Escola de Verão: Tradução automática

J.João Dias de Almeida

11 de Setembro de 2006

Outline

Resumo anterior

- traduzir é difícil (máquinas, humanos)
- diferentes problemas
- subtilezas temporais, sintáticas, semânticas, ambiguidade, aspecto
- diversas armadilhas

Estratégia prevista

I18n e L10n

Definition

Internationalization and localization are means of adapting products such as publications or software for non-native environments, especially other nations and cultures.

terminologia alternativa

- Internationalization = **i18n**
- Localization = **l10n**
- i18n + l10n = globalization = **g11n** = ...
- ... = **p13n** = personalization
- **m17n** = multilingualization
- **r3h** = reach

I18N e L10n (continuação)

Difference

- I18n: adaptation of products for potential use virtually everywhere
- L10n: addition of special features for use in a specific locale

Unique to L10N:

- Language translation
- Special support for certain languages such as East Asian languages (CJK(V))
- Local customs
- Local content
- Symbols
- Aesthetics
- Cultural values and social context

Formatos: perigos vários

- textos a traduzir têm um formato que tem de ser preservado
- *desformata* : $Doc \rightarrow (esqueleto \times F*)$
- *reformata* : $(esqueleto \times F*) \rightarrow Doc$
- como traduzir: Thesaurus, Word, LaTeX, XML, HTML, PO, esquemas, PDF

Formato Thesaurus ISO...

Traduzir

```
cat
FR chat
BT felinos
SN an animal that has seven lives

pussy-cat
USE cat
```

Resultado:

- um ficheiro .doc
- remissivas
- SN
- não preservaram as linhas em branco (ou introduziram novas)

XML

- Que elementos/atributos traduzir?
- entidades?

Documentos proprietários

- Word doc
- ...

Outline

O que fazem as ferramentas de tradução?

4. Tradução assistida
   Exemplo:
   - SDL
   - DejaVu
   - OmegaT
   - *Trados*
   - Kbabel, poedit – traduzir mensagens .po
   - ...
5. Tradução automáticas
   Exemplo:
   - Systran,
   - PowerTranslator,
   - Alguns sistemas disponíveis via web (babelfish, ...)
   - *Apertium*

Tradução assistida

1. ... Baseado em Memórias de tradução + terminologia
2. extrai as frases: *desformatador*
   $desform : doc \rightarrow (esqueletoF*)$
   $F*$: independente do formato original.
3. Permite a tradução de cada frase
4. Recompõe o texto: *reformatador*
   $reform : (esqueleto \times F*) \rightarrow doc$
5. ... Versões

Tradução assistida: frase

Para cada frase:

1. Analisa elementos não texto (datas, url, emails, etc)
   $marcaENT : F \rightarrow F'$
2. Vê se a frase é conhecida (na TM) (fuzzy)
   $conhecido : F' \times TM \rightarrow Bool$
   $conhecido : F' \times TM \rightarrow maybe(F \times Prob \times diff)$
3. Vê se a frase contêm terminologia

Memórias de Tradução (TM)

Enredo concreto: Frigorífico de marca X

Em tradução técnica, grande taxa de repetição (30%, 50%, 70%)
Finalidade: reutilização

- reutilização de frases inteiras iguais
- reutilização de frases semelhantes (Ex 80% de palavras comuns)
- reutilização para consulta (tipo concordâncias em corpora paralelos)
- reutilização para extracção de terminologia e outros recursos

Rever e tornar consistentes TM é pesado:

- $tempo(traduzir) = 5 \times tempo(reverTM)$

Como arranjar memórias de tradução?

- compar, intercâmbio com amigos, somar TM,
- durante a tradução assistida, normalmente são enriquecidas as TM
- alinhar Bitextos anteriormente traduzidos
- Arranjar Bitextos na rede, etc –> alinhar

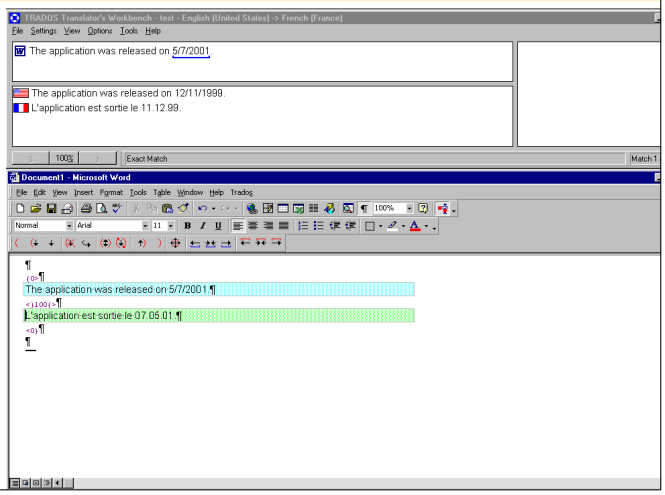
Terminologia

Terminologias : Orientada ao conceito.
Dicionários : Orientados à palavra
Muito importante por:

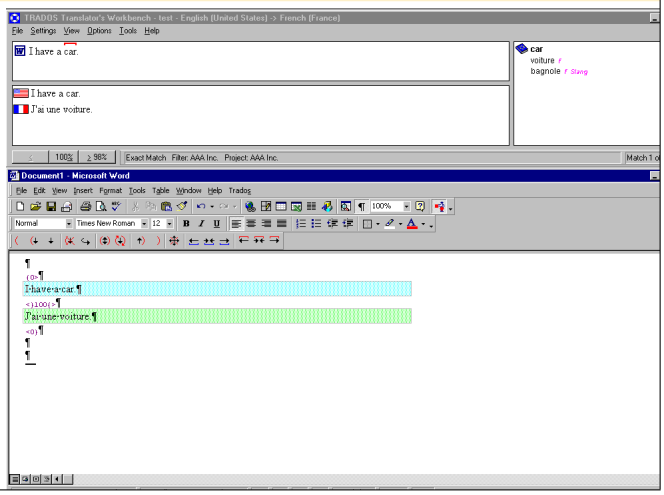
- traduções correctas em relação aos termos técnicos
- coerência nas traduções – termos técnicos normalmente pretende-se que sejam traduzidos da mesma maneira
- é usual gastar imenso tempo para encontrar a tradução certa de certa terminologia

Exemplo
Domain: Computer graphics
Gimp offers a variety of plugins that perform a variety of image manipulations. Examples include bumpmap, edge detect, gaussian blur, and many others.
It features a set of drawing and painting tools such as airbrush, clone, pencil, and paint brush.
Do not use shared memory between Gimp and its plugins.

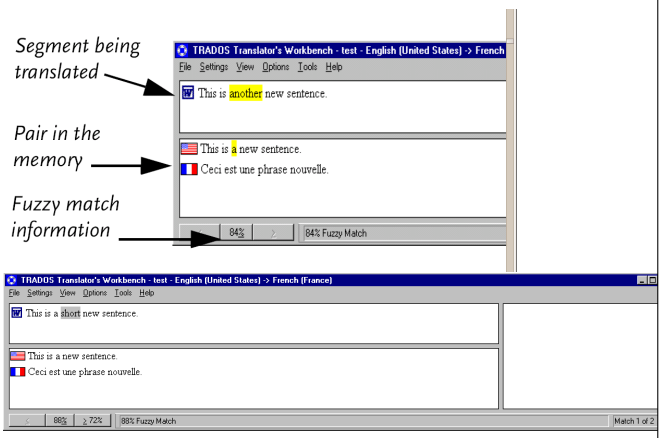
Tradução assistida: Trados 1



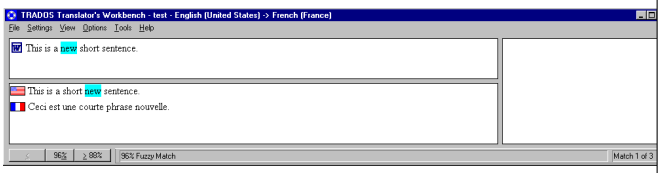
Tradução assistida: trados 2



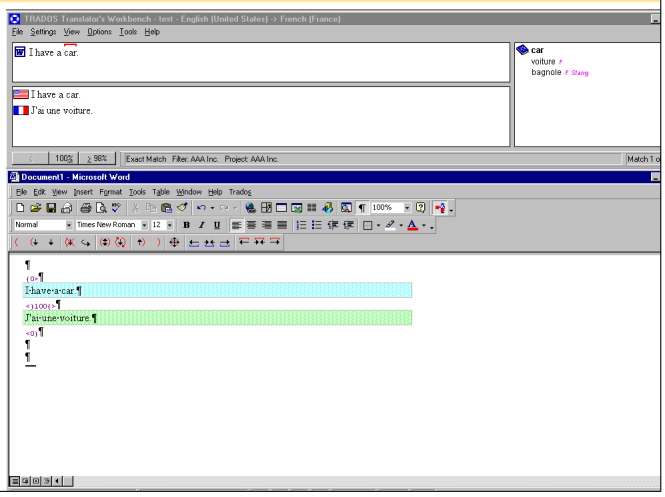
Tradução assistida: fuzzy matching Trados



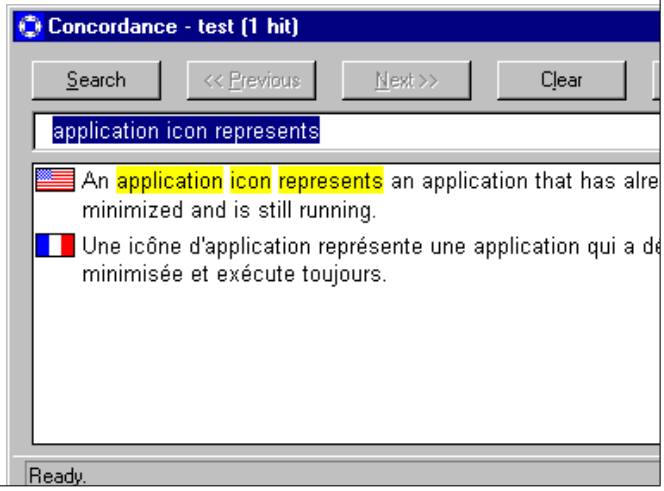
Tradução assistida: fuzzy matching Trados



Tradução assistida: fuzzy matching + terminologia



Tradução assistida: trados Concordância



Tradução assistida: organização do trabalho

4. Gere o trabalho de tradução:
   - trabalho cooperativo,
   - tamanhos, métricas, finanças,
   - workflow, (conhecer a situação actual, estimativas temporais, ...)
   - (norma europeia, certificações)

Tradução assistida: ferramentas associadas

- gestor de TM (coração do sistema)
- gestor de terminologia (Ex: multiterm)
- concordâncias
- extractores e terminologia
- alinhador de textos
- conversores de formatos
  $Any \rightarrow esqueleto \times F*$
  $esqueleto \times F* \rightarrow Any$
- gestor de projecto: workflow + PERT + $ + ...

Tradução assistida: conclusões

- Muito importante para trabalho profissional (inquestionável)
- independências + preservação de formatos

Tradução:
$trad : Texto \times TM \times Terminologia \rightarrow Texto \times TM \times Terminologia$

Alguns formatos imprescindíveis:

- TMX – Translation Memory exange
- XLIFF – (tradutor = editor de XLIFF!!) Mostrar ex.xlf
- termb –

e já agora formatos ligados a SoftWare:

- PO (gettext) – i18n – mensagens programas
- TEI

Mensagens e PO: i18n de programas com gettext

Um ficheiro name.c:

```
printf("My name is %s.\n", my_name);
```

transformado de modo a invocar gettext ...

```
printf(_("My name is %s.\n"), my_name);
```

Corre-se xgettext e produz um template name.pot:

```
#: src/name.c:36
msgid "My name is %s.\n"
msgstr ""

#: src/name.c:49
msgid "I live in %s.\n"
msgstr ""
```

...

corre-se msginit
$msginit : .pot \rightarrow .po$

```
msginit --locale=pt --input=name.pot
```

e cria um pt.po

```
#: src/name.c:36
msgid "My name is %s.\n"
msgstr "My name is %s.\n"
...
```

... traduzido pelos tradutores

```
#: src/name.c:36
msgid "My name is %s.\n"
msgstr "Eu chamo-me %s.\n"
...
```

$msgfmt : pt.po \rightarrow pt.mo$

Como arranjar terminologia?

Tradução automática

Actualmente :

- vários dicionários (médico, informática, jurídico)
- por vezes funciona como um serviço (cliente-servidor)
- "Os tradutores automáticos estão perfeito, os linguistas é que os criticam pos enveja"
- "Os tradutores automáticos são um sistema gerador de anedotas"
- tradução exploratória

Exemplo1 Power-Translator

Demonstração 3: tradução automática e retoques:

- temos um artigo (art_en.doc)
- traduzimos com powerTranslator (art_pt.doc)
- como analisar?:
  - criar tmx (filealigner en.txt pt.txt)
  - ver a tmx
- O que pensam da qualidade obtida?
- como intervir?

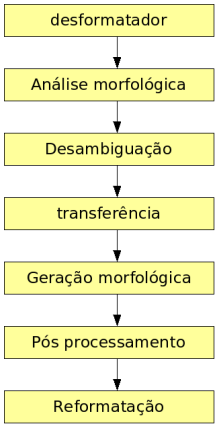
continuação

```
use Text::RewriteRules;
my %dic = impDic("DIC");
while(<>){ print posp(trad($_)); }

MRULES trad
(\w+ \w+ \w+ \w+)==>$dic{$1}!! defined $dic{$1}
(\w+ \w+ \w+)==>$dic{$1}!! defined $dic{$1}
(\w+ \w+)==>$dic{$1}!! defined $dic{$1}
(\w+)==>$dic{$1}!! defined $dic{$1}
(\w+)=e=> ucfirst($dic{lc($1)})!! defined $dic{lc($1)}
(\w+)==>##$1
(.|\n)==>$1

ENDRULES

MRULES posp
...
ENDRULES
```

DIC=

```
Sino de Gordon = Gordon Bell
Sino = Bell
```

Apertium: exemplo de uma arquitectura simples



(estrutura interna do Apertium)

Apertium: scripts internas...

```
apertium-des$FORMATO  f                    |
lt-proc  .../es-pt.automorf.bin            |
apertium-tagger -g .../es-pt.prob          |
apertium-pretransfer                       |
apertium-transfer .../trules-es-pt.xml .../trules-es-pt.bin
    .../es-pt.autobil.bin                  |
lt-proc -g  .../es-pt.autogen.bin          |
lt-proc -p  .../es-pt.autopgen.bin         |
apertium-re$FORMATO                        >     saida
```

Cada bloco pode ser usado independentemente para nosso uso!!!

Apertium: formatos ES-PT

```
<e> <p>
      <l>decir<s n="vblex"/></l>
      <r>dizer<s n="vblex"/></r>
    </p>
</e>
<e>
   <i>mercado<s n="n"/></i>
</e>
<e r="LR">
   <p>
      <l>pulpería<s n="n"/><s n="f"/></l>
      <r>mercado<s n="n"/><s n="m"/></r>
   </p>
</e>
<e r="LR">
    <p>
      <l>añorar<s n="vblex"/></l>
      <r>sentir<g><b/>falta</g><s n="vblex"/></r>
    </p>
</e>
<e>
   <p>
```

Exemplo Apertium

Demonstração Traduzir a constituição portuguesa com o apertium! e aprender algumas das palavras desconhecidas!

```
my $dir = "/home/jj/LN/apertium/apertium-es-pt";
my $tmx = shift
system( "tmxsplit -twente -latin1 $tmx");
system( "apertium-translator $dir pt-es < $tmx-pt > _$tmx.es");
system( "apertium-translator $dir es-pt < $tmx-es > _$tmx.pt");

open(F1,"_$tmx.es") or die;
open(F2,">_$tmx.final.es") or die;
while(<F1>){
  chomp;
  $f= join("  ", (m/\*(\S+)/g) );
  print F2 "$f\n\$\n";
}
close F1;
close F2;
```

Outline

Como ajudar / faça você mesmo

O que é que eu posso fazer de útil no campo dos

- recursos
- ferramentas
- práticas

ligadas à tradução?

Criação de memórias de tradução

1. Encontrar *Bitexto*$*$
2. Alinhar à frase
3. Exportar em formatos públicos (Ex TMX, TEI)

- ... Philip Resnic
- ... Parguess
- Alinhamento de textos paralelos

Philip Resnik – Mining the web for bitexts

1. Expressão de pesquisa tipo

```
      (anchor:"portuguese" OR anchor:"portugues")
  AND (anchor:"english" OR anchor "anglais")
  AND NOT "dictionary"
```

2. produz uma lista de candidatos a Bitextos
3. validação

Estratégia parguess

1. Parte de uma lista de URLs ou de ficheiros paths:
   - Partindo de uma lista of URLs:
     - robot específico
     - contribuição de (....)
   - Partindo de um (ou mais) sites (qualidade mais controlada):
     1. ... eventualmente usar expressões de pesquisa para detectar bisites
     2. wget do site
     3. lista = lista dos ficheiros (`find`)

```
http://www.ex.pt/index_pt.html
http://www.ex.pt/index_en.html
```

2. Analisando Cria uma lista de blocos candidatos blocks

Processamento

1. Procura candidatos a bitextos _pairs:
   $web \vee directory \rightarrow Bitexto*$
2. Validação pairs:
   $Bitexto* \rightarrow Bitexto*$
3. Segmentação à frase:
   $Bitexto* \rightarrow (F^* \times F^*)^*$
4. Alinhamento à frase:
   $(F^* \times F^*)^* \rightarrow ((F \times F)^* \times id^2)^*$
5. seguidamente:
   1. ... cria TMX (em corpus.pt.fr.tmxdir/)
   2. ... alinha à palavra
   3. ... extrai exemplos

Validação de bitextos

- language identification / checking
- file type validation
- file size comparison
- filename similarity checking
- non-text contents comparison

Demonstração

- analisa site http://www.panda.com
- cria várias tmx
- cria uma dicionário probabilístico de tradução (Natools)
- cria um stardict

Demonstração 2

Demonstração criar um dicionário stardict com base em dicionários probabilísticos + exemplo

```
#!/usr/bin/perl
use NAT::Client;

my $dir = shift;
my $dic = do "dir/target-source.dmp";

my $corpus = NAT::Client -> new ( local => $dir );

for $w ( CADA PALAVRA de dic) {
   next unless relevante($w);
   for $t ( CADA TRADUÇÃO COM PROB. > 0.2 ){
      # pesquisa coocorrências  $w $t no servidor
            $concs = $corpus->conc({direction=>'<->'},  $t, $w);
      # guarda o primeiro exemplo
            $dic->{$w}{sample}{$t} = $concs->[0];

print Dumper($dic);

sub relevante {...  }
```

Criação terminologias

- ... ver módulo de terminologia, thesaurus ontologias

Criação de elementos de referência

1. dicionários bilingues
2. corpora temáticos
3. recursos de referências