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

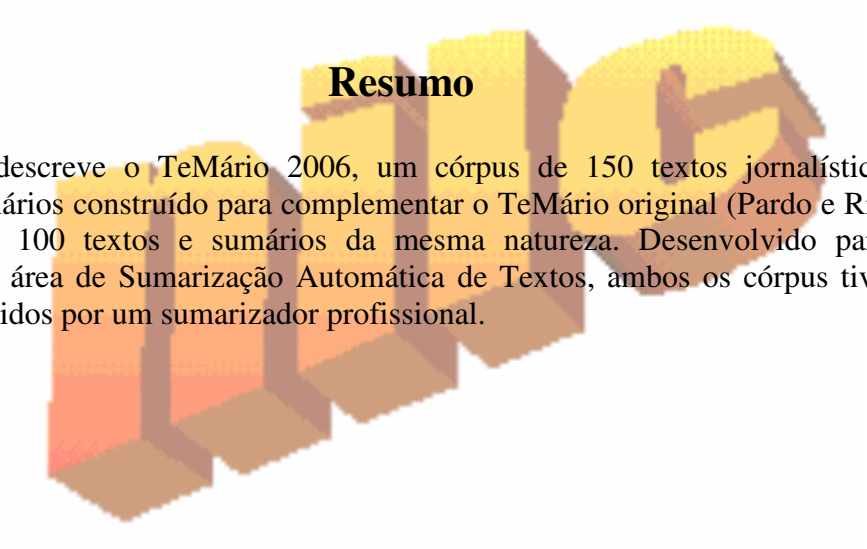
TeMário 2006: Estendendo o Córpus TeMário

Erick Galani Maziero
Vinícius Rodrigues de Uzêda
Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-07-06

Agosto, 2007

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 2006, um córpus de 150 textos jornalísticos e seus respectivos sumários construído para complementar o TeMário original (Pardo e Rino, 2003), o qual contém 100 textos e sumários da mesma natureza. Desenvolvido para fins de investigação na área de Sumarização Automática de Textos, ambos os córpus tiveram seus resumos produzidos por um sumarizador profissional.

ÍNDICE

1. Introdução

Este relatório descreve o córpus TeMário 2006 (sigla de "TExtos com suMÁRIOs"), um córpus criado para estender o córpus TeMário original (Pardo e Rino, 2003), construído com vistas a projetos de Sumarização Automática de Textos (doravante tratado como SA).

Esse córpus é composto, basicamente, de textos jornalísticos e de seus respectivos sumários manuais, construídos por um professor e consultor de editoração de textos em português. Voltado para SA, sendo utilizado no treinamento e na avaliação de sistemas dessa área, também pode ser aplicado a outras tarefas relacionadas, como, por exemplo, as áreas de Recuperação de Informação e de Detecção de Tópicos.

O acréscimo de mais 150 textos a este córpus tem, por finalidade, melhorar os resultados obtidos em suas utilizações. Com um maior conjunto de textos, os sistemas desenvolvidos que fazem uso de Aprendizado de Máquina terão mais bases para adquirir seus conhecimentos. Além disso, a avaliação de qualquer sistema de SA ficará mais elaborada com esta expansão do conjunto de testes. Outras atividades que façam uso do TeMário 2006 também podem ser elaboradas. Por exemplo, estudos teóricos sobre a forma como um sumarizador humano reconhece as informações de um texto que devem compor seus sumários, ou a identificação de algoritmos utilizados no processo de construção desses resumos, para uma posterior modelagem de novos sistemas de SA.

A seguir, apresenta-se uma descrição do córpus e, na Seção 3, algumas considerações finais são feitas.

2. O TeMário 2006

2.1. Características gerais

Para construir o TeMário 2006, foram coletados 150 textos jornalísticos, totalizando 161.289 palavras. Os textos provêm do jornal on-line Folha de São Paulo (doravante, identificado pela sigla FSP) e estão distribuídos entre as seções Brasil, Cotidiano, Dinheiro, Especial, Mundo, Opinião e Tudo. 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 23.041 palavras e a média de palavras por texto é de 1.197 palavras.

Tabela 1 – Características do córpus de textos-fonte

Jornal	Seções	Número de textos	Número de palavras	Média de palavras/texto
Folha de São Paulo	*Brasil*	27	40.951	1.517
	Cotidiano	27	18.908	700
	Dinheiro	30	40.377	1.346
	Especial	10	18.898	1.890
	Mundo	10	16.342	1.634
	Opinião	10	7.928	793
	Tudo	36	17.885	497
Total		**150**	**161.289**	
Média			**23.041**	**1.197**

Foram escolhidos textos jornalísticos para compor o córpus pelo fato de este ser o gênero do TeMário original e pelos textos apresentarem uma linguagem voltada a uma grande audiência de leitores. Desse modo, foram excluídos automaticamente da seleção os suplementos que se destinam a um público mais literato. Essa limitação visa, sobretudo, a facilitar as tarefas relacionadas à SA: 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 SA: os concursos internacionais de avaliação de sumarizadores automáticos, como a SUMMAC (*text SUMMArization evaluation Conference*) e a DUC (*Document Understanding Conference*), 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.

Para a produção do TeMário 2006, 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 a um professor e consultor de editoração de textos em português (o mesmo profissional que construiu o TeMário original) 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.

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 SA, 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.

2.3. Organização do TeMário 2006

Considerando um ambiente hierárquico em que arquivos podem ser armazenados pelo uso do Microsoft Windows, o TeMário 2006 está organizado em uma única pasta, com duas subpastas que agregam, respectivamente, os textos-fonte e os sumários, como mostra a Figura 1.

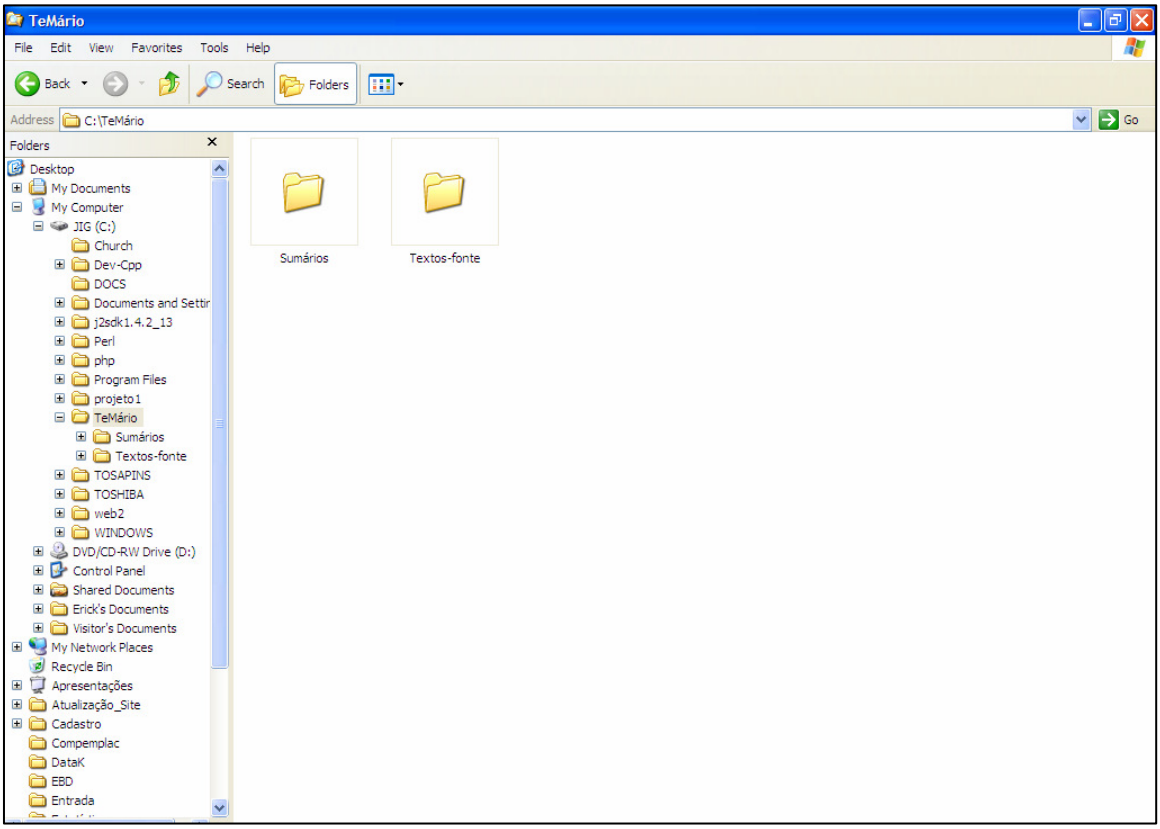


Figura 1 – Diretório "TeMário"

Na pasta de Textos-fonte, há sete pastas (Figura 2), nomeadas de acordo com a seção do Jornal FSP (Brasil, Cotidiano, Dinheiro, Especial, Mundo, Opinião e Tudo), totalizando 150 textos.

Os textos-fonte estão todos em formato txt, já adequado ao 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). Os prefixos indicam as seções do Jornal, como seguem:

- Textos-fonte da seção Brasil têm o prefixo "br";

- Textos-fonte da seção Cotidiano têm o prefixo "co";

• Textos-fonte da seção Dinheiro têm o prefixo “di”;

• Textos-fonte da seção Especial têm o prefixo “ce” (de Caderno Especial);

• Textos-fonte da seção Mundo têm o prefixo “mu”;

• Textos-fonte da seção Opinião têm o prefixo “op”;

• Textos-fonte da seção Tudo têm o prefixo “td”.

Assim, por exemplo, o arquivo de nome ‘br94ab03-22.txt’ indica um texto da seção Brasil da FSP, publicado no dia 03 de abril de 1994; o arquivo ‘mu94ab17-18.txt’ indica um texto da seção Mundo da FSP, publicado no dia 17 de abril de 1994.

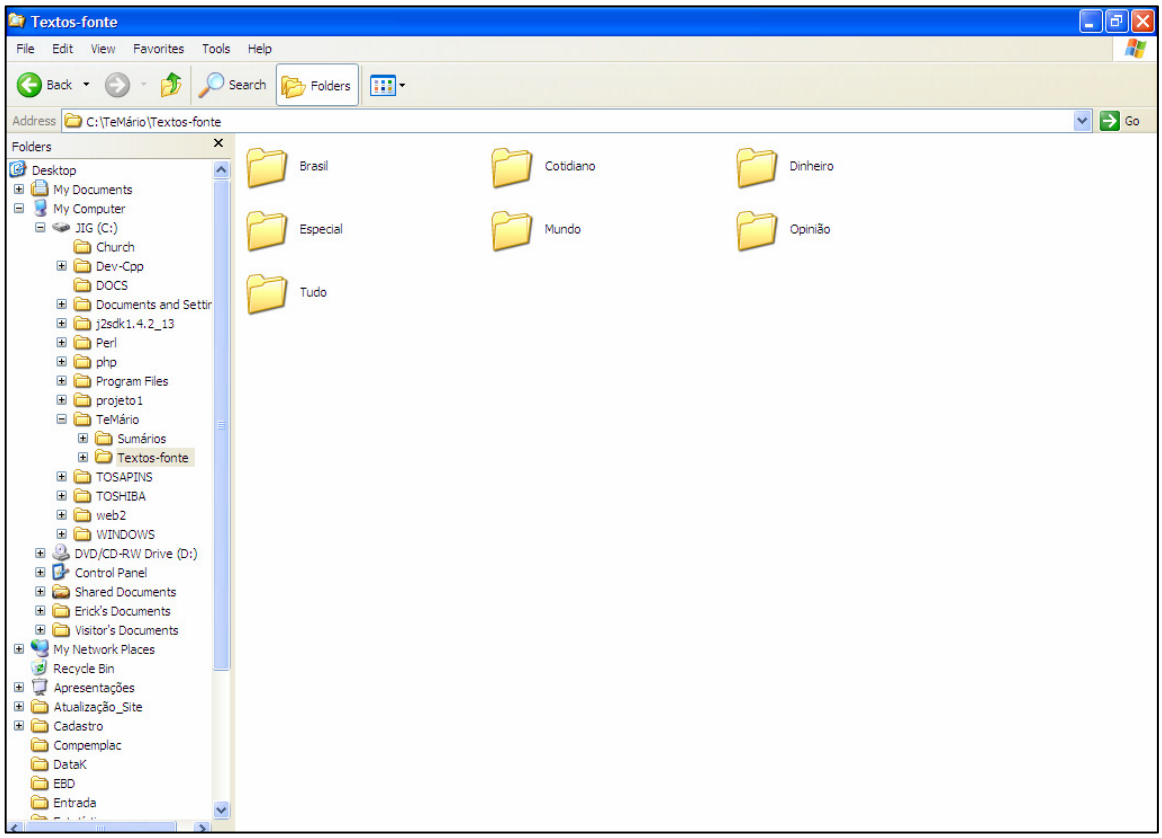


Figura 2 – Subpastas da pasta de “Originais”

Na pasta de Sumários (Figura 3), há também sete subpastas, cada qual, referente à seção do Jornal FSP, conforme descrito anteriormente.

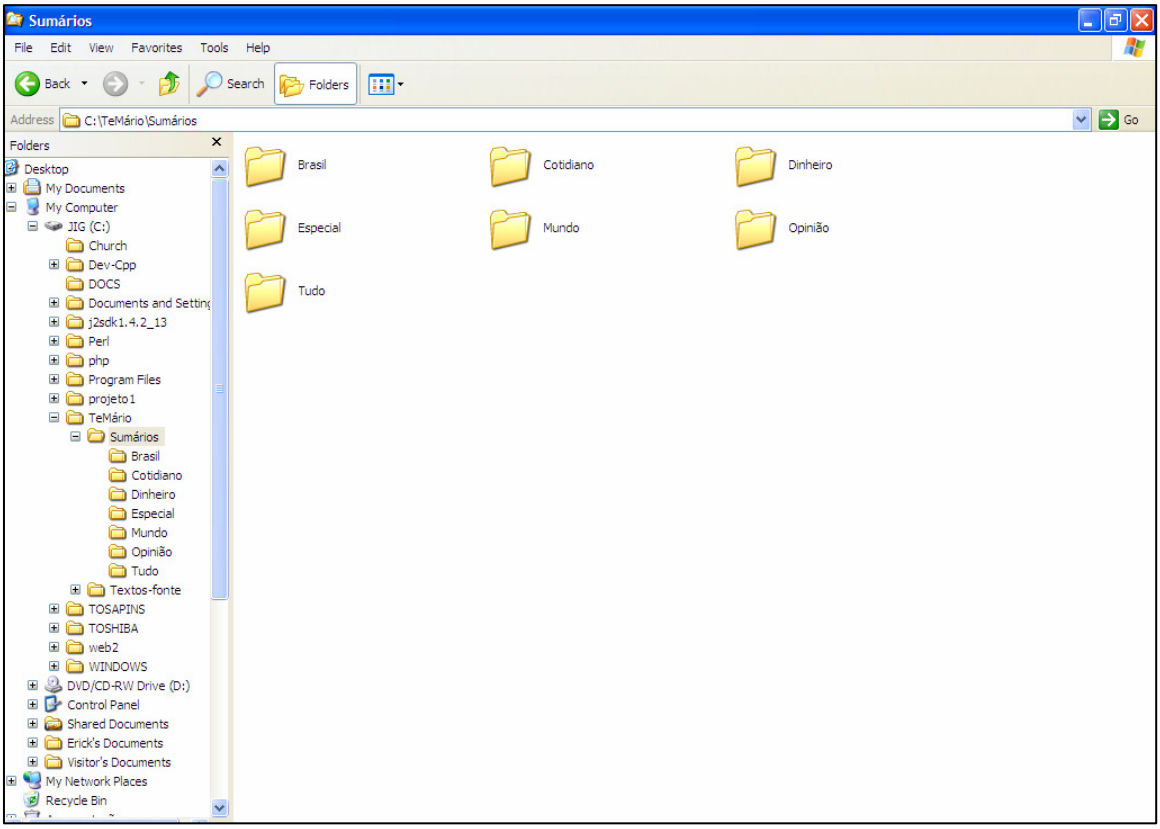


Figura 3 – Subpastas da pasta de "Sumários"

Na pasta de Sumários 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.

3. Considerações finais

Como este relatório descreve, o TeMário 2006 é um córpus de 150 textos jornalísticos e seus correspondentes sumários manuais. Os sumários manuais foram construídos por um profissional em escrita em português para fins de Sumarização Automática.

Com o TeMário original, que contém 100 textos, somam-se para a língua portuguesa 250 textos e seus sumários correspondentes, constituindo o único repositório conhecido para esta língua.

Agradecimentos

Este trabalho contou com o apoio das agências de fomento à pesquisa FAPESP, CAPES e CNPq.

Referências

Pardo, T.A.S. e Rino, L.H.M. (2003). TeMário: Um Corpus para Sumarização Automática de Textos. Série de Relatórios do NILC. NILC-TR-03-09. São Carlos-SP, Outubro, 13p.