Portuguese-English Word Alignment

Some Experiments

Diana Santos and Alberto Simões

1 Background

There are two kinds of word alignment:

token (word) alignment Among tokens of a parallel corpus

type (word) alignment Among types (after processing parallel corpora)

First, we create bilingual dictionaries (type alignment) [PTDs (probabilistic translation dictionaries] using the NATools suite [SA03] and apply those to parallel corpora to perform token alignment, getting token dictionaries.

We use (as data) two very different parallel corpora:

EuroParl Transcription of European parliament debates, 30 Mwords per language, http://www.statmt.org/europarl/ [Koe05]

COMPARA Fiction text, 1.5 Mwords in each language, syntax annotation, http://www.linguateca.pt/COMPARA/ [FGS03, SFG07]

creating a number of subcorpora, and created PTDs from them, and/or token aligned them. "red" stands for reduced, and "tst" for test.

1.1 Alignment tools

NATools essentials: EM algorithm over aligned corpora, creating PTD (our word-type aligment results):

```
** europe ** 42853 occs

      europa: 94.71 %
    europeus:  3.39 %
     europeu:  0.81 %
    europeia:  0.11 %
```

```
** stupid ** 180 occs

    estúpido:  17.55 %
    estúpida:  10.99 %
   estúpidos:   7.41 %
     avisada:   5.65 %
     direita:   5.58 %
     impasse:   4.48 %
     ocupado:   3.75 %
```

Figure 1: Examples of PTD entries

	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**

Figure 2: Word alignment matrix created by `nat-chunker`

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

Table 1: Extracted aligned pairs (only 1-1 go to token dictionary).

1.2 Comparing dictionaries

Some measures provided by `nat-compareDicts` and `nat-descDict`:

- size of token and type dictionaries
- translation fertility: (average) number of different translations in a type dictionary
- alignment density of a token alignment (by types or forms)
- average size of word correspondences after token alignment

2 Research questions

2.1 How relevant is the PTD source?

How do the issues of genre, and size, influence the alignment results? How dependent is an English-Portuguese dictionary from the material it was compiled from?

	COMPARA		EuroParl	
PT	17 433	12 348	17 486	6 570
EN	12 271	9 371	10 927	4 853

Table 2: Size of token dictionaries created after alignment: with dictionaries based on same and other genre, trained on same size corpora.

2.2 Assessing dictionary quality

We used lists of words from a given semantic domain (colour) to assess that domain's preservation.

	Best	Any	Best	Any
Port	172/382 (45%)	271/382 (71%)	47/112 (42%)	70/112 (62%)
Eng	115/455 (25%)	162/455 (36%)	24/96 (25%)	37/96 (39%)

Table 3: Colour correspondences in PTDs

2.3 Is translation direction relevant?

Does it matter, whether the texts have been translated from English or into English? How symmetrical are word (type,token) correspondences?

PTD	language	EtoP	PtoE
same	PT	30 276	20 019
diff	PT	19 798	10 778
same	EN	19 646	12 960
diff	EN	14 279	7 853
full CMP	PT	39 339	38 688
full CMP	EN	25 101	22 429
EuroParl	PT	21 817	19 673
EuroParl	EN	16 124	13 952

Table 4: Size of token dictionaries for the two sections of COMPARA, aligned with several different PTDs.

2.4 Is syntactic analysis relevant?

What is the impact of using lemmatized corpora, PoS-annotated corpora, and/or corpora where sequential multiwords have been connected? How to deal with different morphological richness in the two languages?

Corpus	PT-EN				EN-PT			
	Size	TF	tyAD	toAD	Size	TF	tyAD	toAD
COMPARA	57 198	1.82	1.45	0.03	33 074	2.19	1.59	0.02
CMPred	55 584	2.70	2.20	0.04	32 548	3.97	3.45	0.03
CMPtst	17 433	1.84	1.48	0.10	12 271	2.41	1.82	0.07
EuroParl	115 327	6.97	5.84	0.00	68 090	10.45	8.13	0.00
EuroParlred	115 365	6.93	5.82	0.00	67 421	10.49	8.11	0.00
EuroParlred2	41 551	4.35	3.75	0.02	22 358	6.46	4.86	0.01
EuroParltst	17 486	2.48	2.37	0.06	10 927	3.02	2.64	0.04
EtoP	30 276	2.36	1.47	0.03	19 646	3.16	1.85	0.02
PtoE	20 019	1.83	0.76	0.03	12 960	2.62	1.14	0.02
CMPmwe	60 480	2.73	2.24	0.04	34 209	4.13	3.11	0.02
CMPmweprop	62 681	2.67	2.19	0.04	34 363	4.25	3.22	0.02
CMPlemma	27 348	4.25	3.59	0.02	33 196	3.62	2.64	0.02
CMPpos	30 903	3.90	3.24	0.02	33 087	3.72	2.71	0.02

Table 5: Size, translation fertility and alignment density of the several token dictionaries.

References

[FGS03] Ana Frankenberg-Garcia and Diana Santos. Introducing COMPARA, the Portuguese-English parallel translation corpus. In Federico Zanettin, Silvia Bernardini, and Dominic Stewart, editors, *Corpora in Translation Education*, pages 71–87. St. Jerome Publishing, 2003.

[Koe05] Philipp Koehn. Europarl: a parallel corpus for statistical machine translation. In *MT Summit*, 2005.

[SA03] Alberto M. Simões and J. João Almeida. Natools – a statistical word aligner workbench. *SEPLN*, 31:217–224, Sep. 2003.

[SFG07] Diana Santos and Ana Frankenberg-Garcia. The corpus, its users and their needs: a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12(3):335–374, 2007.