Examples Extraction for Machine Translation

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

Workshop on Language Resources for Teaching and Research

April 23, 2008

What Kind of Resources?

- **Probabilistic Translation Dictionaries**
  Dictionaries obtained automatically from parallel corpora;
- **Translation Dictionaries**
  Filtered Probabilistic Translation Dictionaries;
- **Translation Examples**
  Pairs of mutual translations for word segments;
- **Bilingual Terminology**
  Pairs of mutual translation nominals;
- **Translation Rules**
  Parametric translation rules;

Where to extract resources from? Parallel 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

Number of translation units per corpus.

- ⋆ created by Natura Project;
- ⋆⋆ created by Linguateca and used as a comparison point.

Parallel Corpora Sizes Comparison

Corpus		Tokens		Types	
		Source	Target	Source	Target
Constituição	PT:EN	38 024	40 984	3 761	3 113
	PT:ES	38 024	41 855	3 761	3 817
	PT:FR	38 024	42 484	3 761	3 916
Compara	PT:EN	1 714 049	1 797 976	71 759	45 429
L.M.D.	FR:PT	1 730 166	1 887 250	66 950	59 009
JRC-Acquis	PT:EN	8 248 333	7 797 133	68 325	55 797
	PT:ES	8 005 805	8 333 518	67 314	64 471
	PT:FR	7 934 385	8 134 116	66 939	59 453
EuroParl	PT:EN	29 232 417	28 366 649	137 607	87 511
	PT:ES	29 331 905	29 736 743	142 189	135 126
	PT:FR	29 826 035	33 286 644	148 259	108 356
EurLex	PT:EN	226 600 339	213 832 551	658 601	608 921
	PT:ES	22 904 057	23 724 321	161 804	158 942
	PT:FR	36 589 842	39 799 740	206 467	184 405

Part I

Probabilistic Translation Dictionaries

What are PTD?

- translation dictionaries;
- translation probabilities;
- automatically extracted from sentence-aligned PC;
- map words w_A from the language A to a set of words w_B on B language; de palavras w_B da língua B;
- for most words, $\mathcal{T}(w_A) = w_B$;
- but in some cases, $\mathcal{T}(w_A) = \neg w_B$;
- and other strange results as well.

PTD Examples (1)

```
QUERY> europa
  Occurrences: 39917
  Translations:
      88.50%  europe
       5.73%  european
       2.37%  europa
       0.57%  EU
       0.23%  unece
       0.17%  the
       0.16%  auto
```


PTD Examples (2)

```
QUERY> we
  Occurrences: 300431
  Translations:
      17.81%  (none)
       8.25%  que
       6.02%  temos
```

PTD Examples (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
```


PTD Examples (4)

```
QUERY> aceitável
  Occurrences: 1713
  Translations:
      71.48%  acceptable
       8.56%  unacceptable
```


PTD Examples (5)

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

Floor?? What the...

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 .

PTD Examples (5)

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

Floor?? What the...

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 .

Alberto Simões | Examples Extraction for Machine Translation

PTD Examples (5)

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

Floor?? What the...

```
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 .
```


PTD Extraction Algorithm

a flor cresce / a casa é grande / a casa azul tem flores

the flower grows / the house is big / the blue house has flowers

	a	flor	cresce	casa	é	grande	azul	tem	flores
the	3	1	1	2	1	1	1	1	1
flower	1	1	1	0	0	0	0	0	0
grows	1	1	1	0	0	0	0	0	0
house	2	0	0	2	1	1	1	1	1
is	1	0	0	1	1	1	0	0	0
big	1	0	0	1	1	1	0	0	0
blue	1	0	0	1	0	0	1	1	1
have	1	0	0	1	0	0	1	1	1
flowers	1	0	0	1	0	0	1	1	1

PTD Extraction Algorithm

a flor cresce / a casa é grande / a casa azul tem flores

the flower grows / the house is big / the blue house have flowers

	a	flor	cresce	casa	é	grande	azul	tem	flores
the	3	1	1	2	1	1	1	1	1
flower	1	1	1	0	0	0	0	0	0
grows	1	1	1	0	0	0	0	0	0
house	2	0	0	2	1	1	1	1	1
is	1	0	0	1	1	1	0	0	0
big	1	0	0	1	1	1	0	0	0
blue	1	0	0	1	0	0	1	1	1
have	1	0	0	1	0	0	1	1	1
flowers	1	0	0	1	0	0	1	1	1

PTD Extraction Algorithm

a flor cresce / a casa é grande / a casa azul tem flores

the flower grows / the house is big / the blue house have flowers

	a	flor	cresce	casa	é	grande	azul	tem	flores
the	3	1	1	2	1	1	1	1	1
flower	1	1	1	0	0	0	0	0	0
grows	1	1	1	0	0	0	0	0	0
house	2	0	0	2	1	1	1	1	1
is	1	0	0	1	1	1	0	0	0
big	1	0	0	1	1	1	0	0	0
blue	1	0	0	1	0	0	1	1	1
have	1	0	0	1	0	0	1	1	1
flowers	1	0	0	1	0	0	1	1	1

a	the	100%				
casa	house	100%				
flor	flower	50%	grows	50%		
cresce	flower	50%	grows	50%		
é	is	50%	big	50%		
grande	is	50%	big	50%		
azul	blue	33%	has	33%	flowers	33%
tem	blue	33%	has	33%	flowers	33%
flores	blue	33%	has	33%	flowers	33%

```
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
```

```
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
```


a	the	100%				
casa	house	100%				
flor	flower	50%	grows	50%		
cresce	flower	50%	grows	50%		
é	is	50%	big	50%		
grande	is	50%	big	50%		
azul	blue	33%	has	33%	flowers	33%
tem	blue	33%	has	33%	flowers	33%
flores	blue	33%	has	33%	flowers	33%

```
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
```

```
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
```


Part II

Translation Dictionaries

- filter (cut entries):
  - remove numbers;
  - remove non-words;
  - remove low probability translations;
  - remove low occurring words;
  - remove the "empty" translation;
  - remove empty entries;
- or just search for entries such that:

$$a \in \mathcal{T}_{B \to A}\left(\mathcal{T}_{A \to B}\left(a\right)\right)$$

Part III

Translation Examples

Translation Examples: What?

- Translation Units can be useful to translators;
- but have low reusability;
- Translation Examples are small Translation Units: Word Segments and their translation.
- being smaller, have higher reusability;
- Two approaches:
  - Segmentation based on Marker Hypothesis;
  - Segmentation based on the Translation Matrix main diagonal;

Marker Hypothesis in One Slide

Natural Language uses words from closed classes (like pronouns, prepositions or adverbs) to delimit sentences phrases. (Green 1979)

Esta norma foi objecto de diversas alterações nos dois últimos anos , pelo que é oportuno actualizar consequentemente a norma comunitária .

Marker Hypothesis in One Slide

Natural Language uses words from closed classes (like pronouns, prepositions or adverbs) to delimit sentences phrases. (Green 1979)

Esta norma foi objecto de diversas alterações nos dois últimos anos , pelo que é oportuno actualizar consequentemente a norma comunitária .

Marker Hypothesis: Word List

- English List: offered by Andy Way (MaTrEx Project);
- Portuguese List: constructed upon the English List by Luís Gomes;

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

Marker Hypothesis on two languages

- the number of segments it not necessarilly the same between languages;
- we need to align (make correspondencies) between sgments;
- use the available resources: probabilistic translation dictionaries;

	this decision shall take effect	as soon	as possible
a presente decisão produz efeitos	23.18%	5.86%	7.93%
o mais rapidamente possível	0.00%	76.41%	83.10%

$$\mathcal{P}\left(\mathcal{T}\left(s_\beta\right) \subseteq s_\alpha\right) \qquad s_\beta \leq s_\alpha$$

Marker Hypothesis — Some results

36902	senhor presidente	mr president
3527	da união europeia	of the european union
3149	espero	i hope
2995	da comissão	of the committee
2942	gostaria	i would like
2911	na europa	in europe
2771	o debate	the debate
2515	penso	i think
2366	da comissão	of the commission
2356	está encerrado	is closed
2269	do conselho	of the council
1937	penso	i believe
1885	muito obrigado	thank
1882	em segundo lugar	secondly
1791	a favor	in favour
1734	da união	of the union
1489	do meio ambiente	on the environment
1442	a comissão	the commission
1423	infelizmente	unfortunately
1345	creio	i believe
1287	à comissão	the commission
1278	em nome	on behalf
1277	do tratado	of the treaty
1235	do parlamento europeu	of the european parliament
1210	a votação terá lugar amanhã	the vote will take place tomorrow

Marker Hypothesis — Some results

377	caros colegas	commissioner and gentlemen
254	caros colegas	ladies and gentlemen
221	e outros , em nome	and others , on behalf
153	o mais rapidamente possível	as quickly as possible
149	senhores deputados	ladies and gentlemen
147	o mais rapidamente possível	as soon as possible
146	devo dizer	i have to say
94	e senhores deputados	ladies and gentlemen
90	vamos agora proceder	we shall now proceed
70	passamos agora	we shall now proceed
65	não há dúvida	there is no doubt
61	a votação terá lugar quinta-feira	the vote will take place on thursday
57	pergunta n ° 4	question no 4

Marker Hypothesis — Some results

305	dos direitos do homem	of human rights
296	da comissão da comissão	of the committee
242	dos direitos da mulher	on women 's rights
193	a proposta da comissão	the commission 's proposal
186	dos direitos do homem	on human rights
171	a proposta da comissão	the commission proposal
167	senhor presidente em exercício	mr president-in-office
163	e da política de defesa	and defence policy
151	da sessão de ontem	of yesterday 's sitting
142	o parlamento aprova a acta	the minutes were approved
135	e da política do consumidor	and consumer policy
128	de decisão do conselho	for a council decision
106	período de perguntas	question time
102	dos direitos do homem	for human rights
101	da pena de morte	of the death penalty
98	na irlanda do norte	in northern ireland
98	e da defesa do consumidor	and consumer protection
95	do tratado de amesterdão	of the amsterdam treaty

Translation Matrix: Why and How?

Create segments from a Translation Unit by:

- create a matrix and probabilities of translation between each word pair;
- consider cells with higher probabilities as anchors;
- use the assumption that translation is done from left to right;
- use the assumption that correct translation relationships are on the main diagonal of the matrix.

Translation Matrix

	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

discussion about	---	discussão sobre
sources of	---	fontes de
of financing	---	de financiamento
sources of financing	---	fontes de financiamento

Translation Matrix and Translation Patterns

As we know, translation order changes,

- but most cases are easy to predict, as they are closely related to grammar details;
- thus, it is possible to formalize these changes using patterns;
- these patterns can be used to help extracting examples;
- these patterns are also a good source of bilingual terminology.

Pattern Example 1: ABBA

	Jogos	Olímpicos
Olympic		X
Games	X	

Written in our DSL as:

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

Formally...

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

Pattern Example 2: IDH

	índice	de	desenvolvimento	humano
human				X
development			X	
index	X			

Written as:

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

Formally...

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

Pattern Example 3: FTP

	protocolo	de	transferência	de	ficheiros
file					X
transfer			X		
protocol	X				

Written as:

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

Formally...

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

Pattern Example 4: NPoV

	ponto	de	vista	neutro
neutral				X
point	X			
of		Δ		
view			X	

Written as:

```
[POV] P "de" V N = N P "of" V
```

Formally...

$$\mathcal{T}(P \cdot "de" \cdot V \cdot N) = \mathcal{T}(N) \cdot \mathcal{T}(P) \cdot "of" \cdot \mathcal{T}(V)$$

Translation Matrix: Using Patterns

	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

```
discussion about                  --- discussão sobre
alternative sources of financing  --- fontes de financiamento alternativas
for the                           --- para a
european radical alliance         --- aliança radical europeia
```

`for the european radical alliance --- para a aliança radical europeia`

Part IV

Bilingual Terminology

Terminology (noun phrases) extraction

If we specify alignment patterns with some care, it is possible to extract automatically high quality bilingual terminology (noun phrases).

What kind of Care?

Terminology (noun phrases) extraction

If we specify alignment patterns with some care, it is possible to extract automatically high quality bilingual terminology (noun phrases).

What kind of Care?

Patterns Languages and Restrictions

Morphological Restrictions

Specify the morphological properties we are expecting on each place-holder:

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

Generic Restrictions

Define Perl functions that check if the rule should be applied.

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

Infer word characteristics

Use these predicates to infer, instead of just restrict applicability.

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


Patterns Languages and Restrictions

Morphological Restrictions

Specify the morphological properties we are expecting on each place-holder:

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

Generic Restrictions

Define Perl functions that check if the rule should be applied.

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

Infer word characteristics

Use these predicates to infer, instead of just restrict applicability.

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

Alberto Simões | Examples Extraction for Machine Translation

Patterns Languages and Restrictions

Morphological Restrictions

Specify the morphological properties we are expecting on each place-holder:

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

Generic Restrictions

Define Perl functions that check if the rule should be applied.

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

Infer word characteristics

Use these predicates to infer, instead of just restrict applicability.

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

Alberto Simões | Examples Extraction for Machine Translation

Terminology Counting

```
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
```

Alberto Simões | Examples Extraction for Machine Translation

Terminology Examples: 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
```


Terminology Examples: 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
```


Terminology Examples: 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
```


Terminology Examples: 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
```

Terminology Examples: P "de" V N = N P "of" V

93 tribunal de justiça europeu => european court of justice
51 tribunal de contas europeu => european court of auditors
33 fontes de energia renováveis => renewable sources of energy
27 ponto de vista ambiental => environmental point of view
26 ponto de vista económico => economic point of view
21 ponto de vista jurídico => legal point of view
20 declaração de fiabilidade positiva => positive statement of assuranc
18 ponto de vista político => political point of view
13 ponto de vista técnico => technical point of view
10 ponto de vista institucional => institutional point of view
9 ponto de vista orçamental => budgetary point of view
8 sistema de preferências generalizadas => generalised system of prefe
8 método de coordenação aberto => open method of coordination
7 ponto de vista social => social point of view
7 ponto de vista democrático => democratic point of view

Terminology Examples: P "de" T "de" F = F T P

41 emissões de dióxido de carbono => carbon dioxide emissions
22 sistema de informação de schengen => schengen information system
8 sistema de comércio de emissões => emissions trading system
8 plano de acção de viena => vienna action plan
8 cartão de prestação de serviços => service provision card
8 agenda de desenvolvimento de doha => doha development agenda
7 política de espectro de radiofrequências => radio spectrum policy
6 sistema de transporte de mercadorias => freight transport system
6 dispositivos de limitação de velocidade => speed limitation devices
5 plataforma de acção de pequim => beijing action platform
5 operações de gestão de crises => crisis management operations
5 critérios de convergência de maastricht => maastricht convergence cr
4 política de mercado de trabalho => labour market policy
4 normas de protecção de dados => data protection rules
4 grupo de trabalho de alto => high-level working group

Terminology Extraction - Quick and Dirty evaluation

- EuroParl PT-EN: 1 000 000 TUs
- 700 000 TUs processed
- 578 103 pattern occurrences
- 139 781 different examples
- 103 617 examples after filtering (removal of stop words, noise,..)
- 77 497 examples by A B = B A (938/2/1) (86%)
- 12 694 examples by A "de" B = B A (204/2/1) (95%)
- 7 700 examples by A B C = C B A (40/1/1) (93%)
- 3 336 examples by H "de" D H = H D I (21/1/1) (100%)
- 564 examples by P "de" V N = N P "of" V (6/1/1) (98%)
- 360 examples by P "de" T "de" F = F T P (3/1/1) (96%)

Part V

Translation Rules

Simple translation rules for non-words

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


Simples translation rules with Word Classes

```
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
```

Generalize and create translation rules:

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


Word Classes Creation

Fix a word (noun, for instance) on an alignment rule, and cycle all words that co-occur with it:

```
'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)'
  ...
```


Create rules automatically

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

where the class X.acidClass includes pairs:

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

And more generalization...

- noun/adjective swap:
  - Rules
    - `($w)#a ($w)#sms ==> $2+$1#sms`
    - `($w)#a (#w)#sfp ==> $2+($1#TO#fp)#sfp`
  - So we can translate:
    - abusive aid → auxílio abusivo
    - abusive alteration → alteração abusiva
    - dynamic access → acesso dinâmico
    - dynamic adaptations → adaptações dinâmicas
- prepositional phrases from noun sequences:
  - Rules (simplfied)
    - `($w)#s ($w)#s ==> $2#s+de+$1`
  - So we can translate:
    - embarkation areas → zonas de embarque
    - embarkation deck → pavimento de embarque
    - abandonment measures → medidas de abandono
    - abandonment programme → programa de abandono

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia

contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia

documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia

contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia

documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia

contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia

documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia
contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia
documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia
contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia
documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Alberto Simões | Examples Extraction for Machine Translation

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia
contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia
documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Alberto Simões | Examples Extraction for Machine Translation

Text::Translator Roadmap

$\mathcal{T}$(accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia
contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia
documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Alberto Simões | Examples Extraction for Machine Translation

Text::Translator Roadmap

$\mathcal{T}$ (accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia
contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia
documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Text::Translator Roadmap

$\mathcal{T}$ (accounting documents of the European Union)

Use examples, translation and probabilistic translation dictionaries

Generate all ambiguous translations.

contabilístico#a documento#s de o#art União Europeia
contabilidade#s documento#s de o#art União Europeia

Use rules to re-order words.

documento contabilístico da União Europeia
documento de contabilidade da União Europeia

Evaluate sentences legibility using language models (n-grams, HMM).

documento contabilístico da União Europeia

Alberto Simões | Examples Extraction for Machine Translation

Future Developments

Use the resources in a translation tool:

- Text::Translator
  - a tool for translation systems prototyping;
  - hybrid system: Rules, SMT and EBMT;
  - Perl based — easy to configure;
- Apertium
  - a rule-based system for translation between close languages;
  - although some good results between distant languages;
  - fully configurable (based on heavy XML syntax);

Alberto Simões | Examples Extraction for Machine Translation

http://natools.sf.net/

http://natura.di.uminho.pt/

Alberto Simões | Examples Extraction for Machine Translation