- Textos-fonte da seção Política do JB têm o prefixo “po”.

Assim, por exemplo, o arquivo de nome ‘in96fe29-a.txt’ indica um texto da seção Internacional do JB, publicado no dia 29 de fevereiro de 1996; o arquivo ‘mu94ag07-b.txt’ indica um texto da seção Mundo da FSP, publicado no dia 07 de agosto de 1994.

Adicionalmente, os textos-fonte sem título têm seus arquivos denotados por “St-” antes dos prefixos acima descritos.

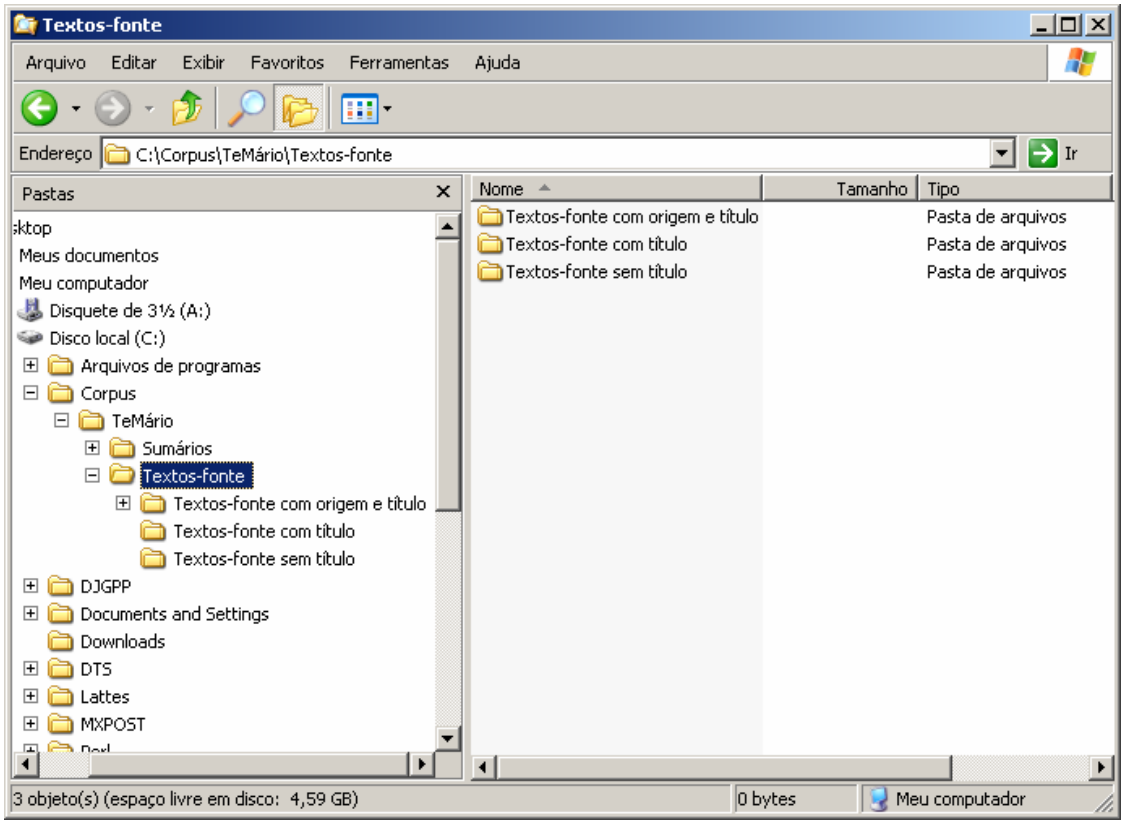

Figura 2 – Diretório “Textos-fonte”

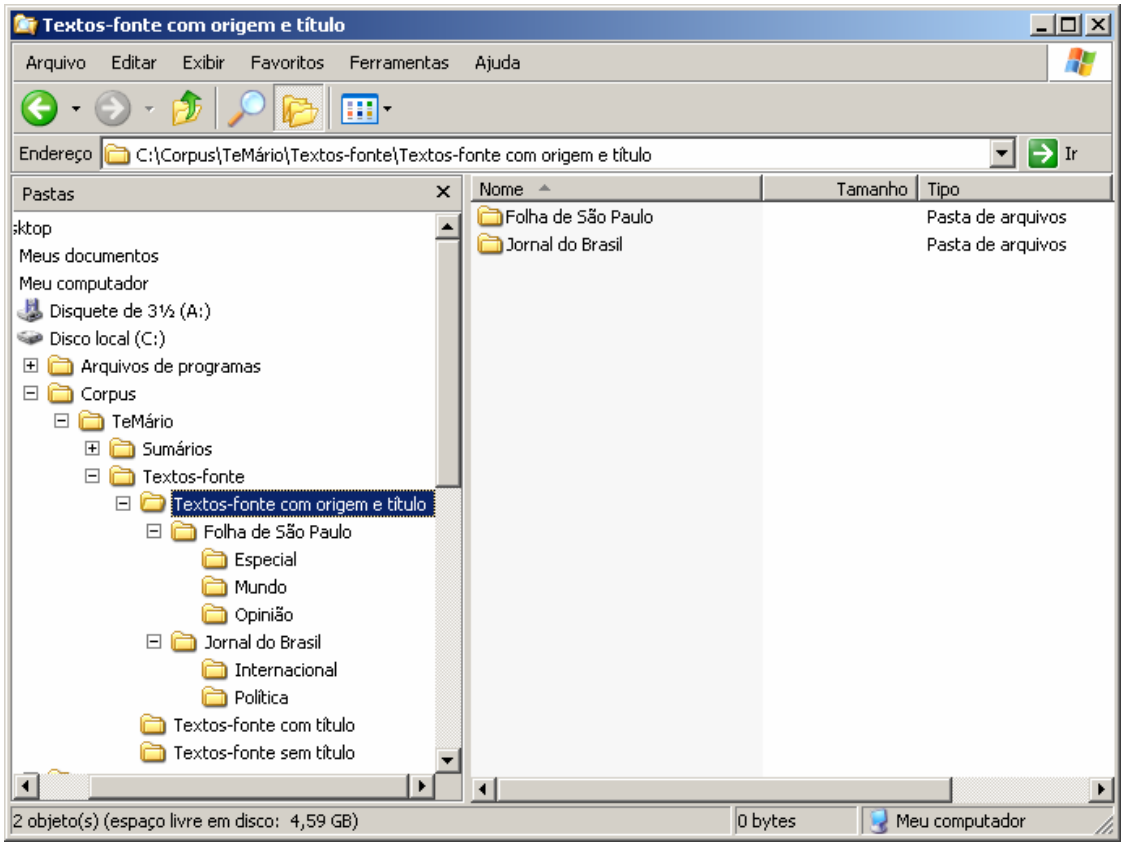

Figura 3 – Diretório de textos-fonte completos

Essa subdivisão dos textos-fonte em pastas de diversas naturezas tem o propósito de auxiliar a recuperação dos dados segundo interesses específicos. Por exemplo, se o objetivo for elaborar um processo de avaliação, os textos-fonte com título permitirão associarem-se sumários ou extratos gerados automaticamente com o próprio título, para comparação. Neste caso, pode-se considerar que o titulo é um representante legítimo da idéia principal do texto-fonte, por ter sido escolhido pelo próprio escritor (no caso, o jornalista), a fim de verificar se os resultados automáticos a preservam. Já para estudos lingüísticos, tais como verificação de características particulares de um gênero ou domínio, um analista pode visar à recuperação dos cadernos específicos, os quais já indicam uma classificação genérica. Adicionalmente, para o próprio processo de sumarização, manual ou automática, é desejável que os textos-fonte estejam visíveis sem quaisquer títulos. No caso manual, para que estes não induzam o escritor a privilegiar uma informação mais explicitamente relacionada ao titulo. No caso automático, a razão é outra: a sumarização de um texto não inclui o tratamento de seu titulo.

Na pasta de sumários (Figura 1), há também três subpastas (Figura 4): uma com os sumários manuais, outra com os sumários manuais marcados e, finalmente, outra com os extratos ideais produzidos pelo gerador de extratos ideais, conforme descrito anteriormente.

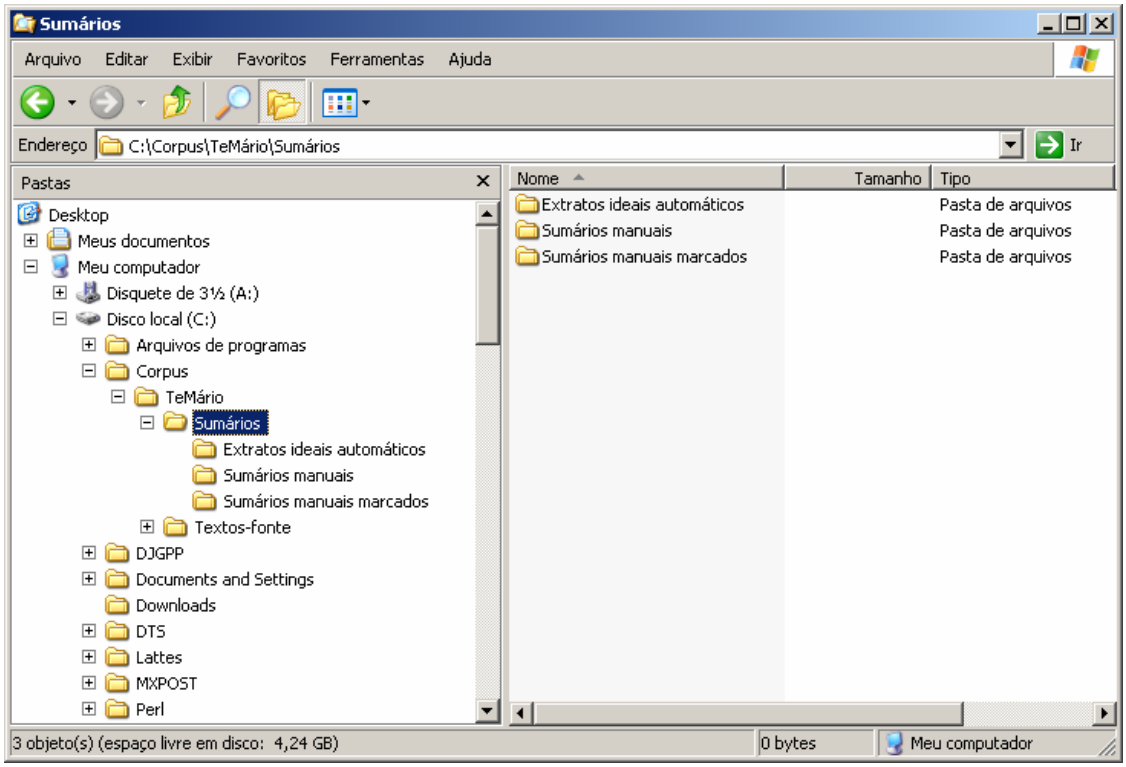

Figura 4 – Subpastas da pasta de "Sumários"

Na pasta de sumários manuais estão os arquivos em formato txt que contêm os sumários construídos pelo profissional. Seus nomes contêm exatamente os nomes dos arquivos dos textos-fonte correspondentes, acrescidos do prefixo "Sum-", para indicar o fato de se tratarem de sumários e não de textos inteiros. Na pasta de sumários manuais marcados estão os mesmos sumários, porém, agora com indicação, em vermelho, das sentenças que indicaram ao sumarizador profissional as idéias principais dos textos-fonte correspondentes (conforme Tarefa 2 solicitada ao sumarizador profissional – Apêndice A). Esses arquivos também são nomeados como os textos-fonte, porém com prefixo "Summ-" (para **S**umários **M**anualmente **M**arcados). Sua extensão .doc é justamente para que a formatação seja preservada e, assim, as sentenças marcadas não sejam alteradas. Portanto, recomenda-se que esses dados sejam sempre preservados.

Para diferenciar os sumários manuais (prefixo "Sum") e os sumários manualmente marcados (prefixo "Summ") dos extratos ideais, os arquivos da pasta 'Extratos ideais automáticos' possuem o prefixo "Ext-" (estes são também arquivos sem qualquer formatação, isto é, arquivos com extensão txt).

3. Considerações Finais

Como este relatório descreve, o TeMário é um corpus de 100 textos jornalísticos e seus correspondentes sumários manuais e extratos ideais. Os sumários manuais foram construídos por um profissional em escrita em português, enquanto os extratos ideais foram produzidos automaticamente, para fins de Sumarização Automática. Devido a essa natureza específica, o repositório de dados contém, ainda, os textos-fonte com seus títulos e os mesmos sumários manuais, agora marcados com as idéias principais dos textos-fonte que nortearam as decisões de sumarização humana. Enquanto a geração dessas informações adicionais não representou ônus significativo para o

especialista humano e a manutenção do titulo dos textos-fonte consistiu somente uma mera decisão de representação, do ponto de vista da Sumarização Automática elas constituem um rico repositório de dados, quer para estudos comparativos durante a avaliação de resultados automáticos, quer para a exploração de outras técnicas de Sumarização Automática. Por exemplo, os próprios títulos podem servir de base para a escolha de segmentos textuais relevantes para compor um sumário: um título pode ser considerado uma *gist sentence*, como o faz o GistSumm (Pardo et al., 2003a), neste caso.

O potencial de uso do TeMário pode ainda ser aumentado ao considerar-se uma marcação XML (e**X**tensible **M**arkup **L**anguage), conforme visa o Projeto LACIO-WEB[10]. Neste caso, os arquivos com extensão .doc (os que indicam as idéias que nortearam as decisões do sumarizador profissional) podem ser também convertidos para a notação XML, sem perda de significado, já que, nesta linguagem, as marcas de estilo são preservadas.

Referências Bibliográficas

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

Martins, C.B. (2002). *UNLSumm: Um Sumarizador Automático de Textos UNL*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

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.

Módolo, M. (2003). *SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português.* Dissertação de Mestrado. Departamento de Computação, UFSCar. São Carlos - SP.

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.; Rino, L.H.M.; Nunes, M.G.V. (2003a). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* - PROPOR, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. (2003b). NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. *Anais do IV Encontro Nacional de Inteligência Artificial*. Campinas-SP.

Rino, L.H.M. e Pardo, T.A.S. (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. *Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII*: *III Jornada de Minicursos de Inteligência Artificial (III MCIA),* pp. 203-245. Campinas-SP.

Salton, G. (1989) *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

[10] http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Apêndice A - Especificação da tarefa de sumarização manual

Tarefa 1 (principal): Sumarizar o corpus de 100 textos, segundo as seguintes características:

1. Tamanho dos sumários: aproximadamente 25-30% do texto-fonte
2. Sumários informativos

Tarefa 2 (adicional):

Indicar a(s) sentença(s) que lhe indicam a idéia principal de cada texto-fonte, i.e., aquela(s) que lhe dão a diretriz para a construção do(s) sumário(s).

A indicação pode ser feita nos próprios arquivos, de preferência, com tarjas coloridas.

Por favor, inclua seus resumos em arquivos txt, um para cada resumo.