Capítulo 2

Extracção de recursos de tradução

Alberto Simões

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 2, p. 7–11, 2008

Este documento resume a dissertação na extracção de recursos de tradução (Simões, 2008) e a sua integração nos objectivos da Linguateca. A dissertação teve como principal objectivo o estudo de métodos para a extracção de recursos de tradução para a língua portuguesa, uma vez que a principal investigação na tradução automática não tem dado a atenção merecida a esta língua.

A tradução automática tem vindo a dar cada vez mais atenção aos métodos de tradução baseados em dados. Estes métodos reaproveitam as traduções que já foram realizadas (no mesmo ou noutros contextos) para realizar as novas traduções. O principal problema desta abordagem é conseguir emparelhar as traduções já realizadas com a frase a traduzir. Por exemplo, nos sistemas de tradução assistida por computador (em inglês, CAT – *Computer Assisted Translation*) é habitual que só sejam reaproveitadas frases muito semelhantes às já traduzidas. Para os sistemas de tradução automática pretende-se aumentar a aplicabilidade das frases já traduzidas, aplicando algoritmos que dividam as traduções já realizadas em segmentos mais pequenos (sintagmas ou simples segmentos de palavras paralelos) com maior reutilização (e que são chamados de *exemplos de tradução*).

Em trabalho anterior (Simões e Almeida, 2003; Simões, 2004) tinham sido estudados métodos para a extracção de dicionários probabilísticos de tradução. Estes dicionários são associações entre palavras na língua de origem com um conjunto de possíveis traduções na língua de destino juntamente com a respectiva probabilidade de tradução (ver figura 2.1).

$$\mathcal{T}\,(\text{codificada}) = \begin{cases} \text{codified} & 62.83\% \\ \text{uncoded} & 13.16\% \\ \text{coded} & 6.47\% \\ \cdots & \end{cases}$$

Figura 2.1: Extracto de um dicionário probabilístico de tradução para a palavra "codificada".

Embora extraídos automaticamente e sem garantias de grande qualidade, mostraram-se extremamente úteis para a extracção de novos recursos de tradução. Além das várias avaliações reportadas na dissertação de doutoramento, foram feitas algumas comparações (Santos e Simões, 2008) de resultados destes dicionários com dicionários de tradução de cores obtidos manualmente a partir do COMPARA (Frankenberg-Garcia e Santos, 2003).

Para a extracção dos vários recursos foram usados vários corpora. Para além do COMPARA foram utilizados o EuroParl v2 (Koehn, 2005), o JRC-Acquis (Steinberger et al., 2006), El Monde Diplomatique (Correia, 2006) e o EurLex, um corpus construido pelo Projecto Natura, com mais de um milhão de unidades de tradução. Algumas versões alinhadas destes corpora, bem como os respectivos dicionários de tradução, estão acessíveis para consulta interactiva em `linguateca.di.uminho.pt/nat/`.

Todo os recursos extraídos durante a dissertação usaram como base os dicionários pro-

babilísticos de tradução para estabelecer pontes entre palavras de duas línguas, e foram aplicaradas diferentes metodologias para a extracção de exemplos de tradução:

- o uso da hipótese das palavras-marca (*Marker Hypothesis*) como mecanismo de segmentação dos corpora paralelos, e o uso das probabilidades de tradução constantes nos dicionários probabilísticos de tradução para o alinhamento destes segmentos (Simões, 2007b).

  Este método baseia-se num conjunto de palavras (pronomes, artigos, alguns advérbios, etc) que, de acordo com Green (1979), podem ser usados como um método eficaz de segmentação:

  O João passou toda a tarde a brincar com os colegas.
  ⇓
  O João passou toda a tarde a brincar com os colegas.
  ⇓
  (*O João passou*) (*toda a tarde*) (*a brincar*) (*com os colegas.*)

  Esta abordagem já tinha sido usada para a segmentação para tradução automática (Armstrong et al., 2006) mas sem terem sido realizadas experiências com a língua portuguesa, nem usando dicionários probabilísticos de tradução para o alinhamento dos segmentos extraídos. A tabela 2.1 apresenta os exemplos (1:1) mais ocorrentes extraídos do EuroParl PT:EN com base na hipótese das palavras-marca.

Ocorrências	Português	Inglês
36886	senhor presidente	mr president
8633	senhora presidente	madam president
3152	espero	I hope
2930	gostaria	I would like
2572	o debate	the debate
2511	penso	I think
2356	está encerrado	is closed
1939	penso	I believe
1932	muito obrigado	thank
1854	em segundo lugar	secondly
$\bar{x} = 1.6654$	Total de 1 507 225 exemplos 1:1	

Tabela 2.1: Exemplos mais ocorrentes extraídos com base na hipótese das palavras marca.

- a construção de uma matriz de alinhamento para cada unidade de tradução, onde cada célula da matriz é preenchida com a probabilidade mútua de tradução entre palavras (ver figura 2.2). Nesta matriz são procuradas as células com probabilidades

mais elevadas, e que correspondem às traduções provavelmente correctas. Estas traduções são extraídas e são criados exemplos de tradução (Simões e Almeida, 2006a). Esta abordagem não é totalmente nova (Melamed, 2001), mas foi introduzido o uso de dicionários probabilísticos de tradução e o uso de padrões de alinhamento.

	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**

Figura 2.2: Matriz de alinhamento.

- a extracção de exemplos, que correspondem a segmentos nominais (próximos de sintagmas nominais, e candidatos a terminologia), com base em padrões de alinhamento, que especificam as trocas de ordem de palavras que ocorrem durante a tradução (Simões e Almeida, 2008).

  Foi desenvolvida uma nova linguagem de domínio específico para a especificação de padrões com objectivos distintos das linguagens de padrões actualmente a serem usadas na área da tradução automática (Och e Ney, 2004; Sánchez-Martínez e Forcada, 2007).

  Seguem-se alguns exemplos de padrões, bem como a respectiva ilustração/interpretação na figura 2.3. Os segmentos nominais extraídos são contados. O número de ocorrências de cada par permite associar-lhe uma noção de qualidade, de acordo com a tabela 2.2. A tabela 2.3 contém algumas medidas de avaliação destes recursos. Consultar Simões (2008) para detalhes sobre a forma como a avaliação foi realizada.

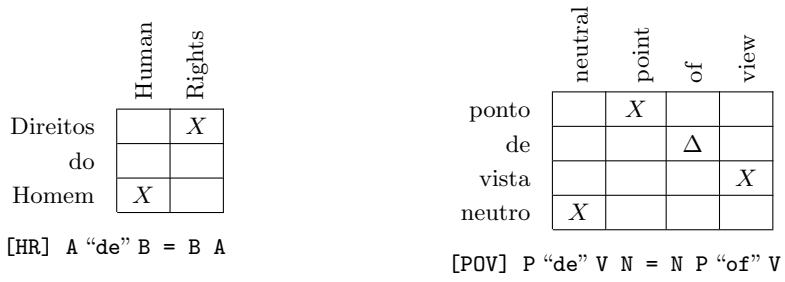

Figura 2.3: Padrões de alinhamento HR e POV.

39214	comunidades europeias	european communities
32850	jornal oficial	official journal
32832	parlamento europeu	european parliament
32730	união europeia	european union
15602	países terceiros	third countries
[...]	[...]	[...]
1	órgãos orçamentais	budgetary organs
1	órgãos relevantes	relevant bodies
1	óvulos de equino	equine ova
1	óxido de cádmio	cadmium oxide
1	óxido de estireno	styrene oxide

Tabela 2.2: Extracto das contagens de unidades nominais.

Padrão	Total	Máx.	Mediana	Min.	Precisão
A B = B A	77 497	938	2	1	86 %
A "de" B = B A	12 694	204	2	1	95 %
A B C = C B A	7 700	40	1	1	93 %
I "de" D H = H D I	3 336	21	1	1	100 %
A B C = C A B	1 466	4	1	1	40 %
P "de" V N = N P "of" V	564	6	1	1	98 %
P "de" T "de" F = F T P	360	3	1	1	96 %

Tabela 2.3: Avaliação de unidades nominais extraídas.

Para além da experimentação dos métodos, estes foram disponibilizados num pacote de ferramentas de código aberto, denominado NATools (Simões e Almeida, 2007). As ferramentas constantes neste pacote foram adaptadas para funcionarem de forma distribuída cliente/servidor (Simões e Almeida, 2006b) e de forma paralela num *cluster* computacional (Simões, 2007a). Além disso, parte destes recursos foram usados no CLEF de 2005 (Cardoso et al., 2006a).