PANACEA's objective is to join together a number of advanced interoperable tools to build a factory of Language Resources
A production line that automates the stages involved in the acquisition, production, updating and maintenance of the LR required by MT and other Language Technologies.
Web service-based workflows
First products ..
The Language Resource Factory ...

- Bilingual lexica
- TMX aligned texts
- Verbal SCF induction
- Intelligent crawling
- MWU detection

Also Available ...
Already available

The PANACEA registry currently has **144 services** and **11 service providers**.

Chunking/Segmentation (3), Corpus Processing (14), Corpus Workbench (2), Crawling (4), Format Conversion (27), Indexing (1), Language Guessing (1), Lexicon/Terminology Extraction (11), Machine Translation (5), Management (1), Morphological Tagging (4), Morphosyntactic Tagging (12), Named Entity Recognition (3), Querying (3), Statistics Analysis (7), Stemming/Lemmatization (9), Syntactic Tagging (9), Text Mining (1), Tokenization (11).

The PANACEA MyExperiment currently has about **60 workflows** shared by users.
Available workflows

Workflow Entry: Classification of nouns in crawled data for Spanish and 9 available classes

Created at: 10/10/12 @ 14:01:48   Last updated: 10/10/12 @ 16:06:39

<table>
<thead>
<tr>
<th>License</th>
<th>Credits (1)</th>
<th>Attributions (0)</th>
<th>Tags (3)</th>
<th>Featured in Packs (0)</th>
<th>Ratings (0)</th>
<th>Attributed By (0)</th>
<th>Favourited By (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Version 2 (latest) (of 2)

Version created on: 10/10/12 @ 16:06:39 by: Muntsa Padró
Revision comments

Title: Classification of nouns in crawled data for Spanish and 9 available classes

Type: Taverna 2

Preview

(Check on the image to get the full size)
Cost and time reduction by automation is the only way to ensure the continuous supply of LR that can guarantee a LT industry covering all languages, all domains, for current and future needs, and in the time required by the market.
• An SMT can gain approximately a 50% relative improvement of BLUE if trained with domain tuned crawled data that has been automatically cleaned (boiler plate removal, segmentation and tokenization) and aligned.

• Bilingual dictionaries of about 100,000 entries could be extracted from crawled data MOSES phrase tables with only a 10% of error.

• To build a rich information lexicon can be achieved reducing manual work by more than a 40%: Verbal subcategorization frames, lexical semantics.
Achievements

- Automatic lexical analysis to discover OOV items (especially important for German and other languages with frequent compounding) for domain tuning can be done with crawled texts obtaining an annotated morphosyntactic lexicon with an error of a 2%.

- Lexical entries coming from different language lexica and format can be automatically merged. For 2 morphosyntactic lexica (Apertium and Freeling) a new merged resource was produced with 112,000 entries, and a potential error of a 3%.
The Platform

• Deploying with already existing tools when possible
  – Taverna, BioCatalogue and myexperiment
  – Hunalign, GIZA++, FreeLing, RASP, etc.
  – Web mining technologies and modules (Hadoop-based Bixo crawler, Langdetect, Tika document parser, Boilerpipe “web clutter” remover, Bitextor document pair detector, etc.)

• Integrating converters and new modules
Available ...
Delivering produced resources

- Monolingual n-gram corpora for IT, EL, ES, EN and FR
- Monolingual dependency corpora for EL, IT, ES
- Bilingual dictionaries for EN-EL, IT-DE.
- Transfer grammars for EN-DE
- Monolingual SUBCAT lexica for EN, IT, ES
- Monolingual Noun Lex. Classes for EN, ES
- Monolingual MWU lexica for IT
## Available Resources

### Test sets

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Domain</th>
<th># of sites</th>
<th># of document pairs</th>
<th># of sentence pairs extracted by WP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-FR</td>
<td>ENV</td>
<td>6</td>
<td>559</td>
<td>16,487</td>
</tr>
<tr>
<td></td>
<td>LAB</td>
<td>4</td>
<td>900</td>
<td>33,326</td>
</tr>
<tr>
<td>EN-EL</td>
<td>ENV</td>
<td>14</td>
<td>284</td>
<td>15,628</td>
</tr>
<tr>
<td></td>
<td>LAB</td>
<td>7</td>
<td>203</td>
<td>11,719</td>
</tr>
</tbody>
</table>
Clearing the compiled corpora IPR’s...

• Bilingual corpus EN-EL 500 docs, 27K sentence pairs soon available.

• Case study: IPR issues for crawled data:
  - Monolingual: 14,479 URLs / 190,540pg.
  - Bilingual data: 27 URLs / 1,948 pg.

• Clear exploitation rights with source owners lasted from 8 to 344 days (on average 176 days)
For more information, visit www.panacea-lr.eu

Dare to try it!!
Thanks
This document is part of dissemination material generated in the PANACEA Project, Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition (Grant Agreement no. 248064).

This documented is licensed under a Creative Commons Attribution 3.0 Spain License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/es/.

Please send feedback and questions on this document to: iulatrl@upf.edu

TRL Group (Tecnologies dels Recursos Lingüístics), Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (IULA-UPF)

Barcelona, 2012