Extracção de Recursos de Tradução com base na Hipótese das Palavras-Marca

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

Orientação
José João Almeida

Simpósio Doutoral da Linguateca 2007b

Hipótese das Palavras-Marca — *definição*

- Em (Green, 1979) é definida a Marker Hypothesis, uma restricção psicolinguística na estrutura gramatical de línguas naturais;
- Esta "hipótese" conjectura que todas as línguas naturais têm a sua estrutura gramatical marcada (ou delimitada) por um conjunto fechado de lexemas ou morfemas.
- Este conjunto contém habitualmente preposições, pronomes, locuções, artigos, determinantes e alguns advérbios.
- Tem vindo a ser usada para a divisão em segmentos que em muitos casos se aproximam a sintagmas.
- Chunker básico e razoavelmente eficaz.

Hipótese das Palavras-Marca — *exemplo*

O João passou toda a tarde a brincar com os colegas.

⇓

O João passou toda a tarde a brincar com os colegas.

⇓

(O João passou) (toda a tarde) (a brincar) (com os colegas.)

Hipótese das Palavras-Marca — *exemplo*

O João passou toda a tarde a brincar com os colegas.

⇓

O João passou toda a tarde a brincar com os colegas.

⇓

(O João passou) (toda a tarde) (a brincar) (com os colegas.)

Hipótese das Palavras-Marca — *exemplo*

O João passou toda a tarde a brincar com os colegas.

$\Downarrow$

O João passou toda a tarde a brincar com os colegas.

$\Downarrow$

(O João passou) (toda a tarde) (a brincar) (com os colegas.)

Hipótese das Palavras-Marca — *para quê?*

- Permite segmentar frases de forma eficaz:
  - pelo menos na língua portuguesa e inglesa;
  - algoritmo mais rápido do que chunkers habituais;
- Segmentação com alguma informação sintáctica:
  - contrapor com os exemplos obtidos de forma "ad-hoc";
- Estruturalmente ricos:
  - permitem a extracção de sub-relacionamentos de forma simples;
- Possibilidade de colaboração com peritos na área...

Hipótese das Palavras-Marca — *para quê?*

- Permite segmentar frases de forma eficaz:
  - pelo menos na língua portuguesa e inglesa;
  - algoritmo mais rápido do que chunkers habituais;
- Segmentação com alguma informação sintáctica:
  - contrapor com os exemplos obtidos de forma "ad-hoc";
- Estruturalmente ricos:
  - permitem a extracção de sub-relacionamentos de forma simples;
- Possibilidade de colaboração com peritos na área...

Hipótese das Palavras-Marca — *para quê?*

- Permite segmentar frases de forma eficaz:
  - pelo menos na língua portuguesa e inglesa;
  - algoritmo mais rápido do que chunkers habituais;
- Segmentação com alguma informação sintáctica:
  - contrapor com os exemplos obtidos de forma "ad-hoc";
- Estruturalmente ricos:
  - permitem a extracção de sub-relacionamentos de forma simples;
- Possibilidade de colaboração com peritos na área...

Hipótese das Palavras-Marca — *para quê?*

- Permite segmentar frases de forma eficaz:
  - pelo menos na língua portuguesa e inglesa;
  - algoritmo mais rápido do que chunkers habituais;
- Segmentação com alguma informação sintáctica:
  - contrapor com os exemplos obtidos de forma "ad-hoc";
- Estruturalmente ricos:
  - permitem a extracção de sub-relacionamentos de forma simples;
- Possibilidade de colaboração com peritos na área...

- A lista inglesa: foi oferecida por Andy Way (MaTrEx);
- A lista portuguesa: foi trabalho de estágio de um aluno (LEA);

most	maior; maioria
much	muito
my	meu; minha; meus; minhas
near; nearby	perto; próximo; quase
neither	tão-pouco; também não
next	seguinte; próximo; próxima
nigh	próximo
now	agora; uma vez que; considerando que
of	de; por; em
on	em; sobre; em cima de; de; relativa
once	desde que; uma vez que; se
one	um; uma
only	apenas; todavia; mas; contudo
or	ou; se não
other	outro; outra; outras; outros
our	nosso; nossa; nossos; nossas
ours	o nosso; a nossa; os nossos; as nossas
over	sobre; em cima de; por cima de
owing to	devido a: por consequência de; por causa de
own	próprio; ser proprietário
past	por; para além disso; fora de
per	por; através de; por meio de; devido a acção de
such	este; esse; aquele; isto; aquilo
that	aquele; aquela; aquilo; esse; essa; isso; . . .
the	o; a; os; as

Hipótese das Palavras-Marca — *exemplos de segmentos*

34 137	da	comissão
17 277	do	conselho
16 891	da	união europeia
11 379	em	matéria
9 880	de	trabalho
9 850	da	união
9 479	no	sentido
8 465	da	europa
8 454	da	ue
8 004	do	parlamento
5 332	em primeiro	lugar
3 245	no que	diz respeito
2 214	para o	desenvolvimento
total de 3 070 398 segmentos		

13 566	and	gentlemen
11 466	the	commission
11 079	in	order
9 182	to	make
8 712	to	be
8 356	to	do
7 992	of the	european union
7 941	of the	committee
7 814	to	say
7 574	with	regard
7 814	to	say
7 574	with	regard
3 749	in the	european union
total de 3 103 797 segmentos		

EuroParl

815815	de
557697	,
468409	a
352064	da
297634	do
232629	e
197922	que
196801	o
178537	em
156299	dos
[...]	
35394	para a
33079	que o
32213	de um
31539	nos
31492	muito
30805	às
Total de 243 242 marcas	

541197	to
471332	the
440903	of
400417	,
370161	and
252298	of the
214191	in
152164	a
131225	in the
112446	for
105992	that
92180	on
91033	to the
78264	we
70578	on the
67805	this
65092	that the
Total de 198 050 marcas	

this decision shall take effect on 16 september 1999.
a presente decisão produz efeitos em 16 de setembro de 1999.

this decision shall take effect on 16 september 1999.
a presente decisão produz efeitos em 16 de setembro de 1999.

(this decision shall take effect) (on 16 september 1999.)
(a presente decisão produz efeitos) (em 16) (de setembro) (de 1999.)

Segmentação Bilingue — *exemplo*

this decision shall take effect on 16 september 1999.
a presente decisão produz efeitos em 16 de setembro de 1999.

this decision shall take effect on 16 september 1999.
a presente decisão produz efeitos em 16 de setembro de 1999.

(this decision shall take effect) (on 16 september 1999.)
(a presente decisão produz efeitos) (em 16) (de setembro) (de 1999.)

this decision shall take effect on 16 september 1999.
a presente decisão produz efeitos em 16 de setembro de 1999.

this decision shall take effect on 16 september 1999.
a presente decisão produz efeitos em 16 de setembro de 1999.

(this decision shall take effect) (on 16 september 1999.)
(a presente decisão produz efeitos) (em 16) (de setembro) (de 1999.)

Número de segmentos não é necessáriamente o mesmo!

⇓

É necessário alinhar segmentos!

⇓

Tirar partido do trabalho já realizado!

⇓

Dicionários Probabilísticos de Tradução!

Número de segmentos não é necessáriamente o mesmo!

⇓

É necessário alinhar segmentos!

⇓

Tirar partido do trabalho já realizado!

⇓

Dicionários Probabilísticos de Tradução!

Número de segmentos não é necessáriamente o mesmo!

⇓

É necessário alinhar segmentos!

⇓

Tirar partido do trabalho já realizado!

⇓

Dicionários Probabilísticos de Tradução!

Número de segmentos não é necessáriamente o mesmo!

⇓

É necessário alinhar segmentos!

⇓

Tirar partido do trabalho já realizado!

⇓

Dicionários Probabilísticos de Tradução!

	this decision shall take effect	on 16 september 1999
a presente decisão produz efeitos	$a\%$	$b\%$
em 16	$c\%$	$d\%$
de setembro	$e\%$	$f\%$
de 1999	$g\%$	$h\%$

Como calcular as probabilidades de tradução?

$\Downarrow$

A probabilidade de s_α e s_β serem traduções mútuas?

$\Downarrow$

A probabilidade de a tradução de s_β estar contida em s_α.
(com $s_\alpha > s_\beta$)

	this decision shall take effect	on 16 september 1999
a presente decisão produz efeitos	$a\%$	$b\%$
em 16	$c\%$	$d\%$
de setembro	$e\%$	$f\%$
de 1999	$g\%$	$h\%$

Como calcular as probabilidades de tradução?

$\Downarrow$

A probabilidade de s_α e s_β serem traduções mútuas?

$\Downarrow$

A probabilidade de a tradução de s_β estar contida em s_α.
(com $s_\alpha > s_\beta$)

	this decision shall take effect	on 16 september 1999
a presente decisão produz efeitos	$a\%$	$b\%$
em 16	$c\%$	$d\%$
de setembro	$e\%$	$f\%$
de 1999	$g\%$	$h\%$

Como calcular as probabilidades de tradução?

$\Downarrow$

A probabilidade de s_α e s_β serem traduções mútuas?

$\Downarrow$

A probabilidade de a tradução de s_β estar contida em s_α.
(com $s_\alpha > s_\beta$)

Sejam s_α e s_β dois segmentos tal que $s_\alpha < s_\beta$.

Data: Sejam s_α e s_β dois segmentos, na língua $\mathcal{L}_\alpha$ e $\mathcal{L}_\beta$ respectivamente, tal que $s_\alpha < s_\beta$ e, $\mathcal{D}_{\alpha,\beta}$ o dicionário probabilístico de tradução entre essas línguas.

```
1  function quality(Dic, Set_1, Set_2)
2      Soma ⟵ 0
3      for w_α ∈ Set_1 do
4          Trads_{w_α} ⟵ T_dic(w_α)
5          for w_β ∈ Trads_{w_α} do
6              if w_β ∈ Set_2 then
7                  Soma ⟵ Soma + P(w_β ∈ Trads_{w_α})
8      return Soma / size(Set_1)
9  end
10 ProbMarcas ⟵ quality(D_{α,β}, marcas(s_α), marcas(s_β))
11 ProbTexto ⟵ quality(D_{α,β}, texto(s_α), texto(s_β))
12 Prob ⟵ 0.1 × ProbMarcas + 0.9 × ProbTexto
```

	this decision shall take effect	on 16 september 1999
a presente decisão produz efeitos	23.18%	5.86%
em 16	0.00%	76.41%
de setembro	0.00%	85.60%
de 1999	0.00%	84.10%

a presente decisão produz efeitos
this decision shall take effect

em 16 de setembro de 1999
on 16 september 1999

Resultados (1:1)

```
36883  senhor presidente ==1:1==  mr president
 8633  senhora presidente ==1:1==  madam president
 3152  espero ==1:1==  i hope
 2931  gostaria ==1:1==  i would like
 2572  o debate ==1:1==  the debate
 2511  penso ==1:1==  i think
 2356  está encerrado ==1:1==  is closed
 1939  penso ==1:1==  i believe
 1932  muito obrigado ==1:1==  thank
 1852  em segundo lugar ==1:1==  secondly
 1808  gostaria ==1:1==  i should like
 1638  ) senhor presidente ==1:1==  mr president
 1524  há ==1:1==  there
 1423  infelizmente ==1:1==  unfortunately
 1346  creio ==1:1==  i believe
 1257  estou ==1:1==  i
 1249  finalmente ==1:1==  finally
 1210  a votação terá lugar amanhã ==1:1==
                        the vote will take place tomorrow
 1193  em terceiro lugar ==1:1==  thirdly
 1104  ( aplausos ==1:1==  ( applause
 1069  e senhores deputados ==1:1==  and gentlemen
 1067  em primeiro lugar ==1:1==  firstly
 1021  ( o parlamento aprova ==1:1==  ( parliament adopted
  926  na europa ==1:1==  in europe
```

Resultados (1:2)

```
602  , caros colegas ==1:2==  , commissioner and gentlemen
252  caros colegas ==1:2==  ladies and gentlemen
170  , senhor comissário ==1:2==  you very much , commissioner
147  senhores deputados ==1:2==  ladies and gentlemen
143  devo dizer ==1:2==  i have to say
142  lamento ==1:2==  i am sorry
105  congratulo-me ==1:2==  i am pleased
 95  estou convencido ==1:2==  i am convinced
 90  vamos agora proceder ==1:2==  we shall now proceed
 90  e senhores deputados ==1:2==  ladies and gentlemen
 90  agradeço ==1:2==  i am grateful
 85  , senhoras ==1:2==  , commissioner , ladies
 82  , senhores deputados ==1:2==  , commissioner and gentlemen
 79  e outros , em nome ==1:2==  and others , on behalf
 76  refiro-me ==1:2==  i am referring
 72  muito obrigado ==1:2==  thank you very
 71  congratulo-me ==1:2==  i am glad
 70  passamos agora ==1:2==  we shall now proceed
 66  não há dúvida ==1:2==  there is no doubt
 62  , senhora comissária ==1:2==  you very much , commissioner
 61  a votação terá lugar quinta-feira ==1:2==
                          the vote will take place on thursday
```

Resultados (2:1)

```
986  segue-se na ordem ==2:1==  the next item
324  ( a sessão , suspensa ==2:1==  ( the sitting was suspended
230  ( o presidente retira a palavra ==2:1==  ( the president cut
222  ( a sessão é suspensa ==2:1==  ( the sitting was closed
187  senhor presidente , senhor presidente ==2:1==  mr president
169  senhor presidente em exercício ==2:1==  mr president-in-office
148  da sessão de ontem ==2:1==  of yesterday 's sitting
142  ( o parlamento aprova a acta ==2:1==  ( the minutes were approved
138  dos assuntos económicos e monetários ==2:1==  and monetary affairs
113  a proposta da comissão ==2:1==  the commission 's proposal
113  a proposta da comissão ==2:1==  the commission proposal
106  período de perguntas ==2:1==  question time
101  , em nome , sobre a proposta ==2:1==  , on behalf
100  dos direitos do homem ==2:1==  of human rights
 84  dos direitos da mulher ==2:1==  on women 's rights
 72  da direita do hemiciclo ==2:1==  from the right
 67  por interrompida do parlamento europeu ==2:1==
                  of the european parliament adjourned
 67  É muito importante ==2:1==  it is very important
 67  da comissão da comissão ==2:1==  of the committee
 64  estamos a falar ==2:1==  we are talking
```

Resultados (outros...)

```
363  segue-se na ordem a discussão conjunta ==3:1==
                                          the next item
 83  ( o presidente retira a palavra à oradora ==3:1==
                                       ( the president cut
 59  segue-se na ordem do dia ==3:1==  the next item
 42  que recebi de resolução , apresentadas ==3:1==
                                        have received
 39  de aplicação do processo de urgência ==3:1==
                                     for urgent procedure
 36  , de pé um minuto de silêncio ==3:1==
                                      a minute 's silence
 32  está encerrado o período de perguntas ==3:1==
                           that concludes question time
 31  nos termos do artigo 37 º do regimento ==3:1==
                                      pursuant to rule 37
 30  ( a sessão , suspensa às 15h00 ==3:1==
                                  ( the sitting was suspended
 29  segue-se na ordem o período ==3:1==  the next item
 28  está encerrado o período de votações ==3:1==
                               that concludes voting time
 26  está encerrado o período de votação ==3:1==
```

- Hipótese das Palavras-Marca permite segmentação do português e do inglês em segmentos lógicos;
- É possível o alinhamento destes segmentos para extracção de relacionamentos bilingues;
- Mais de 4 milhões de segmentos 1:1 extraídos;
- Pouco mais de 2 milhões destes segmentos são diferentes;
- Um bug detectado leva a que:
  - a qualidade baixe
    (exemplos com segmentos repetidos numa das línguas);
  - o número de segmentos diferentes aumente;

Trabalho Futuro

- Corrigir bug detectado ontem!!
- Associar categoria/tipo a cada marca.
- Agrupar exemplos por marcas e tipo de marcas:
  - generalização de exemplos;
  - extracção de terminologia/nominais;
- Avaliação dos exemplos por amostragem;
- Comparar o uso deste tipo de exemplos com os extraídos com base em âncoras;