Combinatory Examples Extraction

Using Probabilistic Translation Dictionaries to Extract Examples

Alberto Simões and José João Almeida

1. **Extract Probabilistic Translation Dictionaries** from sentence-aligned parallel corpora — this results in translation probabilities between words (Hiemstra 98).
2. **Create Alignment Matrices** for each corpora translation unit, and try to find the best relationship between words, detecting blocks of translations.
3. **Extract and Consolidate Examples** from the translation blocks. Count their occurrence number and assign a quality stamp accordingly.

1. Probabilistic Translation Dictionaries

$$w_\alpha \rightharpoonup \left(occur \times w_\beta \rightharpoonup P\left(\mathcal{T}\left(w_\alpha\right) = w_\beta\right)\right)$$

The following example is from EuroParl: more than a million translation units, and 30 million words in each language. The resulting PTD include about 100 000 entries, each with 1 to 8 possible translations.

```
** Word: europe                    ** Word: stupid
** OccurrenceCount: 42853          ** OccurrenceCount: 180

      europa: 94.71 %                   estúpido:  17.55 %
    europeus:  3.39 %                   estúpida:  10.99 %
     europeu:  0.81 %                 estúpidos:   7.41 %
    europeia:  0.11 %                   avisada:   5.65 %
                                        direita:   5.58 %
                                        impasse:   4.48 %
```

2. Diagonal Finder

	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

Create a matrix and **fill** it with word translation probabilities in each cell: the average of $\mathcal{P}(\mathcal{T}(w_{TL}) = w_{SL})$ and $\mathcal{P}(\mathcal{T}(w_{SL}) = w_{TL})$

Smooth the matrix, enhancing values near the main diagonal.

	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**

Find **anchor points** on the matrix: values with high probability relatively to other values in the same row and column. These points are considered good translations, and as such, the number of anchor points in the final diagonal should be maximized.

For each language pair we define a **set of rules** to detect specific **patterns** where the main diagonal fails. These rules are defined using a **Domain Specific Language**.

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

```
[ABCCBA] A B C = C B A
```

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

	european	economic	area
espaço			X
económico		X	
europeu	X		

The language is powerful enough not just to define patterns but to **restrict** the cases when they can be applied, and to **infer** properties about words:

```
[ABBA] A[CAT -> ADJ] B[CAT <- N] = B[CAT <- N] A[CAT -> ADJ];
[POV]  N (P "of" V)[CAT -> N] = (P "de" V)[CAT -> N] N;
```

On the ABBA rule, we are defining that the word that matches the placeholder B need to be a noun (this is validated with a morphological analyzer). Also, we are inferring that words matching A are adjectives. This is added in a dictionary generated during this process. In the POV rule the same approach is used: the string that matches P "of" V will be considered a Noun.

	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**

The blocks defined by the DSL are constructed and matched against each possible cell set on the matrix. **Matching patters** will be considered as single cells and as single anchor points.

	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	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**

The **diagonal finding** algorithm will start in the up-left corner of the matrix and will try to reach the lower-right corner passing by the more number of anchor points possible, without getting too far from the main diagonal. When no adjacent cell is found, a multi-cell block is constructed

3. Examples Extraction

For each block on the matrix extract its relationship. We extract the first level example but also join them two or more times. This results in more and bigger examples. This process is important as it can raise the quality of other extracted examples.

Level 1

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

Level 2

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

Level 3

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

Examples Consolidation

```
parlamento europeu                              (2937)
       2855 european parliament
          1 european parliament and the
          1 european parliament considered the
          1 european parliament during the
          1 european parliament has the
          [...]

carta dos direitos fundamentais                  (385)
        333 charter of fundamental rights
         14 charter of fundamental
          9 charter on fundamental rights
          5 charter of fundamental human rights
          [...]
```