Summ-it

1 BASIC INFORMATION

1.1 *Corpus composition*

The corpus was developed as a linguistic resource for Automatic Summarization research and his relation with different issues to engage studies on the discourse treatment.
Summ-it consists of fifty texts from Science domain extracted from Science section of Brazilian daily newspaper *Folha de São Paulo* (FSP), compose by:
I. Human summaries produced by experts in summarization (Coelho, 2007), rewriting the original texts in a compressed format.
II. Automatic summaries, obtained by GistSumm (Pardo *et al.*, 2002, and Pardo *et al.*, 2003) and SuPor-2 (Leite and Rino, 2006a, Leite and Rino, 2006c, and Leite and Rino, 2006b). All summaries were generated with a 70% compression rate, which means that the summaries correspond to roughly 30% of the original texts.
III. Manual underline sentences which contain relevant informations from the original texts (see 3.2).
IV. Texts semi-automatically annotated with morpho-syntactic informations, assisted by the syntactic parser PALAVRAS (available at: http://visl.sdu.dk/visl/pt/) and Xtractor converter (available at: http://abc.di.uevora.pt/xtractor/).
V. Texts semi-automatically annotated with co-reference informations (MMAX) and with rhetorical relations (RST) (cf. Carbonel *et al.*, 2007, Fuchs, 2008, and Collovini *et al.*, 2007) of noun phrases. The first process intents the identification of the entities in the discourse (e.g. noun phrases) referred or recovered in the text and, the second one, permits to structure a text by relating their discursive units through RST relations.

This resource was built under the ProCaCoSA (Processamento de Cadeias de Co-referência para a Sumarização Automática de Textos em Português) and PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil) projects (more informations available at http://www.inf.pucrs.br/~linatural/procacosa.html), aiming the analysis and the diagnosis of some problems caused to Automatic Summarization and Anaphoric Resolution areas by the occurrence of co-reference chains (CCR) unresolved during the selection and structure of the summaries content. The main goal is to use co-reference information to enrichment of summaries automatically generated.

1.2 *Representation of the corpora (flat files, database, markup)*

The corpus is represented in XCES format.

1.3 *Character encoding*

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 *Contact person*

Name: Sandra Collovini
Address:
PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul,
Faculdade de Informática
Av. Ipiranga, 6681, prédio 32, sala 630
Partenon
90619-900 – Porto Alegre, RS - Brasil
Affiliation: PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul, Faculdade de Informática
Position: Ph.D Student
Telephone: +51 33538630
e-mail: sandra.abreu@acad.pucrs.br; sandrinha.collovini@gmail.com

2.2 *Delivery medium (if relevant; description of the content of each piece of medium)*

The resource will be uploaded on the META-SHARE platform as an archive.

2.3 Copyright statement and information on IPR

This resource is a free license-based for research purposes only, with atribution and distribution allowed under the original license (license at http://creativecommons.org/licenses/by-nc-sa/2.5/br/ - CC BY-NC-SA 2.5).

3 TECHNICAL INFORMATION

3.1 *Directories and files*

The archive that will be uploaded on the META-SHARE platform is a .zip file and contain one "LEIAME.pdf" file, one folder with the resource documentation and five folders with the annotation of co-reference (<corpusAnotado_CCR>), the statistical information of that annotation¸ the annotation of relations RST, and the summaries, the original texts.
Below, the content of files:

<corpusAnotado_CCR>

50 folders corresponding to each text of the Summi-it corpus that contains:

• one text file <file.txt> (with the original text);

• five XML files:

1. <name_file.txt.words.xml> (with the words of the text);

2. < name_file.txt.pos.xml> (with morpho-syntactic informations);

3. < name_file.txt.chunk.xml> (with the syntactic structure of each sentences);

4. < name_file.txt.markables.xml> (with the result of the manual annotation);

5. <lel_coref_scheme.xml> (schema of MMAX annotation);

• two folders with annotation reports (HTML files):

1. < name_file _Rel_Member.htm> (report of co-reference chains from nominal phrases (members), containing the identification of the co-reference chain (set), the classification and the nominal phrase);

2. < name_file _Rel_Pointer.htm> (report of the associative relations from the nominal phrases (pointers), with nominal phrases related (anaphora and antecedent) and the type of associative relation envolved).

<estatisticas_corpusAnotado_CCR>

In this folder, are available statistical data concerning to the annotated corpus with co-reference informations.

<corpusAnotado_RST>

Contains all of data from RST analysis, divided into three folders, as well as the related documentation (reports and adenda of the manual RST annotation):

1. <Parcial-12txts>: 12 files, or trees RST, corresponding to 12 texts-source (cf. Carbonel *et al.*, 2007).

2. <Parcial-38txts>: 38 files, or trees RST, corresponding to remaining 38 texts-source restantes (cf. Fuchs, 2008).

3. <AnotacaoAdicional-RhetDB>: 1 file with the database (<Parcial-38txts-RhetDB>) and 1 <leiame> file.

<sumarios>

1. <sumarios_manuais>: manual summaries in .txt files (e.g <sum-CIENCIA_2000_6381.txt) done by experts in summarizaton (cf. Coelho, 2007);

2. <textos_originais_tarjados>: .doc files (e.g. <tarjado-CIENCIA_2000_6381>) containing the same texts-source (<textos_originais folder);

3. <extratos_automaticos_Gistsumm>: .txt files (e.g. <ext-CIENCIA_2000_6381>) with automatic summaries produced by GistSumm summarizer (cf. Pardo *et al.*, 2002, and Pardo *et al.*, 2003);

4. <extratos_automaticos_SuPor-2>: .txt files (<CIENCIA_2000_6381.txt.ext>) with automatic summaries produced by SuPor-2 summarizer (Leite and Rino, 2006a, Leite and Rino, 2006c, and Leite and Rino, 2006b).

<textos_originais>

Raw text files for queries and other processings.

3.2 Data structure of an entry

<sumarios>

- <extratos_automaticos_Gistsumm>

* 50 files (summuries produced by Gistsumm tool)

- <extratos_automaticos_Supor2>

* 50 files (summuries produced by Supor2 tool)

- <sumarios_manuais>

* 50 files of manual summaries

- <textos_originais_tarjados>

* 50 files in .doc fomat with the content of summaries underlined and in red color.

Following samples of all files described previously:

<corpusAnotado_RST>

- <AnotacaoAdicional-RhetDB> (37 files in .rhetdb format)

elaboration#True#False#---#NS#<2> A informação foi dada na conferência "Da Origem do Universo aos Grandes Telescópios", <3> ministrada ontem por João Steiner, astrofísico da USP, durante a 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência.#2#3#<4> A reunião começou anteontem no campus da UnB <5> (Universidade de Brasília).#4#5#C:\Documents and Settings\ju\Meus documentos\Juliana\ProCaCoSA\RST\Pacote 4\Textos analisados\CIENCIA_2000_6380.rs3.rhetdb#0######

- <Parcial-12> (12 files in .rst format)

 prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos,

 se a emissão de gases por queimadas permanecer nos níveis atuais.

 Esse é o pior panorama climático previsto pelo instituto,

- **<Partial-38>** (38 files in .rst format)

 A informação foi dada na conferência "Da Origem do Universo aos Grandes Telescópios",

 ministrada ontem por João Steiner, astrofísico da USP, durante a 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência.

 A reunião começou anteontem no campus da UnB

 (Universidade de Brasília).

<corpusAnotado_CCR> (50 archives)

- **Report**

* Rel_Member (rel_member.png)

* Rel_Pointer (rel_pointer.png)

- **txt**

* raw texts.

- **anno**

* list of files.

- **chunks** (XML format)

</chunk>
<chunk id="chunk_17" ext="advl" form="pp" span="word_11..word_20">
<chunk id="chunk_18" ext="h" form="prp" span="word_11"></chunk>

```
      <chunk id="chunk_19" ext="p" form="np" span="word_12..word_20">
        <chunk id="chunk_20" ext="n" form="art" span="word_12"></chunk>
        <chunk id="chunk_21" ext="h" form="n" span="word_13"></chunk>
        <chunk id="chunk_22" ext="n" form="pp" span="word_14..word_20">
          <chunk id="chunk_23" ext="h" form="prp" span="word_14"></chunk>
          <chunk id="chunk_24" ext="p" form="np" span="word_15..word_20">
            <chunk id="chunk_25" ext="n" form="art" span="word_15"></chunk>
            <chunk id="chunk_26" ext="n" form="adj" span="word_16"></chunk>
            <chunk id="chunk_27" ext="h" form="n" span="word_17"></chunk>
            <chunk id="chunk_28" ext="n" form="pp" span="word_18..word_20">
              <chunk id="chunk_29" ext="h" form="prp" span="word_18"></chunk>
              <chunk id="chunk_30" ext="p" form="np" span="word_19..word_20">
                <chunk id="chunk_31" ext="n" form="art" span="word_19"></chunk>
                <chunk id="chunk_32" ext="h" form="n" span="word_20"></chunk>
            </chunk>
          </chunk>
        </chunk>
      </chunk>
    </chunk>
</chunk>
```

- **markables** (XML format)

```
<markable id="markable_8" span="word_26..word_27" status="deictic" np_n="yes" np_form="def-np"/>

<markable id="markable_9" span="word_29..word_30" is_anaphoric="encapsulation" status="old" comment="sen anterior" np_n="yes" np_form="def-np"/>

<markable id="markable_80" span="word_34..word_54" status="new" np_n="yes" np_form="def-np"/>

<markable id="markable_81" span="word_41..word_46" member="set_14" np_n="yes" np_form="pn"/>

<markable id="markable_82" span="word_45..word_46" status="new" np_n="yes" np_form="def-pn"/>

<markable id="markable_83" span="word_49..word_54" member="set_2" status="new" np_n="yes" np_form="def-pn"/>
```

- **POS** (XML format)

```
<word id="word_1">
  <n canon="astrônomo" gender="M" number="P">
    <secondary_n tag="Hprof"/>
  </n>
</word>
```

```
<word id="word_2">
   <adj canon="brasileiro" gender="M" number="P"/>
</word>
<word id="word_3">
   <v canon="esperar">
      <fin tense="PR" person="3P" mode="IND"/>
   </v>
</word>
<word id="word_4">
   <v canon="fotografar">
      <inf person="0-1-3S"/>
   </v>
</word>
```

- **Words** (XML format)

```
<word id="word_1">Astrônomos</word>
<word id="word_2">brasileiros</word>
<word id="word_3">esperam</word>
<word id="word_4">fotografar</word>
```

- **annotation scheme** (lel_coref_sheme.xml)

```
<attribute id="level_4" name="np_n" text="Esta expressão é um np c/ nucleo nome">
        <value id="value_4" name="yes" next="level_5"/>
        <value id="value_5" name="no" next="level_6"/>
</attribute>
```

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 50 texts available with different types of summaries and annotations (see 3.2) and needs 4.3 MB compressed (13.9 MB uncompressed) for disk storage.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is Portuguese from Brazil.

4.3 *Domain(s)/register(s) of the corpus*

The corpus contains summaries of Science domain from the daily newspaper *Folha de São Paulo*.

4.4 *Annotations in the corpus (if an annotated corpus)*

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

It is important take into account different types of annotation concerning the different types of treatment level of the 50 texts: process that implies two types of annotation, manual and automatic, aiming the automatic summaries performance and the comparison between summaries made manually and summaries built on the basis of annotation and tools:

I. PALAVRAS and Xtractor converter:

a) semi-automatic annotation: the articles were annotated with morpho-syntactic informations (with PALAVRAS help), after converted to three different formats: *words.xml*, *pos.xml*, and *chunks.xml*.

II. CCR annotation:

Aiming the automatic summaries performance and the comparison between summaries made manually and summaries built on the basis of annotation and tools, a semi-automatic annotation with co-reference information was made by 12 experts annotators with MMAX tool (http://www.eml.org/) support (each text was annotated by 2 annotators):

– Selection of interesting units *(markables)*; identification of their morpho-syntactic configurations; indication of the relations between of several markables; their classification as well as the classification of anaphoric co-referential and associative links.

III. RST annotation:

Semi-automatic annotation by 2 experts annotators in RST, which gave structure to each text with RST tool support choosing from a set of 32 relations (Pardo, 2005).

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

The 50 texts were tagged at the word level by the parser PALAVRAS, with grammatical and semantic tagset. Please, find it at http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html.

4.4.3 Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable.

4.4.4 Attributes and their values (if annotated)

Following the attributes and their values for CCR annotation:

- markables
comment: field for annotator comments
np_form: (np-n=yes) def-np; def-pn; indef-np; dem-np; poss-np; int-np; num-np; quant-np; coord-np; bare-np; pn
pro_form: (np=no) indef-pro; dem-pro; pes-pro; poss-pro; int-pro; num-ana
member: when co-reference relation
pointer: when anaphoric relation
status: new; old; associative; deictic
is_bridging: element-of/element-of-inv; subset-of; part-of/part-of-inv; entiy-attribute; possessor-thing; other-brigding
is_anaphoric: direct, indirect, pronominal; encapsulation (when ”status=old”)

For more informations detailed, see Coelho *et al.* (2006) and Coelho (2007).

4.5 *Intended application of the corpus*

Due to the different levels of annotation, Summ-it corpus can be used in the study of various issues of the discourse modeling and to test and engage the Automatic Summarization models and tools in order to improve the coherence and the informative level of the automatic summaries.

4.6 *Reliability of the annotations (automatically/manually assigned) – if any*

The summaries manual and automatically done and the annotation was automatically done with a manual review.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Carbonel, T., Fuchs, J., and Rino, L. (2007). “Anotação Parcial de Estruturas Retóricas (RST) do Corpus Summ-it.” In *Technical report*, NILC-TR-07-04. São Carlos-SP.

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.

Coelho, J. C. B., Collovini, S., Vieira, R. (2006). “Instruções para anotação de relações anafóricas e referência dêitica”, Laboratório de Engenharia da Linguagem, Universidade de Vale do Rio dos Sinos, Brazil.

Collovini, S., Carbonel, T., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007), “Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática”. In *5.º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL’2007)*,Rio de Janeiro, RJ. Proceedings of the SBC.

Fuchs, J. (2008). “Summ-it: Relatório de anotação RST.” In *Technical report*, NILC-TR-08-02. São Carlos-SP.

Leite, D. and Rino, L. (2006a). “Supor: extensões e acoplamento a um ambiente para mineração de dados.” In *Technical report*, Departamento de Computação, Universidade Federal de São Carlos. NILC-TR-06-07. São Carlos-SP. Agosto, p. 18.

Leite, D. S. and Rino, L. (2006b). “A migração do SuPor para o WEKA: potencial e abordagens.” In *Technical report*, NILC-TR-06-03. São Carlos - SP, Fevereiro, p. 35.

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.

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, 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.

Pardo, T. (2005). *Métodos para Análise Discursiva Automática*. PhD thesis, Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP.