

Corpus Summ-it

Renata Vieira e Lucia Rino

Junho de 2008

Resumo

Este corpus foi criado no âmbito dos projetos ProCaCoSA (Processamento de Cadeias de Co-referência para a Sumarização Automática de Textos em Português) e PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil) .

Definição e Origem

Este corpus foi criado no âmbito dos projetos ProCaCoSA (Processamento de Cadeias de Co-referência para a Sumarização Automática de Textos em Português)¹ e PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil) .

O corpus foi desenvolvido como um recurso lingüístico para pesquisa em sumarização automática e sua relação com diferentes aspectos do estudo do discurso. O corpus consiste de 50 textos jornalísticos do caderno de Ciência da Folha de São Paulo (FSP) referentes ao corpus PLN-BR, anotados automaticamente com informações morfossintáticas e anotados manualmente com informações de correferência dos sintagmas nominais e com informações de relações retóricas (RST). O corpus Summ-it conta também com sumários manuais e extratos automáticos (resumos gerados automaticamente). Em [Collovini et al., 2007], é detalhado o processo de anotação de correferência e de relações retóricas.

Os sumários do corpus Summ-it foram construídos e organizados de acordo com as orientações da comunidade da área, registradas por Pardo e Rino ([Pardo and Rino, 2003] e [Rino and Pardo, 2003]). Os sumários manuais foram elaborados por sumarizadores profissionais e os extratos automáticos foram gerados pelos sumarizadores GistSumm ([Pardo et al., 2002] e [Pardo et al., 2003]) e SuPor-2 ([Leite and Rino, 2006a],[Leite and Rino, 2006c] e [Leite and Rino, 2006b]). O processo de elaboração dos sumários manuais é detalhado em [Coelho, 2007].

¹ProCaCoSA: CNPq, Proc. Nro. 507030/2004-4; PLN-Br: CNPq, Proc. Nro. PDITI 550388/2005-2.

Características do Summ-it

Características gerais

Os textos do Summ-it foram coletados no âmbito do projeto PLN-BR. Esse projeto compilou um corpus é formado por 103.080 textos de diversos cadernos (Casa, Ciência, Esporte, Moda, Política, Veículos etc.) do jornal Folha de São Paulo (FSP), compreendidos entre os anos de 1994 a 2005. Desse corpus maior foram retirados os textos que constituem o corpus Summ-it, precisamente, 50 textos do caderno Ciências. Cada documento corresponde a um arquivo texto (ASCII) com tamanho entre 1 Kbytes e 4 Kbytes (de 127 a 654 palavras).

Para a anotação automática (informações morfossintáticas) do Summ-it foi utilizado o analisador sintático PALAVRAS e o conversor Xtractor. O resultado são três arquivos XML²: words.xml (com as palavras do texto), pos.xml (com as categorias morfossintáticas) e chunks.xml (com as estruturas sintáticas das sentenças).

Anotação de Correferência

A anotação manual de informações de correferência seguiu um guia de instruções definido pelos próprios responsáveis por esta anotação [Laboratório de Engenharia da Linguagem – LEL, 2006].

A anotação de correferência seguiu várias etapas: seleção das unidades de interesse (markables), identificação de suas configurações morfossintáticas, indicação das relações entre os diversos markables, classificação dos mesmos e classificação dos relacionamentos anafóricos correferenciais e associativos. A própria ferramenta MMAX permite codificar as marcações indicadas pelos anotadores como elementos markables, associando-os a vários atributos.

A anotação foi realizada em parte de forma manual (com apoio da MMAX) e em parte de forma semi-automática (a identificação dos sintagmas foi feita inicialmente pelo PALAVRAS e passou por uma revisão manual).

O corpus Summ-it foi anotado com informações de correferência por uma equipe de doze anotadores, sendo que cada texto foi anotado por dois anotadores. Cabe salientar que o processo de anotação foi dividido em duas etapas. Primeiramente, cada um dos dois anotadores realizou uma anotação inicial dos textos. Depois, cada par de anotações foi comparado, para se obter um consenso e, se necessário, revisar a anotação.

Como resultado final da anotação, tem-se um total de 590 cadeias de correferência (CCRs), tendo cada CCR uma média de 3 membros e contendo a, mais extensa, 16 membros.

Anotação de Relações RST

A anotação manual de relações RST foi realizada por dois analistas, especialistas em RST, os quais estruturaram cada texto do corpus com o apoio da ferramenta de suporte a anotação RST Tool, v3.45 [O'Donnell, 1997]. Ela seguiu um conjunto de 32 relações, as mesmas adotadas por [Pardo, 2005] em sua tese de doutorado (disponível em <http://www.icmc.usp.br/~taspardo/>).

A anotação seguiu os critérios de [Mann and Thompson, 1988] e [Carlson et al., 2002], além de algumas diretrizes adicionais elaboradas pelos próprios anotadores para dirimir dúvidas.

²eXtensible Markup Language

O processo de anotação é detalhado em duas partes: a primeira contendo a anotação de 12 textos [Carbonel et al., 2007] e a segunda, a anotação dos 38 textos restantes [Fuchs, 2008].

Construção dos Sumários

Os sumários compreendem três tipos de textos: textos tarjados, sumários manuais e extratos automáticos. Os textos tarjados correspondem aos textos-fonte (em formato ASCII) com marcações de destaque para algumas sentenças do texto.

Cabe ressaltar que os textos tarjados e os sumários manuais foram produzidos por uma equipe de três sumarizadores profissionais. Todos eles, pós-graduados, trabalham a mais de dez anos na seleção, produção e sumarização de textos para o poder público.

Os extratos automáticos do Summ-it foram gerados pelo GistSumm([Pardo et al., 2002] e [Pardo et al., 2003]), um sumarizador extrativo projetado para mapear a idéia central de um texto. Para isso, ele combina métodos estatísticos simples. Visando simular o processo humano, ele inicialmente busca a sentença que “melhor expressa a idéia principal” para, então, selecionar as demais sentenças que farão parte do extrato.

Todos os sumários foram gerados com uma taxa de compressão de 70%, isto é, seu tamanho é de, aproximadamente, de 30% do tamanho dos textos-fonte correspondentes.

Ferramentas utilizadas na Anotação do Corpus Summ-it

- MMAX³ versão 0.94 - <http://www.eml.org/>
- PALAVRAS - <http://visl.sdu.dk/visl/pt/>
- Xtractor - <http://abc.di.uevora.pt/xtractor/>
- RSTTool versão 3.45 -<http://www.wagsoft.com/RSTTool/>
- RSTToolkit -http://www.icmc.usp.br/~taspardo/RSTToolkit_Install.rar
- RhetDB - Ferramenta embutida no pacote RSTToolkit
- GistSumm - <http://www.icmc.usp.br/~taspardo/>

³Nota: utilizaram-se as opções de anotação do MMAX: member e pointer.

Organização do Summ-it

Organização do Summ-it disponível para download

O Summ-it está organizado em uma única pasta, com um arquivo “LEIAME.pdf” e seis sub-pastas que agregam, respectivamente, a anotação de correferência, estatísticas desta anotação, a anotação das relações RST, os sumários, os textos originais para consulta e a documentação que acompanha este relatório.

corpusAnotado_CCR

Na pasta “corpusAnotado_CCR”, há 50 pastas correspondentes a cada texto do corpus Summ-it contendo:

- um arquivo texto, com o nome nome_arquivo.txt (arquivo com o texto original);
- cinco arquivos XML
 1. nome_arquivo.txt.words.xml (arquivo com as palavras do texto);
 2. nome_arquivo.txt.pos.xml (arquivo com informações morfosintáticas);
 3. nome_arquivo.txt.chunk.xml (arquivo com estruturas sintáticas das sentenças);
 4. nome_arquivo.txt.markables.xml (arquivo com o resultado da anotação manual);
 5. lel_coref_scheme.xml (esquema de anotação do MMAX);
- duas subpastas com os relatórios da anotação (arquivos HTML)
 1. nome_arquivo_Rel_Member.htm (relatório das cadeias de correferência dos sintagmas nominais (members), contendo a identificação da cadeia de correferência (set), a classificação e o sintagma nominal);
 2. nome_arquivo_Rel_Pointer.htm (relatório dos relacionamentos associativos dos sintagmas nominais (pointers), contendo os sintagmas nominais relacionados (anáfora e antecedente) e o tipo de relação associativa envolvida).

estatisticas_corpusAnotado_CCR

Na pasta “estatisticas_corpusAnotado_CCR”, são disponibilizadas estatísticas referentes ao corpus anotado com informações de correferência.

corpusAnotado_RST

A pasta “corpusAnotado_RST” contém todos os dados da análise RST, divididas em três pastas, assim como a documentação correspondente (relatórios explicativos e um adendo à anotação RST manual. As pastas são as seguintes:

1. Parcial-12txts: contém 12 arquivos, ou árvores RST, correspondentes a 12 textos-fonte, cf. relatado em [Carbonel et al., 2007].

2. `Parcial-38txts`: contém 38 arquivos, ou árvores RST, correspondentes aos 38 textos-fonte restantes, cf. relatado em [Fuchs, 2008].

Essas árvores RST foram construídas manualmente com o apoio da RSTTool [O'Donnell, 1997].

3. `AnotacaoAdicional-RhetDB`: contém informações subjetivas que complementam a análise retórica da 2^a. etapa (38 textos-fonte) e foram introduzidas pela analista com o uso da RhetDB, ferramenta de acesso ao banco de dados de árvores RST embutida no RST Toolkit (de autoria de Thiago Pardo, disponível em <http://www.icmc.usp.br/~taspardo/>). Essa pasta contém os dados da base, propriamente ditos (pasta “`Parcial-38txts-RhetDB`”) e um arquivo `leiam`.

sumarios

Na pasta “`sumarios`”, há quatro pastas assim organizadas:

1. `sumarios_manuais`: arquivos em formato texto que contém sumários construídos por sumarizadores humanos profissionais, cf. relatado em [Coelho, 2007]. São dados que reproduzem, portanto, a tarefa de reescrita, em formato condensado, dos textos-fonte. Os arquivos têm nomes `sum-nome_arquivo.txt` (`nome_arquivo` é o nome do texto original acrescido do prefixo `sum`);
2. `textos_originais_tarjados`: arquivos em formato Word (.doc) que consistem dos mesmos textos-fonte (pasta “`originais`”), com marcações em destaque para algumas sentenças dos textos. Essas anotações foram construídas pelos sumarizadores profissionais, e correspondem às sentenças que eles consideraram mais relevantes para a produção de seus sumários manuais. Nesse caso, essas informações indicam não só a idéia central, mas também todo o conteúdo julgado pertinente e representado, de alguma forma, em sua reescrita em sumários. Os arquivos têm os nomes `tarjado-nome_arquivo.txt` (`nome` do texto original acrescido do prefixo `tarjado`);
3. `extratos_automaticos_Gistsumm`: arquivos em formato texto que contém os extratos automáticos produzidos pelo sumarizador GistSumm([Pardo et al., 2002] e [Pardo et al., 2003]), com os nomes `ext-nome_arquivo.txt` (`nome` do texto original acrescido do prefixo `ext`);
4. `extratos_automaticos_SuPor-2`: arquivos em formato texto que contém os extratos automáticos produzidos pelo sumarizador SuPor-2([Leite and Rino, 2006a],[Leite and Rino, 2006c] e [Leite and Rino, 2006b]), com os nomes `ext-nome_arquivo.txt` (`nome` do texto original acrescido do prefixo `ext`).

textos_originais

Por fim, na pasta “`textos_originais`” estão disponíveis os textos originais em formato texto plano para consultas e outros processamentos.

Referências Bibliográficas

- [Carbonel et al., 2007] Carbonel, T., Fuchs, J., and Rino, L. (2007). Anotação Parcial de Estruturas Retóricas (RST) do Corpus Summ-it. Technical report, NILC-TR-07-04. São Carlos-SP.
- [Carlson et al., 2002] Carlson, L., Marcu, D., and Okurowski, M. (2002). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current Directions in Discourse and Dialogue*, Jan van Kuppevelt and Ronnie Smith eds., Kluwer Academic Publishers. To appear.
- [Coelho, 2007] Coelho, J. C. B. (2007). Uso de informação de correferência e anáfora para verificação da coesão e coerência textual na sumarização automática. Trabalho de Conclusão de Curso de Letras. Unisinos - São Leopoldo.
- [Collovini et al., 2007] Collovini, S., Carbonel, T., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007). Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In *5º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007)*, Rio de Janeiro, RJ. Proceedings of the SBC.
- [Fuchs, 2008] Fuchs, J. (2008). Summ-it: Relatório de anotação RST. Technical report, NILC-TR-08-02. São Carlos-SP.
- [Laboratório de Engenharia da Linguagem – LEL, 2006] Laboratório de Engenharia da Linguagem – LEL (2006). Instruções para anotação de relações anafóricas e referência dêitica.
- [Leite and Rino, 2006a] Leite, D. and Rino, L. (2006a). Supor: extensões e acoplamento a um ambiente para mineração de dados. Technical report, Departamento de Computação, Universidade Federal de São Carlos. NILC-TR-06-07. São Carlos-SP. Agosto, 18 p.
- [Leite and Rino, 2006b] Leite, D. S. and Rino, L. (2006b). A migração do SuPor para o WEKA: potencial e abordagens. Technical report, NILC-TR-06-03. São Carlos - SP, Fevereiro, 35 p.
- [Leite and Rino, 2006c] Leite, D. S. and Rino, L. H. M. (2006c). Selecting a Feature Set to Summarize Texts in Brazilian Portuguese. In *IBERAMIA-SBIA*, pages 462–471.
- [Mann and Thompson, 1988] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. In *Text 8 (3)*, pages 243–281.
- [O'Donnell, 1997] O'Donnell, M. (1997). Rst-tool: An RST analysis tool.
- [Pardo, 2005] Pardo, T. (2005). *Métodos para Análise Discursiva Automática*. PhD thesis, Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP.
- [Pardo and Rino, 2003] Pardo, T. and Rino, L. (2003). TeMário: Um Corpus para Sumarização Automática de Textos. Technical report, NILC-TR-03-09. São Carlos, Outubro, 12p.
- [Pardo et al., 2002] Pardo, T., Rino, L., and Nunes, M. (2002). Extractive summarization: how to identify the gist of a text. In *1st International Information Technology Symposium I2TS*, Florianopolis-SC, Brazil.
- [Pardo et al., 2003] Pardo, T., Rino, L., and Nunes, M. (2003). Gistsumm: A summarization tool based on a new extractive method. In *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, Verlag, Germany.

[Rino and Pardo, 2003] Rino, L. and Pardo, T. (2003). A sumarização automática de textos: Principais características e metodologias. In *Anais do XXIII Congresso da Sociedade Brasileira de Computação. Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA)*. Agosto., pages 203–245, Campinas-SP.