PTD Algebra

Alberto Simões <ambs@ilch.uminho.pt>
José João Almeida <jj@di.uminho.pt>

November 23rd, 2011

What are PTDs?

- PTD stands for Probabilistic Translation Dictionary;
- Can be seen as a common translation dictionary, but. . .
- . . . is extracted automatically from parallel corpora;
- . . . for each word suggests a set of possible translations;
- . . . together with each translation presents a certainty level;
- . . . does not present just translations, but also "related words;"
- Can be extracted by some different tools:
  - NATools (**?**);
  - GIZA++ (**?**);

```
europe =>
    count => 42853,
    trans =>
        europa   => 94.7 %
        europeus =>  3.4 %
        europeu  =>  0.8 %
        europeia =>  0.1 %
```

```
stupid =>
    count => 180,
    trans =>
        estúpido  => 47.6 %
        estúpida  => 11.0 %
        estúpidos =>  7.4 %
        avisada   =>  5.6 %
        direita   =>  5.6 %
        impasse   =>  4.5 %
        ocupado   =>  3.8 %
```

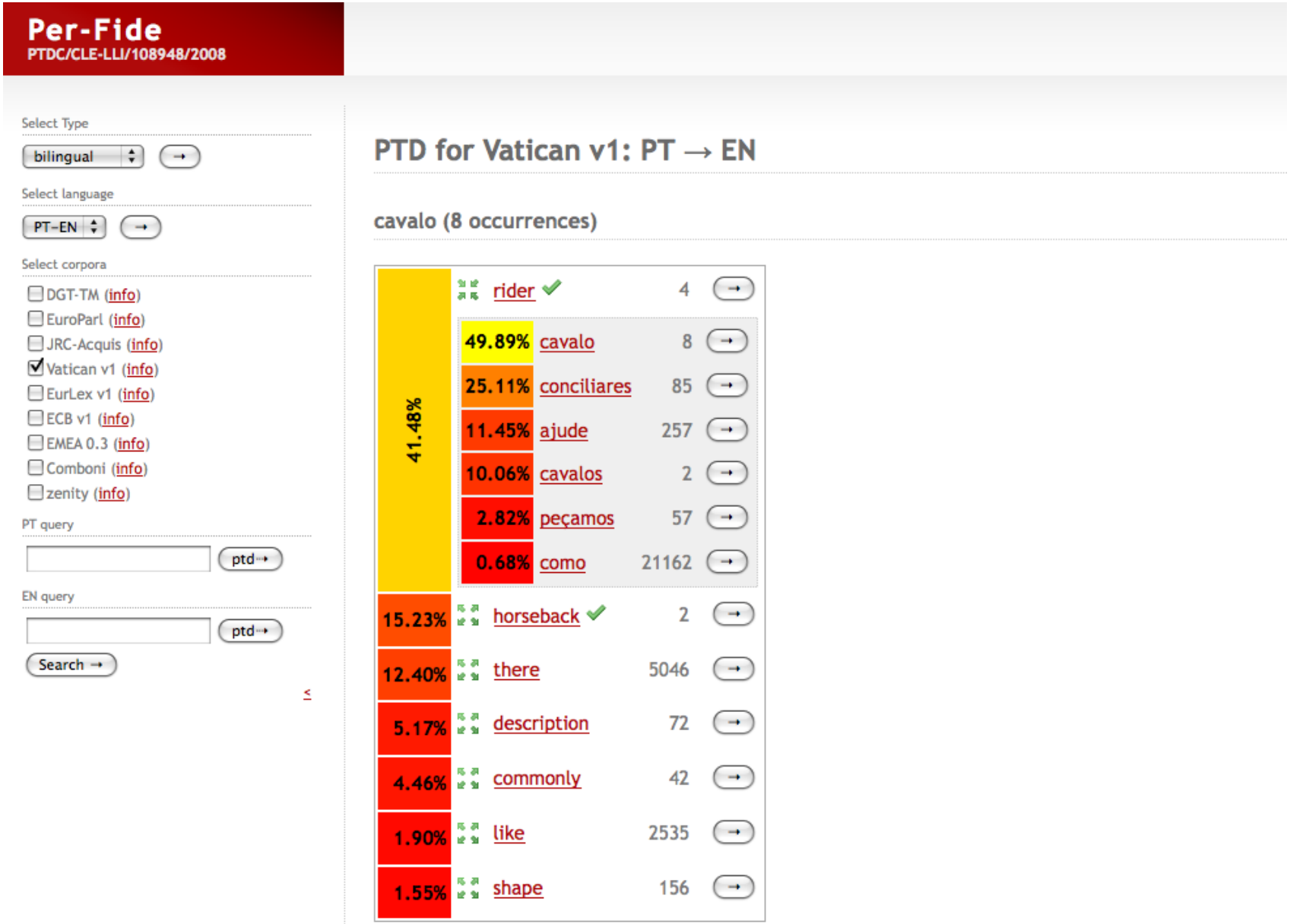
Extraction process

$$align : TU^{\star} \longrightarrow Dic(\mathcal{L}_{\alpha}, \mathcal{L}_{\beta}) \times Dic(\mathcal{L}_{\beta}, \mathcal{L}_{\alpha})$$

$$\begin{array}{ll} TU & \cong sentence_{\mathcal{L}_{\alpha}} \times sentence_{\mathcal{L}_{\beta}} \\ Dic(A, B) & \cong word_{A} \rightharpoonup wordInfo(B) \\ wordInfo(B) & \cong occur \times word_{B} \rightharpoonup probabilidade \end{array}$$

Dictionary structure

$$word_{\mathcal{L}_{\alpha}} \mapsto (occurrences \times word_{\mathcal{L}_{\beta}} \mapsto \mathcal{P}(\mathcal{T}(word_{\mathcal{L}_{\alpha}}) = word_{\mathcal{L}_{\beta}}))$$


Per-Fide
PTDC/CLE-LLI/108948/2008
Select Type
bilingual
Select language
PT–EN
Select corpora
DGT-TM (info)
EuroParl (info)
JRC-Acquis (info)
Vatican v1 (info)
EurLex v1 (info)
ECB v1 (info)
EMEA 0.3 (info)
Comboni (info)
zenity (info)
PT query
ptd
EN query
ptd
Search
PTD for Vatican v1: PT → EN
cavalo (8 occurrences)
41.48%
rider 4
49.89% cavalo 8
25.11% conciliares 85
11.45% ajude 257
10.06% cavalos 2
2.82% peçamos 57
0.68% como 21162
15.23% horseback 2
12.40% there 5046
5.17% description 72
4.46% commonly 42
1.90% like 2535
1.55% shape 156

There is a list of operations one can perform with PTDs:

Description	Notation
union	$d1 \cup d2$
interception	$d1 \cap d2$
domain restrict	$d1 \mid d2$
domain subtract	$d1 \setminus d2$
PTD composition	$d1 \circ d2$
PTD distance	$distance(d1, d2)$
PTD totalize	$totalize(d1)$
PTD filtering	$filter(d, entry \rightarrow bool)$

PTD Union

- PTD Union sums two or more dictionaries;
- The base of NATools algorithm for scalability (**?**);
- This operation:
  - is applied to dictionaries with the same source and target languages;
  - gives different weights to probabilities accordingly to corpora sizes and word occurrence counts;
  - is applied to multi-sets. This means that

$$D \cup D = 2 \times D,$$

where the multiplication is the duplication of occurrences counts (translations probabilities are kept).

$$\underbrace{\underbrace{\text{doente}}_{(10964)} \begin{cases} \text{patient} & 80.4\ \% \\ \text{patients} & 4.5\ \% \\ \text{card} & 0.6\ \% \\ \text{ill} & 0.6\ \% \\ \text{you} & 0.4\ \% \\ \text{she} & 0.3\ \% \end{cases} \bigcup \underbrace{\text{doente}}_{(404)} \begin{cases} \text{patient} & 24.2\ \% \\ \text{ill} & 18.7\ \% \\ \text{sick} & 17.7\ \% \\ \text{patients} & 10.1\ \% \\ \text{illness} & 2.0\ \% \\ \text{well} & 1.7\ \% \end{cases}}_{=}$$

$$\underbrace{\text{doente}}_{(11368)} \begin{cases} \text{patient} & 79.8\ \% \\ \text{patients} & 4.6\ \% \\ \text{ill} & 0.8\ \% \\ \text{card} & 0.6\ \% \\ \text{you} & 0.4\ \% \\ \text{shee} & 0.3\ \% \end{cases}$$

PTD Intersection

Intersection is a dictionary enhancing mechanism:

- intersects the domain of both dictionaries (removing any word that does not occur in one of the dictionaries);
- uses, for each maintained word, the minimum number of occurrences in the two dictionaries;
- intersects the probable translation sets, maintaining words that are probable translations on both dictionaries;
- associates for each translation a probability that is the minimum in the two dictionaries.

This process:

- makes a stronger dictionary;
- reduces probability values;
- resulting values lose significance as a probability;

PTD Intersection

Intersection can be useful for:

- the intersection of dictionaries obtained from different domain corpora, to compute the shared or base lexicon;
- the intersection with a specific small hand-controlled dictionary can be used to tune PTD extraction algorithms;

$$\underbrace{\underbrace{\text{doente}}_{(10964)} \begin{cases} \text{patient} & 80.4\ \% \\ \text{patients} & 4.5\ \% \\ \text{card} & 0.6\ \% \\ \text{ill} & 0.6\ \% \\ \text{you} & 0.4\ \% \\ \text{she} & 0.3\ \% \end{cases} \cap \underbrace{\text{doente}}_{(404)} \begin{cases} \text{patient} & 24.2\ \% \\ \text{ill} & 18.7\ \% \\ \text{sick} & 17.7\ \% \\ \text{patients} & 10.1\ \% \\ \text{illness} & 2.0\ \% \\ \text{well} & 1.7\ \% \end{cases}}_{=}$$

$$\underbrace{\text{doente}}_{(404)} \begin{cases} \text{patient} & 80.4\ \% \\ \text{patients} & 4.5\ \% \\ \text{ill} & 0.6\ \% \end{cases}$$

PTD Domain Restriction and Subtraction

Domain restriction,

- restricts the PTD domain to a specific set of words;
- can be used to restrict a dictionary to some specific words;

Domain subtraction,

- removes from the PTD domain a specific set of words;
- can be used to remove from a dictionary some common words;
- . . . and that can be used to detect terminology;

Recalculate probabilities:

- Most of the operations described lose translations, or change translation probabilities;
- For example, when removing some words from the list of possible translations, the sum of translation probabilities is no longer 100 %;
- This operation recalculate all probabilities in a way they sum up 100 % again.

$$\text{totalize}\left(\underbrace{\text{doente}}_{(404)}\begin{cases}\text{patient} & 80.4\ \% \\ \text{patients} & 4.5\ \% \\ \text{ill} & 0.6\ \%\end{cases}\right) = \underbrace{\text{doente}}_{(404)}\begin{cases}\text{patient} & 94.0\ \% \\ \text{patients} & 5.3\ \% \\ \text{ill} & 0.7\ \%\end{cases}$$

PTD Composition

PTD composition is the most interesting operation:

- take two dictionaries, one $A \rightarrow B$ and another $B \rightarrow C$;
- compose the two dictionaries, obtaining a $A \rightarrow C$ dictionary;
- can be useful to calculate translation dictionaries for one pair of languages that does not have a parallel corpora available (**?**);

It can be used in other situations, like:

- take one dictionary, $A \rightarrow B$;
- compute $A \rightarrow A$ (compose with inverse dictionary);
- can be useful to calculate related-word sets.

PTD Composition: Triangulation Example

afluencia	influx	18.6 %	afflusso	48.9 %
			flusso	12.7 %
			flussi	4.7 %
	flow	12.9 %	flusso	46.9 %
			flussi	9.9 %
			gravi	1.7 %
	inflow	6.1 %	sfogo	24.2 %
			afflusso	16.8 %
			ascritto	14.7 %
	flood	5.9 %	inondazioni	5.6 %
			flusso	4.4 %
			alluvione	2.8 %
	flows	4.7 %	flussi	72.3 %
			flusso	1.6 %
			ondate	1.5 %

Results from the Triangulation Result

afluencia	afflusso	10.08 %
	flusso	8.73 %
	flussi	5.51 %
	sfogo	1.46 %
	ascritto	0.89 %
	inondazioni	0.33 %
	gravi	0.22 %
	alluvione	0.16 %
	ondate	0.07 %

Composing with the Inverse Dictionary

casa	house	48 %	casa	76 %	=	**36.5 %**
			house	5 %	=	**2.4 %**
			casiña	2 %	=	**1.0 %**
			vivenda	2 %	=	**1.0 %**
	home	35 %	casa	52 %	=	**18.2 %**
			fogar	11 %	=	**3.9 %**
			país	2 %	=	**0.7 %**
			domicilio	2 %	=	**0.7 %**
			terra	1 %	=	**0.4 %**
	cottage	2 %	casa	79 %	=	**1.60 %**
			cabana	9 %	=	**0.20 %**
	houses	1 %	casa	80 %	=	**0.80 %**
			casopa	5 %	=	**0.05 %**
			moradas	2 %	=	**0.02 %**
	homes	1 %	casa	19 %	=	**0.20 %**
			asilos	5 %	=	**0.05 %**
			domicilios	3 %	=	**0.03 %**

Result of inverse composition

casa	casa	57.3 %
	fogar	3.9 %
	house	2.4 %
	casiña	1.0 %
	vivenda	1.0 %
	país	0.7 %
	domicilio	0.7 %
	terra	0.4 %
	cabana	0.2 %
	casopa	0.05 %
	asilos	0.05 %
	domicilios	0.03 %
	moradas	0.02 %

PTD Filtering

- big corpora PTDs tend to be big;
- PTDs tend to be used for word translations:
  - so we can remove non-words, like numbers and punctuation;
- PTDs have a probability value associated with translations:
  - so we can use it to reduce the PTD size, removing non probable translations;
- PTDs have a number of occurrences:
  - so we can use it to remove infrequent words...

$$\text{filter}\left(\underbrace{\text{doente}}_{(10964)}\left\{\begin{array}{ll}\text{patient} & 80.4\ \% \\ \text{patients} & 4.5\ \% \\ \text{card} & 0.6\ \% \\ \text{ill} & 0.6\ \% \\ \text{you} & 0.4\ \% \\ \text{she} & 0.3\ \%\end{array}\right.\right) = \underbrace{\text{doente}}_{(10964)}\left\{\begin{array}{ll}\text{patient} & 80.4\ \% \\ \text{patients} & 4.5\ \%\end{array}\right.$$

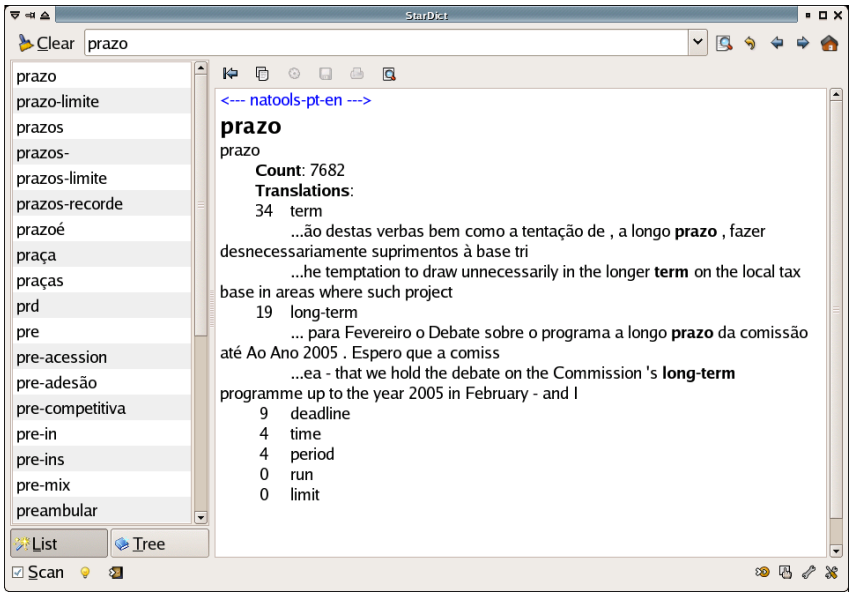
- PTDs can be used for terminology extraction (**?**; **?**);
- The process can be quite good if used together with a morphological analyzer;
- It uses translation patterns:

$$\mathcal{T}(A \cdot \text{“de”} \cdot B) = \mathcal{T}(B) \cdot \mathcal{T}(A)$$

	Human	Rights
Direitos		X
do		
Homem	X	

PTD can also be used directly by end-users.

- We've shown how to query them using Per-Fide corpora interface;
- We can also use a command-line tool: nat-ptd query;
- We can create dictionaries to be queried offline:



Bibliography