Extracção de Recursos com base em Dicionários Probabilísticos de Tradução

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

Workshop de Tecnologia da Linguage da Humana
25 de Outubro de 2007

Recursos Monolingues e Bilingues:

- Dicionários Probabilísticos de Tradução;
- Palavras Aparentadas;
- Exemplos de Tradução;
- Terminologia Bilingue;
- Dicionários StarDict;
- Classes de Palavras;
- Exemplos de Tradução Genéricos;

Matéria Prima: Corpora

	PT–EN	PT–ES	PT–FR
Constituição Portuguesa⋆	2 013	2 011	2 013
COMPARA⋆⋆	97 215	—	—
Le Monde Diplomatique⋆	—	—	68 231
JRC	286 008	281 185	277 754
EuroParl	(300 MB) 998 830	1 006 895	1 023 841
EurLex⋆	(3 GB) 10 394 893	1 111 068	1 710 760

Valores correspondem ao número de unidades de tradução.

⋆ Criados pelo projecto Natura;

⋆⋆ Criado pela Linguateca, só disponível via Web. (elemento de comparação)

Parte I

Dicionários Probabilísticos de Tradução

Dicionários Probabilísticos de Tradução (Ex 1)

```
QUERY> europa
  Occurrences: 39917
  Translations:
      88.50%  europe
       5.73%  european
       2.37%  europa
       1.16%  (none)
       0.57%  eu
       0.23%  unece
       0.17%  the
       0.16%  auto
```

Dicionário Probabilístico de Tradução (Ex 3)

```
QUERY> read
  Occurrences: 2435
  Translations:
      29.32%  ler
      13.75%  li
       8.36%  read
       5.96%  lido
       3.54%  lemos
       1.60%  leio
       1.46%  estar
       1.45%  leu
```

```
QUERY> represent
  Occurrences: 2538
  Translations:
      17.87%  representam
      11.57%  representar
       8.93%  represento
       7.54%  representamos
       4.93%  constituem
       3.63%  representa
       3.37%  (none)
       2.35%  representante
```

Dicionário Probabilístico de Tradução (Ex 5)

```
QUERY> palavra
  Occurrences: 6337
  Translations:
      35.75%  floor
      16.88%  word
      13.57%  (none)
       9.28%  speak
```

Floor??

```
Tem a palavra , em nome da comissão , o senhor comissário Barnier .
Mr Barnier has the floor on behalf of the Commission .

Tem a palavra , em nome da comissão , a senhora comissária wallström .
Mrs wallström has the floor on behalf of the Commission .
```

Dicionários Probabilísticos de Tradução: Conclusões

- um dicionário probabilístico de tradução não pode ser visto como um dicionário tradução convencional;
- é possivel a partir de um dicionário probabilístico de tradução obter um dicionário de tradução convencional;
- os dicionários probabilísticos de tradução são úteis para a tradução manual e automática
- os dicionários probabilísticos de tradução são úteis para a criação/bootstraping de dicionários manuais;

Parte II

Palavras Aparentadas

Construção de listas de palavras associadas

```
use NAT::Client;

my $client = NAT::Client->new( crp => "EuroParl-PT-EN" );
my %r = ();

my $a1 = $client->ptd( $_ );
for my $b1 (keys %{$a1->[1]}) {
  my $c = $client->ptd( { from => 'target' }, $b1);
  for my $d ( keys %{$c->[1]} ){
    $r{$d}    +=  $a1->[1]{$b1} * $c->[1]{$d};
  }
}

for(( sort {$r{$b} <=> $r{$a}} keys %r)[0..9]) {
    printf "  %s (%.3f) BSn", $_, $r{$_}*100
}
```

Resultados de listas de palavras associadas

país		povo	
país	(62.511)	pessoas	(36.158)
países	(8.153)	povo	(9.914)
estado	(0.453)	cidadãos	(5.934)
território	(0.427)	população	(5.321)
turquia	(0.412)	popular	(3.872)
de	(0.332)	povos	(3.237)
nacionais	(0.277)	nação	(1.830)
há	(0.170)	os	(1.748)
em	(0.145)	nacionais	(0.388)

Parte III

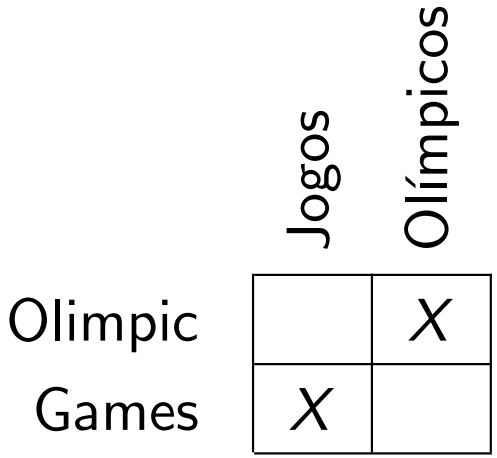
Exemplos de Tradução

Diagonalização da Matriz de Tradução

Segmentar unidades de tradução:

- construir uma matriz com probabilidades de tradução mútua;
- considerar células de alta probabilidade como âncoras;
- assumir que as traduções se fazem da esquerda para a direita;
- assumir portanto que as maiores probabilidades se encontram na diagonal principal da matriz.

Exemplo 1: Padrão ABBA



Descrita como:

```
[ABBA] A B = B A
```

Formalmente...

$$\mathcal{T}(A \cdot B) = \mathcal{T}(B) \cdot \mathcal{T}(A)$$

Exemplo 2: Padrão IDH

	índice	de	desenvolvimento	humano
human				X
development			X	
index	X			

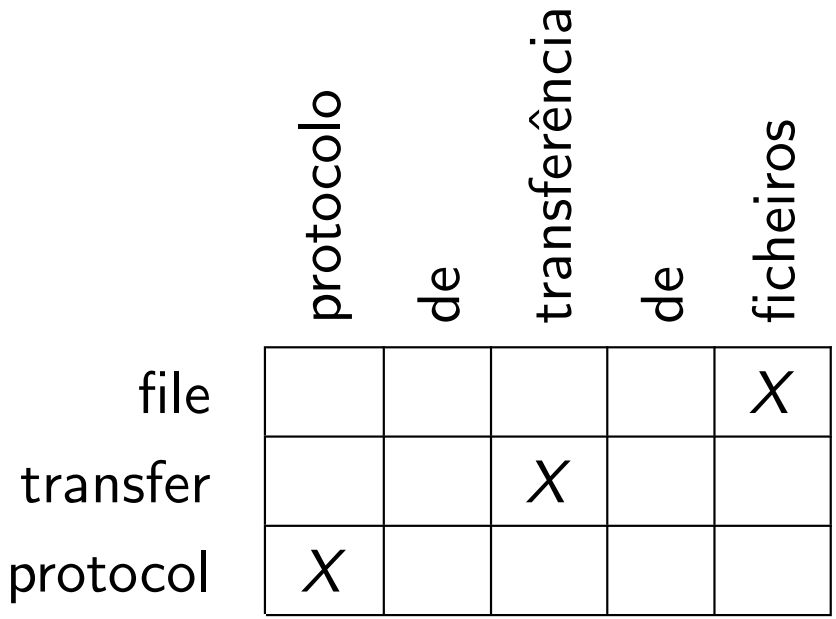
Descrita como:

```
[IDH] I "de" D H = H D I
```

Formalmente...

$$\mathcal{T}(I \cdot "de" \cdot D \cdot H) = \mathcal{T}(H) \cdot \mathcal{T}(D) \cdot \mathcal{T}(I)$$

Exemplo 3: Padrão FTP



Descrita como:

```
[FTP] P "de" T "de" F = F T P
```

Formalmente...

$$\mathcal{T}(P \cdot "de" \cdot T \cdot "de" \cdot F) = \mathcal{T}(F) \cdot \mathcal{T}(T) \cdot \mathcal{T}(P)$$

Parte IV

Extracção de Terminologia

Padrões com Restrições

A linguagem de padrões permite:

Restrições Morfológicas

Definição de propriedades morfológicas que devem ser válidas para a expressão em causa.

```
[ABBA] A B[CAT<-adj] = B[CAT<-adj] A
```

Predicados Genéricos

Definição de predicados genéricos, definindo funções Perl sobre palavras ou conjunto de palavras.

```
[ABBA] A B.is_adj = B.is_adj A
```

Inferência

Realizar inferência de propriedades (normalmente morfológicas) a partir das expressões encontradas.

```
[ABBA] A B[CAT->adj] = B[CAT->adj] A
```

Consolidação de Terminologia

```
39214 = comunidades europeias =!ABBA!= european communities
32850 = jornal oficial =!ABBA!= official journal
32832 = parlamento europeu =!ABBA!= european parliament
32730 = união europeia =!ABBA!= european union
31650 = comunidade europeia =!ABBA!= european community
15602 = países terceiros =!ABBA!= third countries
 [...]
 3614 = livro verde =!ABBA!= green paper
 3520 = saúde pública =!ABBA!= public health
 3434 = direito comunitário =!ABBA!= community law
 3243 = conselho europeu =!ABBA!= european council
 3227 = nível comunitário =!ABBA!= community level
 3179 = comité permanente =!ABBA!= standing committee
 3038 = nomenclatura combinada =!ABBA!= combined nomenclature
 [...]
    1 = órgãos orçamentais =!ABBA!= budgetary organs
    1 = órgãos relevantes =!ABBA!= relevant bodies
    1 = óvulos de equino =!A!= equine ova
    1 = óxido de albendazole =!A!= albendazole oxide
    1 = óxido de cádmio =!A!= cadmium oxide
    1 = óxido de estireno =!A!= styrene oxide
```

Exemplos A B = B A

```
21007 união europeia => european union
 9301 parlamento europeu => european parliament
 4171 direitos humanos => human rights
 3504 estados unidos => united states
 2353 mercado interno => internal market
 1911 posição comum => common position
 1826 países candidatos => candidate countries
 1776 comissão europeia => european commission
 1708 conselho europeu => european council
 1629 saúde pública => public health
 1558 direitos fundamentais => fundamental rights
 1546 nações unidas => united nations
 1337 países terceiros => third countries
 1294 conferência intergovernamental => intergovernmental conference
 1258 fundos estruturais => structural funds
```

Exemplos A "de" B = B A

729 plano de acção => action plan
722 conselho de segurança => security council
680 processo de paz => peace process
582 mercado de trabalho => labour market
580 pena de morte => death penalty
492 pacto de estabilidade => stability pact
431 política de defesa => defence policy
353 acordo de associação => association agreement
348 protocolo de quioto => kyoto protocol
343 programa de acção => action programme
259 branqueamento de capitais => money laundering
258 comité de conciliação => conciliation committee
241 política de concorrência => competition policy
226 processo de conciliação => conciliation procedure
217 requerentes de asilo => asylum seekers

Exemplos A B C = C B A

```
531 política agrícola comum => common agricultural policy
418 banco central europeu => european central bank
329 tribunal penal internacional => international criminal court
166 aliança livre europeia => european free alliance
156 modelo social europeu => european social model
153 partidos políticos europeus => european political parties
 83 fundo monetário internacional => international monetary fund
 75 política externa comum => common foreign policy
 66 organização marítima internacional => international maritime organi
 65 própria união europeia => european union itself
 65 fundo social europeu => european social fund
 55 direitos humanos fundamentais => fundamental human rights
 45 relações económicas externas => external economic relations
 45 homens e mulheres => women and men
 45 agência espacial europeia => european space agency
```

Exemplos I "de" D H = H D I

95 mandato de captura europeu => european arrest warrant
85 fontes de energia renováveis => renewable energy sources
80 mandado de captura europeu => european arrest warrant
67 sistemas de segurança social => social security systems
64 zona de comércio livre => free trade area
55 força de reacção rápida => rapid reaction force
54 orientações de política económica => economic policy guidelines
46 planos de acção nacionais => national action plans
46 direitos de propriedade intelectual => intellectual property rights
33 sistema de alerta rápido => rapid alert system
29 política de defesa comum => common defence policy
29 método de coordenação aberta => open coordination method
27 método de coordenação aberto => open coordination method
27 conselho de empresa europeu => european works council
25 acordo de comércio livre => free trade agreement

Extracção de Terminologia — Avaliação

- EuroParl PT-EN: 1 000 000 TUs
- 700 000 unidades de tradução processadas
- 578 103 ocorrências de padrões
- 139 781 exemplos diferentes
- 103 617 exemplos depois de filtrados (stop words, ruído,...)
- 77 497 ex. com a regra A B = B A (938/2/1) (86%)
- 12 694 ex. com a regra A "de" B = B A (204/2/1) (95%)
- 7 700 ex. com a regra A B C = C B A (40/1/1) (93%)
- 3 336 ex. com a regra H "de" D H = H D I (21/1/1) (100%)
- 1 466 ex. com a regra A B C = C A B (4/1/1) (40%)
- 564 ex. com a regra P "de" V N = N P "of" V (6/1/1) (98%)
- 360 ex. com a regra P "de" T "de" F = F T P (3/1/1) (96%)

Parte V

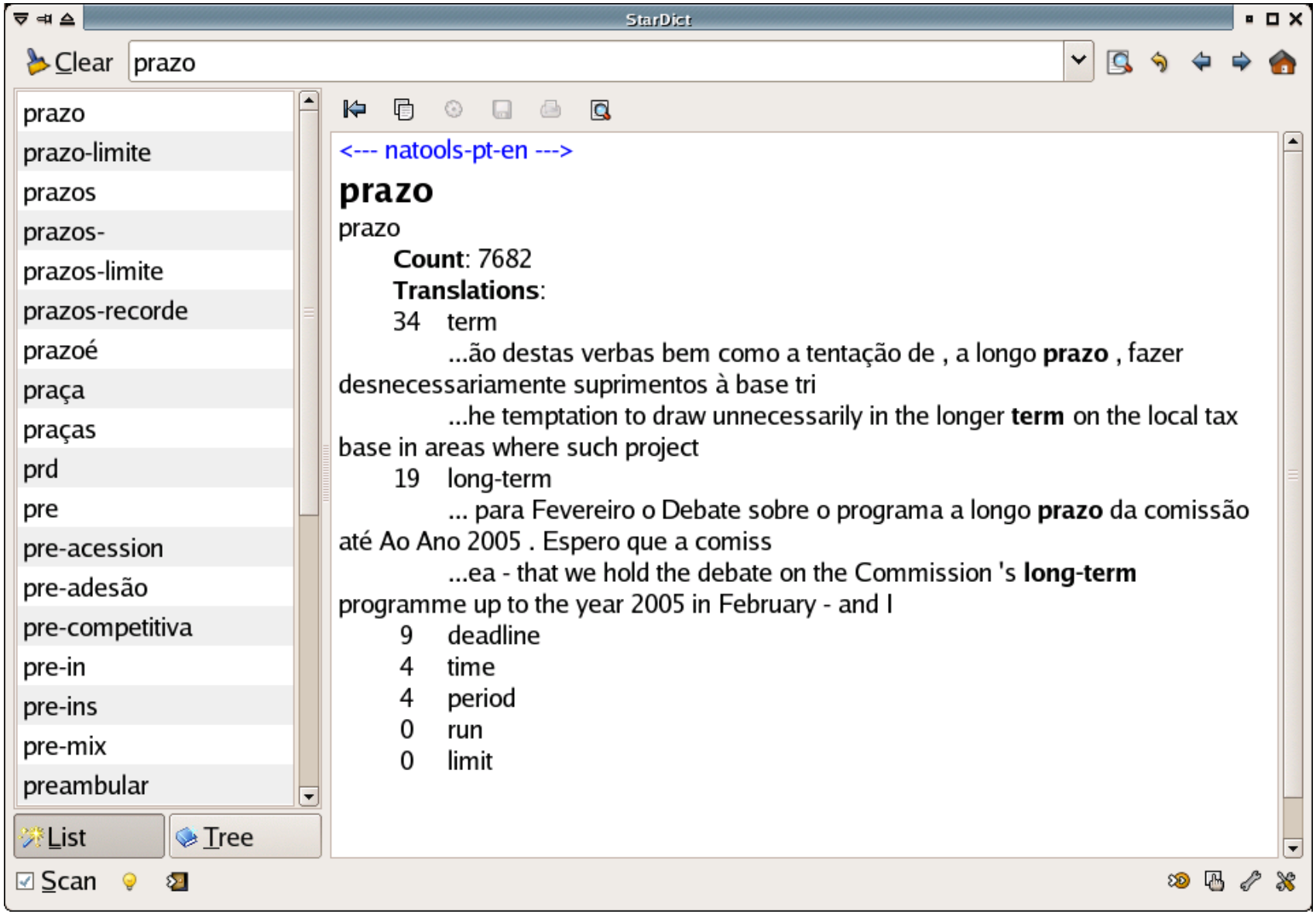
Dicionário StarDict

Construção de um dicionário StarDict

Queremos construir um dicionário StarDict com exemplos de uso:

```
use NAT::Client;

$client = NAT::Client -> new ( crp => "EuroParl-PT-EN" );

$client-> iterate( { Language => "PT" },
            sub {
                my %param = @_;

                for my $trans (keys %{$param{trans}}) {
                    if ($param{trans}{$trans} > 0.1) {
                        my $concs = $client-> conc({concordance => 1},
                                                   $param{word}, $trans);
                        $stardict{$param{word}}{$trans} = $concs -> [0];
                    }
                }
    });

print StarDict($stardict);
```

Construção de um dicionário StarDict



Parte VI

Classes de Palavras

```
'ácido' => [ 'clorídrico (hydrochloric acid)',
             'sulfúrico  (sulphuric acid)',
             'acético    (acetic acid)',
             'fólico     (folic acid)',
             'cítrico    (citric acid)',
             'nítrico    (nitric acid)',
             'tartárico  (tartaric acid)',
             'benzóico   (benzoic acid)',
             'fórmico    (formic acid)',
             'málico     (malic acid)',
             'sulfúrico  (sulfuric acid)',
             'erúcico    (erucic acid)',
  ...
'livro' => [ 'verde      (green paper)',
             'branco     (white paper)',
             'azul       (blue paper)',
             'aberto     (open book)',
             'azul       (blue book)',
             'branco     (white book)',
             'laranja    (orange book)',
             'vermelho   (red book)'
  ...
```

Parte VII

Generalização de Exemplos

Generalização de Exemplos/Terminologia (1)

399	às _horaA_	_horaB_
187	orçamento de _ano_	_ano_ budget
136	_int_ euros	eur _int_
135	_int_ euros	eur _int_
127	directiva de _ano_	_ano_ directive
51	orçamento _ano_	_ano_ budget
46	_int_ de setembro	september _int_
31	partir de _ano_	_ano_ onwards
29	convenção de _ano_	_ano_ convention
26	eleições de _ano_	_ano_ elections
25	período _ano_-_ano_	_ano_-_ano_ period
25	_int_ dólares	usd _int_
24	relatório de _ano_	_ano_ report
21	convenção de genebra de _ano_	_ano_ geneva convention
17	período de _ano_-_ano_	_ano_-_ano_ period

Generalização de Exemplos/Terminologia (2)

2	povo português	portuguese	people
2	povo paraguaio	paraguayan	people
2	povo nigeriano	nigerian	people
2	povo mexicano	mexican	people
2	povo marroquino	moroccan	people
2	povo mapuche	mapuche	people
2	povo indígena	indigenous	people
2	povo holandês	dutch	people
2	povo húngaro	hungarian	people
2	povo hmong	hmong	people
2	povo guatemalteco	guatemalan	people

generalizando...

povo	X: gentílico(X)	T(X)	people
governo	X: gentílico(X)	T(X)	governo

```
ácido X.acidClass = X.acidClass acid
```

em que X.acidClass contém:

```
clorídrico = hydrochloric
sulfúrico  = sulphuric
acético    = acetic
fólico     = folic
cítrico    = citric
nítrico    = nitric
tartárico  = tartaric
benzóico   = benzoic
fórmico    = formic
málico     = malic
sulfúrico  = sulfuric
erúcico    = erucic
 ...
```

http://natools.sf.net/