The architecture of SUPeRB

We maintain a large web portal on the computational processing of the Portuguese language, with more than 2,000,000 visits so far.

We list resources, tools and services, as well as actors and publications, and we offer a repository in the area.

We also answer questions and help users about any related subject.

We make available already existing resources and develop new, as well as their full documentation.

Linguateca, a distributed resource center for language technology for Portuguese

www.linguateca.pt

Information

- Corpora (large bodies of text):
  - AC/DC: allows one to query syntactically annotated texts (up to 250 million words) online
  - COMPARA: the largest post-edited parallel corpus in the world: Portuguese and English source texts and their translations
  - Floresta sintáctica: treebank
  - CETEMPúblico, CETENFolha

- IR collections
  - WPT03: all Portuguese Web
  - CHAVE: newspaper doc.s and topics

- Tools
  - Question answering (Esfinge)
  - Named entity recognition (SIEMÉS)
  - Tokenizers, sentence separators
  - Morphological analysers (AnELL)
  - Spellcheckers (Jspell)
  - Word aligners (NATools)

- Other resources
  - Corpógrafo (a full-fledged system for terminology and knowledge management)
  - GKB (Geographic Knowledge Base) and Geo-Net-PT01
  - REPENTINO: a NER gazetteer
  - BACO: database of collocations

- Research tools or resources
  - Example-based machine translation
  - Ontology extraction from text
  - Ontology building from dictionaries
  - SUPERB: Extraction and quality checking of publication citations

Resources

Evaluation

- Organization of evaluation contests
  - Compare several systems around a shared task
  - Create evaluation resources
  - Create evaluation programs
  - Organize a workshop to discuss the results and the evaluation

- Evaluation contests
  - Morfolimiadias (morphological analysis out of context): 2003

- Other evaluation activities
  - MT from English into Portuguese: evaluating the performance of actual Web translation engines
  - Unobtrusive user evaluation of Web services
  - Component evaluation of Esfinge

The architecture of Esfinge

<table>
<thead>
<tr>
<th>Tools</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question answering (Esfinge)</td>
<td>Organization of evaluation contests</td>
</tr>
<tr>
<td>Named entity recognition (SIEMÉS)</td>
<td>Compare several systems around a shared task</td>
</tr>
<tr>
<td>Tokenizers, sentence separators</td>
<td>Create evaluation resources</td>
</tr>
<tr>
<td>Morphological analysers (AnELL)</td>
<td>Create evaluation programs</td>
</tr>
<tr>
<td>Spellcheckers (Jspell)</td>
<td>Organize a workshop to discuss the results and the evaluation</td>
</tr>
<tr>
<td>Word aligners (NATools)</td>
<td>Evaluation contests</td>
</tr>
<tr>
<td>Other resources</td>
<td>Morfolimiadias (morphological analysis out of context): 2003</td>
</tr>
<tr>
<td>GKB (Geographic Knowledge Base) and Geo-Net-PT01</td>
<td>HAREM (Named entity recognition): 2005, 2006</td>
</tr>
<tr>
<td>REPENTINO: a NER gazetteer</td>
<td>Other evaluation activities</td>
</tr>
<tr>
<td>BACO: database of collocations</td>
<td>MT from English into Portuguese: evaluating the performance of actual Web translation engines</td>
</tr>
<tr>
<td>Research tools or resources</td>
<td>Unobtrusive user evaluation of Web services</td>
</tr>
<tr>
<td>Example-based machine translation</td>
<td>Component evaluation of Esfinge</td>
</tr>
<tr>
<td>Ontology extraction from text</td>
<td></td>
</tr>
<tr>
<td>Ontology building from dictionaries</td>
<td></td>
</tr>
<tr>
<td>SUPERB: Extraction and quality checking of publication citations</td>
<td></td>
</tr>
</tbody>
</table>

The architecture of SUPeRB