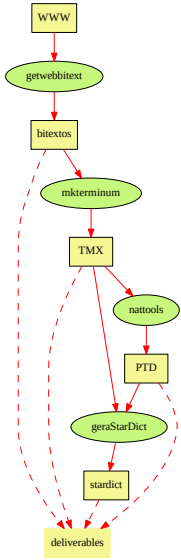


Per-Fide

J.J. Almeida Alberto Simões

23 de Setembro de 2011





- 1 algumas Ferramentas ligadas ao ciclo de vida Per-Fide
- 2 Get Web Bitexts
 - Motivação
 - Corora a partir da Web
 - Problemas
 - Algorithm and Pipelines
 - Pipeline elements
 - Examples
 - Example 1: CFP
 - Example 2: Drinks Corpus
- 3 Some paternalistic suggestions
- 4 Pos Scriptum TMX
- 5 Pos Scriptum PTD



Bitext = a text and its translation

- ① Procura de bitextos na web:
 - de modo controlado
- ② Construir corpora temáticos, úteis, ...:
 - alinhados (etc)
- ③ Gerar o que for possível

E já agora (underground) goals

- Modularidade / funcional programming
- métodos work flow
- types + API + scripting oriented
- quick feedback
- try to build something I can use!





Several WACi

- STRAND (Resnick):
 - partindo de páginas contendo **Português - English**
 - criar uma base de dados com biurl - (bitext candidates)
- BootCat (Baroni):
 - partindo de um conjunto de palavras semente (domain specific)...
$$\left\{ \begin{array}{l} \textbf{foreach } a \leftarrow [\text{some subgroups of seeds}] \\ \quad \{ \text{text}_c \leftarrow \text{text}_c \cup \text{websearch}(a) \\ \quad \text{texts} \leftarrow [\text{Reject bad texts}](\text{text}_c) \\ \quad \text{cor} \leftarrow [\text{Build a corpus with } \text{texts}] \\ \quad [\text{Do miracles with } \text{cor}] \end{array} \right.$$
 - WebBootCat - ... mais um interface Web
- Parguess: $(\text{URL}^*) \rightarrow (\text{biurl candidates})^*$
- MkTerminum: pipelines of tools



Qualidade / ruído

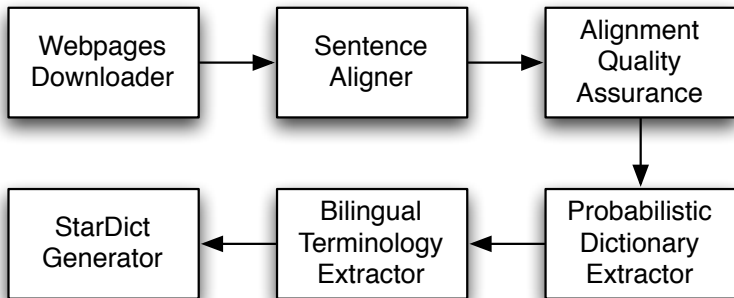
- conteúdo:
 - Riqueza (Terminologically)
 - Fiabilidade da informação
 - tradução literal/criativa
- Estrutura:
 - Formato
 - Qualite das correspondências:
 - traduções em falta ou parciais?
- traduções automáticas

Claim

Precisamos de controle sobre as fontes!



- Controlar e escolher as fontes
- Pipeline para extracção e construção:
 - translation memories (=parallel corpora)
 - terminologia
 - ...
- Modular
- Reutilizando:
 - Open-Corpus-Workbench,
 - Easy-Align,
 - NATools,
 - *Yahoo!* API or Google::Search Ajax
 - StarDict
 - ...



- 1 document URLs that contain all the keywords K and is cataloged by the search engine as being in language L_1 .

$$Docs_{L_1} = yahoo(K, site : D, lang : L_1)$$

- 2 for each URL in $Docs_{L_1}$ try to guess the corresponding URL with the document in language L_2 :

$$Docs_{L_2} = urlguess(Docs_{L_1}, L_2)$$

- 3 retrieve all documents pointed by the obtained URL (if they exist), and convert them to a textual format (PML):

$$Bitexts = retrieve(Docs_{L_1}, Docs_{L_2})$$

- 4 build a parallel corpora PC aligning at the sentence level the retrieved documents. Note that this is done for each document pair.

$$PC = align(Bitexts)$$



- 1 filter the parallel corpora discarding translation units or documents with low alignment quality:

$$PC = \text{filter}(PC)$$

- 2 extract probabilistic translation dictionaries from the aligned corpora:

$$PTD = \text{extractPTD}(PC)$$

- 3 extract bilingual terminology using the probabilistic dictionaries and a set of alignment patterns:

$$Terms = \text{terms}(PC, PTD, Patterns)$$

- 4 create a StarDict dictionary for off-line usage based on the bilingual terminology and dictionaries extracted:

$$StarDict = \text{mkSD}(PTD, Terms)$$



`http://dom.ini.o/en/file.html`

↓ split

`http://dom.ini.o/ + en/file + .html`

↓ url_guesser

... + pt/file + ...
por/file
portuguese/file

↓ join

`http://dom.ini.o/pt/file.html`



```
Fun mkterminum(pairs : bitext*) : TMX  
{ foreach (t1, t2)  $\in$  pairs  
  { [Verify language and compatibility(t1,t2)]  
    { if { [not suitable]  $\Rightarrow$  remove(t1, t2)  
  }  
  { foreach (t1, t2)  $\in$  pairs  
    { tmxi = align(t1, t2)  
      { if { [found many non 1:1 corresp]  $\Rightarrow$  remove(tmxi)  
    }  
  }  
  ret :  $\sum_i$  tmxi
```



- 1 Save some micro TMX (bilingual Call for papers)
`getwebbitext -s "ciawi-conf.org-l pt:es -until tmx trabalhos`
- 2 Build a small corpus of alcoholic drinks
`getwebbitext -l pt:en vodka cerveja`



Example 1: CFP

Extraction of some topics (form a CFP)

`getwebbitext -s "ciawi-conf.org" → the source -l pt:es → languages -until`

`tmx → sub pipeline trabalhos → keywords`

metrics:

800 translation tools, (TMX format) 50 s 10k words



getwebbitext -l pt:en vodka cerveja → default source = european eur-lex
→ default until = ... the end of pipeline

- 3m21s to execute this task,
- 37 MB of bitext candidates (12 bitexts in HTML format, and 22 in PDF)
- Excluding the PDF documents the final TMX:
 - 32 941 TU (about 9MB).
- Including PDF documents (several of the PDF documents were rejected – format or alignment problems)
- the final TMX:
 - had 81 844 translation units
 - (about 22MB of text — about 1 300 000 tokens)

cervejas (29) $\left\{ \begin{array}{ll} \textit{beer} & \rightarrow 98\% \\ \textit{actual} & \rightarrow 2\% \end{array} \right.$

cerveja (53) $\left\{ \begin{array}{ll} \textit{beer} & \rightarrow 62\% \\ \textit{brewing} & \rightarrow 24\% \\ \textit{distilling} & \rightarrow 3\% \\ \textit{coloured} & \rightarrow 1\% \end{array} \right.$

vodka (139) $\left\{ \begin{array}{ll} \textit{vodka} & \rightarrow 94\% \\ \textit{flavoured} & \rightarrow 2\% \\ \textit{vodkas} & \rightarrow 1\% \end{array} \right.$

licor (73) $\left\{ \begin{array}{ll} \textit{liqueur} & \rightarrow 95\% \\ \textit{licor} & \rightarrow 2\% \\ \textit{liqueurs} & \rightarrow 1\% \end{array} \right.$

rum (99) { *rum* → 96%
produced → 1%
word → 1%
solbaerrom → 1%

vinho (271) { *wine* → 81%
vinho → 7%
aromatised → 2%
wines → 2%
wine – based → 1%

vinagres (38) { *vinegar* → 96%

malte (208) { *malt* → 95%

aguardente (226) { *spirit* → 70%
aguardente → 14%
spirits → 13%
diluted → 1%
distilled → 1%



Example 2 conclusion:

The full process took near 30 minutes...

In the end we obtained:

- a 81K translation unit TMX file (22MB);
- a pair of probabilistic translation dictionaries PTD;
- a StarDict dictionary

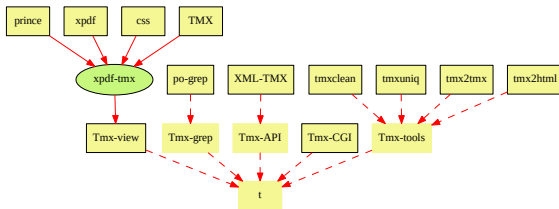


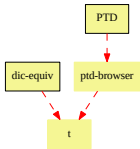


Some paternalistic suggestions

- 1 whenever possible chose sources that you know;
- 2 if you build (small) TMX with good quality translation you can consult and query them now and join them in the future.
- 3 try to select seed terms that just exist in one of the languages
- 4 if you suspect that the sources contain appendices, chose a set of extreme seed terms (eg. "ácido sulfúrico" and "nitrato de prata") to find those rich terminological documents









```
nat-create -id=name -lang=pt..en -tmx file.tmx  
→ fork Natools 33
```

