Automatic Extraction of Translation Resources from Parallel Corpora

Alberto Manuel Simões

ambs@di.uminho.pt

CCTC / Computer Science Department

Motivation

- translation resources are needed for:
  - Computer Assisted Translation (CAT);
  - Machine Translation (MT);
- their creation by hand is expensive;

— BUT —

- parallel corpora are available in more quantity and quality;
  - European Union, multilingual organizations, ...
- they comprise a lot of hidden bilingual information;

— THUS —

- automatic extraction of translation resources is of great importance;

Motivation

- translation resources are needed for:
  - Computer Assisted Translation (CAT);
  - Machine Translation (MT);
- their creation by hand is expensive;

— BUT —

- parallel corpora are available in more quantity and quality;
  - European Union, multilingual organizations, ...
- they comprise a lot of hidden bilingual information;

— THUS —

- automatic extraction of translation resources is of great importance;

Motivation

- translation resources are needed for:
  - Computer Assisted Translation (CAT);
  - Machine Translation (MT);
- their creation by hand is expensive;

— BUT —

- parallel corpora are available in more quantity and quality;
  - European Union, multilingual organizations, ...
- they comprise a lot of hidden bilingual information;

— THUS —

- automatic extraction of translation resources is of great importance;

Parallel Corpora

- parallel corpora are defined as:
  - collections of pairs of texts translated from one language into another;
- a sentence aligned parallel corpora are:
  - parallel corpora where correspondences between sentences where identified;
  - based on a simple statistical model of character lengths of sentences/paragraphs (normally without any lexical clues);
  - each pair of sentences is normally called a translation unit.
- typicall parallel corpora sizes (in TUs):

Corpus	PT:EN	PT:ES	PT:FR
COMPARA	97 215	—	—
Le Monde Diplomatique	—	—	68 231
JRC-Acquis	286 008	281 185	277 754
EuroParl	998 830	1 006 895	1 023 841
EurLex	10 394 893	1 111 068	1 710 760

Parallel Corpora

- parallel corpora are defined as:
  - collections of pairs of texts translated from one language into another;
- a sentence aligned parallel corpora are:
  - parallel corpora where correspondences between sentences where identified;
  - based on a simple statistical model of character lengths of sentences/paragraphs (normally without any lexical clues);
  - each pair of sentences is normally called a translation unit.
- typicall parallel corpora sizes (in TUs):

Corpus	PT:EN	PT:ES	PT:FR
COMPARA	97 215	—	—
Le Monde Diplomatique	—	—	68 231
JRC-Acquis	286 008	281 185	277 754
EuroParl	998 830	1 006 895	1 023 841
EurLex	10 394 893	1 111 068	1 710 760

Parallel Corpora

- parallel corpora are defined as:
  - collections of pairs of texts translated from one language into another;
- a sentence aligned parallel corpora are:
  - parallel corpora where correspondences between sentences where identified;
  - based on a simple statistical model of character lengths of sentences/paragraphs (normally without any lexical clues);
  - each pair of sentences is normally called a translation unit.
- typicall parallel corpora sizes (in TUs):

Corpus	PT:EN	PT:ES	PT:FR
COMPARA	97 215	—	—
Le Monde Diplomatique	—	—	68 231
JRC-Acquis	286 008	281 185	277 754
EuroParl	998 830	1 006 895	1 023 841
EurLex	10 394 893	1 111 068	1 710 760

Sentence Aligned Parallel Corpora

Estes resultados constituem a base do Programa Europeu de defesa do Mar de Barents e, por esse motivo, peço-lhe que analise um projecto de carta que lhe expõe os factos mais importantes, e que, de acordo com as decisões do Parlamento, torne clara esta posição na Rússia.

No entanto, somos também da opinião de que deveria haver um debate sobre esta estratégia da comissão que seguisse um procedimento ordenado, e não só com base numa declaração oral pronunciada aqui no Parlamento Europeu, mas também com base num documento que seja decidido na comissão e que apresente uma descrição deste programa para um período de cinco anos.

These findings form the basis of the European Programmes to protect the Barents Sea, and that is why I would ask you to examine a draft letter setting out the most important facts and to make Parliament's position, as expressed in the resolutions which it has adopted, clear as far as Russia is concerned.

We believe, however, that the commission's strategic plan needs to be debated within a proper procedural framework, not only on the basis of an oral statement here in the European Parliament, but also on the basis of a document which is adopted in the commission and which describes this programme over the five-year period .

Translation Resources extraction

Estes resultados constituem a base do Programa Europeu de defesa do Mar de Barents e, por esse motivo, peço-lhe que analise um projecto de carta que lhe expõe os factos mais importantes, e que, de acordo com as decisões do Parlamento, torne clara esta posição na Rússia.

No entanto, somos também da opinião de que deveria haver um debate sobre esta estratégia da comissão que seguisse um procedimento ordenado, e não só com base numa declaração oral pronunciada aqui no Parlamento Europeu, mas também com base num documento que seja decidido na comissão e que apresente uma descrição deste programa para um período de cinco anos.

These findings form the basis of the European Programmes to protect the Barents Sea, and that is why I would ask you to examine a draft letter setting out the most important facts and to make Parliament's position, as expressed in the resolutions which it has adopted, clear as far as Russia is concerned.

We believe, however, that the commission's strategic plan needs to be debated within a proper procedural framework, not only on the basis of an oral statement here in the European Parliament, but also on the basis of a document which is adopted in the commission and which describes this programme over the five-year period .

Probabilistic Translation Dictionaries Extraction (1)

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

Probabilistic Translation Dictionaries Extraction (2)

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
has	1	0	0	1	0	0	1	1	1
flowers	1	0	0	1	0	0	1	1	1

Probabilistic Translation Dictionaries Extraction (3)

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
has	1	0	0	1	0	0	1	1	1
flowers	1	0	0	1	0	0	1	1	1

Resulting Probabilistic Translation Dictionary

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
```

Resulting Probabilistic Translation Dictionary

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
```

Translation Units Extraction

Create a translation matrix with translation probabilities:

	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**

Extract segments:

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

But translation order is not linear...

Translation Units Extraction

Create a translation matrix with translation probabilities:

	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**

Extract segments:

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

But translation order is not linear...

Translation Units Extraction

Create a translation matrix with translation probabilities:

	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**

Extract segments:

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

But translation order is not linear...

Define Patterns

Define patterns to specify translation order changes:

	Jogos	Olímpicos
Olimpic		X
Games	X	

Described on a compact Domain Specific Language as:

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

Described Formally. . .

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

Define Patterns

Define patterns to specify translation order changes:

	Jogos	Olímpicos
Olimpic		X
Games	X	

Described on a compact Domain Specific Language as:

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

Described Formally. . .

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

Other Pattern Examples

	índice	de	desenvolvimento	humano
human				X
development			X	
index	X			

	protocolo	de	transferência	de	ficheiros
file					X
transfer			X		
protocol	X				

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

Translation Units Extraction with 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
```

And we can even concatenate them:

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

Patterns ⇒ Nominals ⇒ Terminology

```
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
```

Terminology Evaluation

- EuroParl PT-EN: 1 000 000 TUs
- 700 000 Translation Units Processed
- 139 781 Different Examples

Quantity	Pattern	Evaluation
77 497 ex.	A B = B A	86%
12 694 ex.	A ”de” B = B A	95%
7 700 ex.	A B C = C B A	93%
3 336 ex.	H ”de” D H = H D I	100%
1 466 ex.	A B C = C A B	40%
564 ex.	P ”de” V N = N P ”of” V	98%
360 ex.	P ”de” T ”de” F = F T P	96%

Other work

Other scientific issues:

- Extraction of Examples based on the Marker Hypothesis;
- Extraction of word classes based on n-grams, and extracted terminology;

Engineering issues:

- algorithm scalability to Gigabytes of text;
- algorithm parallelism to be used on Clusters;
- package configuration for easy installation;
- client/server architecture for efficiency;