

Escola de Verão: Corpora, ...

J. João Dias de Almeida

11 de Setembro de 2006

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Estratégia prevista

- 1 Criação de corpus de domínio específico
 - BootCat e NetBootCat
- 2 Criação de corpora: corpora paralelos
 - Philip Resnick – Mining the web for bitexts
 - Estratégia parguess

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Outline

- 1 Criação de corpus de domínio específico
 - BootCat e NetBootCat
- 2 Criação de corpora: corpora paralelos
 - Philip Resnick – Mining the web for bitexts
 - Estratégia parguess

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Corpus de domínio específico

(BootCat e NetBootCat: Baroni, et al)

- dado uma lista **boa** de 20 termos técnicos semente,
- gerar aleatoriamente 100 subconjuntos de 4 termos
- pesquisar o Google e obter 3 URL para cada subconjunto
- ir buscar cada um desse 300 ficheiros
- limpá-los e juntá-los
- processá-los
- -> corpus temático

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Outline

- 1 Criação de corpus de domínio específico
 - BootCat e NetBootCat
- 2 Criação de corpora: corpora paralelos
 - Philip Resnick – Mining the web for bitexts
 - Estratégia parguess

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Criação de corpora paralelos

(= Criação de memórias de tradução)

- 1 Encontrar *Bitexto**
 - 2 Alinhar à frase
 - 3 Exportar em formatos públicos (Ex TMX, TEI)
- ... Philip Resnick
 - ... Parguess
 - Alinhamento de textos paralelos

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Philip Resnick – Mining the web for bitexts

- 1 Algumas páginas bilingues...



- 2 Expressão de pesquisa tipo

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

- 3 produz uma base de dados de candidatas a Bitextos
- 4 quem quiser usa-a
- 5 faz a validação
- 6 problemas: qualidade muito variável

Estratégia parguess

- 1 Parte de uma lista de URLs ou de ficheiros **name.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/ingles/index.html
http://www.ex.pt/portugues/index.html
```

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

- 2 Analisando Cria uma lista de blocos candidatos **name.blocks**

Processamento

- 1 Procura candidatos a bitextos **name.pairs**:

web V directory → *Bitexto**

- 2 Validação **name.pairs**:

*Bitexto** → *Bitexto**

- 3 Segmentação à frase:

*Bitexto** → $(F^* \times F^*)^*$

- 4 Alinhamento à frase:

$(F^* \times F^*)^* \rightarrow ((F \times F)^* \times id^2)^*$

- 5 seguidamente:

- 1 ... cria TMX (em corpus.pt.fr.name.pt.fr.tmxdir/name.pt.fr.tmx)
- 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

Man mkterminum

```
NAME
    mkterminum - makes text alignment, build TMX paralell corpora

SYNOPSIS
    mkterminum [-id=name] [-l1=pt] [-l2=fr] dir [output]
    mkterminum [-l1=pt] [-l2=fr] f.paths [output]
    mkterminum [-l1=pt] [-l2=fr] f.blocks [output]
    mkterminum f.en.pt._pairs [output]

DESCRIPTION
    Depending on the arguments (input / output) some of the following steps are done

    0      dir                directory
    1 dir -> paths            .paths list of files
    2      -> blocks          .blocks list of blocks
    3      -> _pairs          ._pairs list of bitext candidate pairs
    4      -> pairs           .pairs list of bitext
    5      -> tmx             .tmxdir directory with the TMXs

    If we want just to calculate the tmx for a set of bitexts, we can
```

Demonstração

- analisa site <http://www.panda.com>
- extrai, traduz para P, alinha,
- cria várias tmx
- `tmx2tmx -cat *.tmx`
- cria um página de pesquisa HTML
- cria um dicionário probabilístico de tradução (Natools)
- cria um stardict

Demonstração

1. descarregar

```
wget http://www.panda.com (demorou horas)
```

2. Tratar dos zips

```
for a *.zip; do jzip $a ; done
```

3. mkterminum -skipch -patt4pdf -max10

```
./es_des_hl_titanium2005.chm  
./es_des_hl_invent.zip.dir/pinv3_sp.cnt  
./es_des_hl_invent.zip.dir/pinv3_sp.hlp  
./de_des_gu_enterprisecure.zip  
./es_des_gu_webadmin.zip  
./fr_des_hl_platinumis.zip  
./en_des_ft_platinum7.zip  
./es_des_gr_busessecure.zip  
./es_des_hl_trupreventcorp.chm  
./fr_des_hl_titanium2005.zip  
./es_des_gu_titanium.zip  
./en_des_gr_adminsecure.zip  
./de_des_gr_trupreventcorp.pdf
```

Demonst...

```
portuguese => ./pt_des_gr_busessecure.pdf  
french => ./fr_des_gr_busessecure.pdf  
%  
portuguese => ./pt_des_gu_titanium.pdf  
french => ./fr_des_gu_titanium.pdf  
%  
portuguese => ./pt_des_gr_titanium2005.zip  
french => ./fr_des_gr_titanium2005.zip  
%  
portuguese => ./pt_des_hl_titanium.zip.dir/IHFAQS.GIF  
french => ./fr_des_hl_titanium.zip.dir/IHFAQS.GIF  
%  
portuguese => ./pt_des_hl_titanium.zip.dir/IcoOso3.gif  
french => ./fr_des_hl_titanium.zip.dir/IcoOso3.gif  
%  
french => ./fr_des_gr_busessecure.zip  
portuguese => ./pt_des_gr_busessecure.zip  
%  
portuguese => ./pt_des_hl_titanium.zip.dir/IArquivo.GIF  
french => ./fr_des_hl_titanium.zip.dir/IArquivo.GIF
```

Demo...

```
./pt_des_gr_busessecure.pdf ./fr_des_gr_busessecure.pdf  
./pt_des_gu_titanium.pdf ./fr_des_gu_titanium.pdf  
./pt_des_gr_titanium2005.zip ./fr_des_gr_titanium2005.zip  
./pt_des_hl_titanium.zip.dir/IcoOso3.gif  
./fr_des_hl_titanium.zip.dir/IcoOso3.gif  
./pt_des_gr_busessecure.zip ./fr_des_gr_busessecure.zip  
./pt_des_hl_titanium.zip.dir/IArquivo.GIF  
./fr_des_hl_titanium.zip.dir/IArquivo.GIF  
./pt_des_hl_titanium.zip.dir/I0trops.GIF  
./fr_des_hl_titanium.zip.dir/I0trops.GIF  
./_pt.fr.tmxdir/_pt.fr.tmx-pt ./_pt.fr.tmxdir/_pt.fr.tmx-fr  
./_pt.fr.tmxdir/_pt.fr.tmx-fr ./_pt.fr.tmxdir/_pt.fr.tmx-pt  
./pt_des_hl_platinum7.zip ./fr_des_hl_platinum7.zip  
./pt_des_hl_titanium2004.chm ./fr_des_hl_titanium2004.chm  
./pt_des_gr_platinum7.zip ./fr_des_gr_platinum7.zip  
./pt_des_gr_sendmailsecure.pdf ./fr_des_gr_sendmailsecure.pdf  
./pt_des_gr_exchangesecure.zip ./fr_des_gr_exchangesecure.zip  
./pt_des_hl_trupreventcorp.chm ./fr_des_hl_trupreventcorp.chm
```

Dem...

```
./pt_des_gr_busessecure.pdf ./fr_des_gr_busessecure.pdf  
./pt_des_gu_titanium.pdf ./fr_des_gu_titanium.pdf  
./pt_des_gr_sendmailsecure.pdf ./fr_des_gr_sendmailsecure.pdf  
./pt_des_gr_dominosecure.pdf ./fr_des_gr_dominosecure.pdf  
./pt_des_gu_busessecureeex.pdf ./fr_des_gu_busessecureeex.pdf  
./pt_des_gr_platinumis.pdf ./fr_des_gr_platinumis.pdf  
./pt_des_gr_qmailsecure.pdf ./fr_des_gr_qmailsecure.pdf  
./pt_des_gu_titanium2005.pdf ./fr_des_gu_titanium2005.pdf  
./pt_des_gr_titanium2005.pdf ./fr_des_gr_titanium2005.pdf  
./pt_des_gr_pavclsecure.pdf ./fr_des_gr_pavclsecure.pdf  
./pt_des_gu_platinum7.pdf ./fr_des_gu_platinum7.pdf  
./pt_des_gr_adminsecure.pdf ./fr_des_gr_adminsecure.pdf  
./pt_des_gr_platinumis2005.pdf ./fr_des_gr_platinumis2005.pdf  
./pt_des_gu_enterprisecure.pdf ./fr_des_gu_enterprisecure.pdf  
./pt_des_gr_isasecure.pdf ./fr_des_gr_isasecure.pdf  
./pt_des_gr_mimesecure.pdf ./fr_des_gr_mimesecure.pdf  
./pt_des_gu_titanium2004.pdf ./fr_des_gu_titanium2004.pdf  
./pt_des_gr_exchangesecure.pdf ./fr_des_gr_exchangesecure.pdf  
./pt_des_gr_proxysecure.pdf ./fr_des_gr_proxysecure.pdf  
./pt_des_gr_cvpssecure.pdf ./fr_des_gr_cvpssecure.pdf  
./pt_des_gr_l_gdef.pdf ./fr_des_gr_l_gdef.pdf
```

De...

```
***** alinhamento... _(pt fr)
```

```
...  
-rw-rw-r-- 1 jj jj 37199 Jul 12 09:46 _1.pt.fr.tmx  
-rw-rw-r-- 1 jj jj 130012 Jul 12 09:46 _2.pt.fr.tmx  
-rw-rw-r-- 1 jj jj 49519 Jul 12 09:46 _3.pt.fr.tmx  
-rw-rw-r-- 1 jj jj 10653 Jul 12 09:46 _4.pt.fr.tmx  
-rw-rw-r-- 1 jj jj 34449 Jul 12 09:46 _5.pt.fr.tmx  
-rw-rw-r-- 1 jj jj 260554 Jul 12 09:47 _pt.fr.tmx  
5439 31763 260554 _pt.fr.tmx
```

D...

```
<tu><!--1:1-->  
<tuv lang='pt'><seg>INSTALAR O PANDA BUSINESSSECURE REQUISITOS MÍNIMO  
INSTALAÇÃO Requisitos mínimos para instalar o Panda ClientShield Pro  
</seg></tuv>  
<tuv lang='fr'><seg>INSTALLATION DE PANDA BUSINESSSECURE CONDITIONS  
D'INSTALLATION Conditions minimum pour installer Panda ClientShield E  
</seg></tuv>  
</tu>  
  
<tu><!--1:1-->  
<tuv lang='pt'><seg>Disco rígido </seg></tuv>  
<tuv lang='fr'><seg>Disque dur </seg></tuv>  
</tu>  
  
<tu><!--1:1-->  
<tuv lang='pt'><seg>Sistema operativo </seg></tuv>  
<tuv lang='fr'><seg>Système d'exploitation </seg></tuv>
```

Outline

- 3 Caso Estudo 2 – Processamento de Corpora paralelos
 - Exemplo: extracção de dicionários probabilísticos de tradução
 - Exemplo: extracção de subexemplos
 - Exemplo: Juntar exemplos a dicionários probabilísticos e gerar Stardict
- 4 Caso Estudo 3 – Processamento de Treebanks
 - Lingua::Treebank::SimTreeML
 - Padrão de processamento
 - Exemplo 1 – extracção de dicionários simples
 - Exemplo 2 – extrair gramática
 - Exemplo 3 – Extrair atributos de subcategorização verbos
 - Exemplo 4 – Alterar a floresta
- 5 Caso Estudo 4 – pesquisas à medida
 - Exemplo do CondivPT
- 6 Conclusões

Dicionários Probabilísticos de Tradução

```

coruja =>                               enorme =>
count => 36,                             count => 45,
trans =>                                 trans =>
    owl => 97 %                          large => 42 %
    vacuum => 2 %                          huge => 23 %
    forward => 1 %                         enormous => 13 %
                                           (none) => 11 %
                                           deep => 4 %
                                           ...
    
```

Formalmente

$$word_{L_a} \mapsto (occurrences \times word_{L_b} \mapsto \mathcal{P}(\mathcal{T}(word_{L_a}) = word_{L_b}))$$

Extracção de Exemplos

	no	podemos	tolerar	por	mais	tempo	que	a	comissão	continue	este	jogo	do	gato	e	do	rato	!
não	0.4	83.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
podemos	0.0	0.0	73.8	1.8	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
tolerar	0.0	0.0	0.0	72.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
por	0.0	0.0	0.0	0.0	1.5	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mais	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
tempo	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
que	0.0	0.1	0.0	0.0	74.7	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
a	0.3	0.3	0.0	0.0	2.0	32.0	0.0	0.0	1.7	0.0	0.4	1.3	0.0	0.0	0.0	0.0	0.0	0.0
comissão	0.0	0.0	0.0	0.0	0.0	32.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
continue	0.0	0.0	0.0	0.0	0.0	0.0	33.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
este	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
jogo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
do	0.0	0.0	1.0	0.5	0.8	0.0	0.0	0.0	1.5	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0
gato	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	75.6	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
e	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	83.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
do	0.0	0.0	0.7	0.3	0.6	0.0	0.0	1.2	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
rato	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.1	0.0	0.0	0.0	0.0	0.0
!	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	79.1	0.0	0.0	0.0	0.0

```

nãõ                i no
podemos            podemos
tolerar por mais tempo que  tolerar que
a                  la
comissão           comissão
continue este jogo do gato  continúe jugando al gato
e do rato          y al ratón
!                  !
    
```

D 2: dicionários probabilísticos + exemplos -> Stardict

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 );

PARA CADA PALAVRA $w do dic {
    next unless relevante($w);
    PARA CADA TRADUÇÃO $t 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 {... }
    
```

Stardict

The screenshot shows the Stardict application window. The search bar contains the word "prazo". The results list shows various translations and their counts. The main pane displays the word "prazo" with a count of 7682 and 34 translations. The first translation is "term" with a count of 34. The text below the translations shows examples of the word "prazo" in context, such as "...ão destas verbas bem como a tentação de , a longo prazo , fazer desnecessariamente suprlmentos à base tri ...he temptation to draw unnecessarily in the longer term on the local tax base in areas where such project".

Outline

- 3 Caso Estudo 2 – Processamento de Corpora paralelos
 - Exemplo: extracção de dicionários probabilísticos de tradução
 - Exemplo: extracção de subexemplos
 - Exemplo: Juntar exemplos a dicionários probabilísticos e gerar Stardict
- 4 Caso Estudo 3 – Processamento de Treebanks
 - Lingua::Treebank::SimTreeML
 - Padrão de processamento
 - Exemplo 1 – extracção de dicionários simples
 - Exemplo 2 – extrair gramática
 - Exemplo 3 – Extrair atributos de subcategorização verbos
 - Exemplo 4 – Alterar a floresta
- 5 Caso Estudo 4 – pesquisas à medida
 - Exemplo do CondivPT
- 6 Conclusões

Exemplo

CP2-3 Mas como, se muitas não dispõem, nos seus quadros, dos técnicos necessários?

```
fcl
conj-c: Mas adv: como ,
fcl
conj-s: se pron-det: muitas adv: não v-fin: dispõem ,
pp
prp: em
np
art: os pron-det: seus n: quadros
,
pp
prp: de
np
art: os n: técnicos adj: necessários
?
```

Lingua::Treebank::SimTreeML

```
<extract>
<source>CP2-232 ... de idade... </source>
...
<tree cat='pp' fun='N'>
  <t cat='prp' lema='de' fun='H'>de</t>
  <t cat='n' lema='idade'... fun='P'>idade</t>
</tree>
...
<extract>
```

Exemplo de SimTreeML

```
<extract>
<source>CP2-3 Mas como, se muitas não dispõem, ....dos técnicos
necessários? </source>
<tree cat="fcl" fun="QUE">
  <t cat="conj-c" lemma="mas" fun="CO">Mas</t>
  <t cat="adv" lemma="como" fun="ADVL">como</t>
  <punct ort=","/>
  <tree cat="fcl" fun="ADVL">
    <t cat="conj-s" lemma="se" fun="SUB">se</t>
    <t cat="pron-det" lemma="muito" fun="SUBJ">muitas</t>
    <t cat="adv" lemma="não" fun="ADVL">não</t>
    <t cat="v-fin" lemma="dispor" args="PR 3P" fun="P">dispõem</t>
    <punct ort=","/>
    <tree cat="pp" fun="ADVL">
      <t cat="prp" lemma="em" fun="H">em</t>
      <t cat="adj" lemma="necessário" args="M P">necessários</t>
    </tree>
  </tree>
  <punct ort="?"/>
</tree>
</extract>
```

Padrão de processamento

```
func downTr(p:DisTab): ANY
r = <>
ParaTodos ( {tree in treebank : filter(tree)} )
let ( f = p["-end"] or id )
    res = f( proc(tree ,p) )
    in push(r, res)

return r
}

func proc(TXT(pc), p:DisTab) = pc

func proc(XML(ele,att,sons), p:DisTab) =
let( f = p[e] or p["-default"] or toxml,
    v = att
    c = strcat(<proc(son,p) | son in sons >)
    in f(c,v)
}
```

Compilar a gramática

```
use Lingua::Treebank::SimTreeML
dir=> "/opt/treebanks", id => "tb1";

compileTB("/share/tb1.xml")
```

Get the metadata

```
use Lingua::Treebank::SimTreeML
dir=> "/opt/treebanks", id => "tb1";

print Dumper(getmeta())
```

```
Resultado:
{
  trees => 5216
  nt => { np => 26197,
        pp => 20170,
        fcl => 9301,
        icl => 2147,
        vp => 2001, ... },
  t => { n => 24254,
        prp => 20636,
        art => 18569,
        conj-c => 3401 ... }
```

Extracção de dicionários simples

```
use Lingua::Treebank::SimTreeML
dir=> "/opt/treebanks", id => "tb1";

downTr({
  t => sub{ $dic->{$c}{"$v{cat}, $v{lemma}"}++ }
});

print Dumper $dic;
```

produzindo:

```
{a      => { 'art, o'      => 5748,
          'prp, a'     => 2502, },
desaparece => { 'v-fin, desaparecer' => 1 },
reformas  => { 'n, reforma' => 9 },
dotar     => { 'v-inf, dotar' => 3
...
},
```

Navigation icons

Extrair uma gramática probabilística a partir dum treebank

```
downTr({tree => sub{ $c = norm($c);
                  $prod{"$v{cat} --> $c"}++;
                  return $v{cat} },
  t => sub{ return "[$v{cat}] " },
  punct => sub{ return "'$v{ort}' " }});

for(sort {$prod{$b} <=> $prod{$a}} keys %prod)
{ print "$_ $prod{$_}\n";}

sub norm{ ... }
```

Navigation icons

...obtendo-se:

```
pp --> [prp] np      13639
np --> [art] [n]     4248
np --> [art] [n] pp  3527
np --> [art] [prop] 2331
pp --> [prp] [n]    1790
np --> [n] pp       1214
pp --> [prp] icl    1128
pp --> [prp] [prop] 1060
np --> [art] [n] [adj] 964
np --> [pron-det] [n] 872
icl(pcp) --> [v-pcp] pp 679
vp --> [v-fin] [v-pcp] 575
icl --> [v-inf] np  561
np --> [n] [adj]   541
np --> [art] [pron-det] [n] 469
... e muitas mais produções...
```

Navigation icons

Alterar a floresta sintáctica!

```
<t cat="v-fin" lemma="estar" args="..." fun="P">estava</t>
```

...toda de modo a criar novas categorias **ser-v...** para as formas do verbo **ser, estas e ter**

```
%toBeChanged=(ser => 1, ter => 1, estar => 1);

downTr({
  -end => sub { print $c },
  t => sub{ if ($toBeChanged{$v{lemma}}){
            $v{cat}= "$v{lemma}-$v{cat}" };
            return toxml } });
```

Obtendo-se...

```
<t cat="estar-v-fin" lemma="estar" args="..." fun="P">estava</t>
```

Navigation icons

Outline

- 3 Caso Estudo 2 – Processamento de Corpora paralelos
 - Exemplo: extracção de dicionários probabilísticos de tradução
 - Exemplo: extracção de subexemplos
 - Exemplo: Juntar exemplos a dicionários probabilísticos e gerar Stardict
- 4 Caso Estudo 3 – Processamento de Treebanks
 - Lingua::Treebank::SimTreeML
 - Padrão de processamento
 - Exemplo 1 – extracção de dicionários simples
 - Exemplo 2 – extrair gramática
 - Exemplo 3 – Extrair atributos de subcategorização verbos
 - Exemplo 4 – Alterar a floresta
- 5 Caso Estudo 4 – pesquisas à medida
 - Exemplo do CondivPT
- 6 Conclusões

Navigation icons

Outline

- 3 Caso Estudo 2 – Processamento de Corpora paralelos
 - Exemplo: extracção de dicionários probabilísticos de tradução
 - Exemplo: extracção de subexemplos
 - Exemplo: Juntar exemplos a dicionários probabilísticos e gerar Stardict
- 4 Caso Estudo 3 – Processamento de Treebanks
 - Lingua::Treebank::SimTreeML
 - Padrão de processamento
 - Exemplo 1 – extracção de dicionários simples
 - Exemplo 2 – extrair gramática
 - Exemplo 3 – Extrair atributos de subcategorização verbos
 - Exemplo 4 – Alterar a floresta
- 5 Caso Estudo 4 – pesquisas à medida
 - Exemplo do CondivPT
- 6 Conclusões

Navigation icons

Conclusões

- 1 importância de composicionalidade
- 2 importância de dispor de corpora localmente!
- 3 programas podem ser pequenos!
- 4 programar pode ajudar a fazer estragos...
- 5 efeito multiplicativo