Extracção de Recursos de Tradução

Alberto Simões
ambs@di.uminho.pt

http://linguateca.di.uminho.pt/natools/

Este documento resume a dissertação na extracção de recursos de tradução [11] 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 (CAT) 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 [17, 10] 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:

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

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 [9] de resultados destes dicionários com dicionários de tradução de cores obtidos manualmente a partir do COMPARA [4].

Para a extracção dos vários recursos foram usados vários corpora. Para além do COMPARA [4] foram utilizados o EuroParl v2 [6], o JRC-Acquis [19], El Monde Diplomatique [3] e o EurLex, um corpus construido no 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 http://linguateca.di.uminho.pt/nat/.

Todo os recursos extraídos durante a dissertação usaram como base os dicionários probabilísticos de tradução para estabelecer pontes entre palavras de duas línguas, e aplicaram-se 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 [12].

  Este método baseia-se num conjunto de palavras (pronomes, artigos, alguns advérbios, etc) que, de acordo com [5], 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 vindo a ser usada para a segmentação para tradução automática [1] mas sem terem sido realizadas experiências com a língua portuguesa, nem usando dicionários probabilísticos de tradução para o alinhamento dos segmentos extraídos. Seguem-se 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	

- 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. 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 [13].

	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**

Esta abordagem não é totalmente nova [7], mas foi introduzido o uso de dicionários probabilísticos de tradução e o uso de padrões 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 [16].

  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 [8, 20].

  Seguem-se alguns exemplos de padrões, bem como a respectiva ilustração/interpretação:

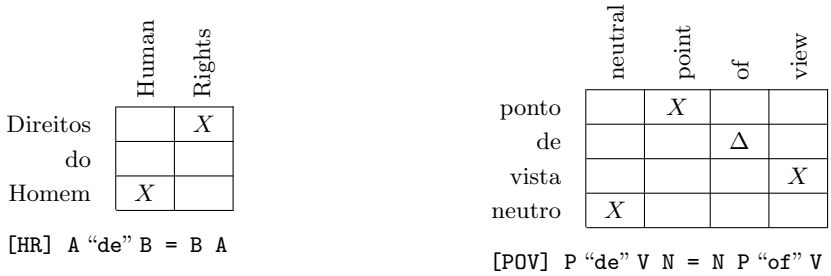

Figura 1: Padrões de alinhamento HR e POV.

As unidades nominais extraidas são contadas. O número de ocorrências de cada par permite associar-lhe uma noção de qualidade, de acordo com a tabela 1.

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 1: Extracto das contagens de unidades nominais.

A tabela 2 contém algumas medidas de avaliação destes recursos. Consultar [11] para detalhes sobre a forma como a avaliação foi realizada.

Para além da experimentação dos métodos, estes foram disponibilizados num pacote de ferramentas de código aberto, denominado NATools [15]. As ferramentas constantes neste pacote foram adaptadas para funcionarem de forma distribuída Cliente/Servidor [14], e de forma paralela

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: Avaliação de unidades nominais extraídas.

num *cluster* computacional [18]. Além disso, parte destes recursos foram usados no CLEF de 2005 [2].

Referências

[1] Stephen Armstrong, Marian Flanagan, Yvette Graham, Declan Groves, Bart Mellebeek, Sara Morrissey, Nicolas Stroppa e Andy Way. MaTrEx: machine translation using examples. Em *TC-STAR OpenLab Workshop on Speech Translation*. 2006.

[2] Nuno Cardoso, Leonardo Andrade, Alberto Simões e Mário J. Silva. The XLDB Group at the CLEF 2005 Ad-Hoc Task. Em C. Peters, F. Gey, J. Gonzalo, H. Mueller, G. Jones, M. Kluck, B. Magnini e M. Rijke, editores, *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*. Setembro de 2005. p. 54–60. `http://alfarrabio.di.uminho.pt/~albie/publications/clef05.pdf`.

[3] Ana Teresa Varajão Moutinho Pereira Correia. Colaboração na constituição do corpus paralelo Le Monde Diplomatique (FR-PT). Relatório de estágio. Universidade do Minho. Dezembro de 2006.

[4] Ana Frankenberg-Garcia e Diana Santos. Introducing COMPARA, the portuguese-english parallel translation corpus. Em Silvia Bernardini Federico Zanettin e Dominic Stewart, editores, *Corpora in Translation Education*. Manchester: St. Jerome Publishing. 2003. p. 71–87. `http://www.linguateca.pt/Diana/download/Frankenberg-GarciaSantos2000.rtf`.

[5] Thomas R. G. Green. The necessity of syntax markers. two experiments with artificial languages. *Journal of Verbal Learning and Behaviour*. 18:481–496. 1979.

[6] P Koehn. Europarl: A parallel corpus for statistical machine translation. Em *In Proceedings of MT-Summit*. 2005. p. 79–86.

[7] I. Dan Melamed. *Empirical Methods for Exploiting Parallel Texts*. MIT Press. 2001.

[8] Franz Josef Och e Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*. 30:417–449. 2004.

[9] Diana Santos e Alberto Simões. Portuguese-English word alignment: some experiments. Em *LREC 2008 — The 6th edition of the Language Resources and Evaluation Conference*. 28–30 de Maio de 2008.

[10] Alberto Simões. Parallel corpora word alignment and applications. Tese de mestrado. Escola de Engenharia - Universidade do Minho. 2004. `http://alfarrabio.di.uminho.pt/~albie/publications/msc.pdf`.

[11] Alberto Simões. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Tese de doutoramento. Escola de Engenharia, Universidade do Minho. Maio de 2008.

[12] Alberto Simões. Segmentação bilingue com base na marker hypothesis. Em César Analide, Paulo Novais e Pedro Henriques, editores, *Simpósio Doutoral em Inteligência Artificial 2007*. Dezembro de 2007. p. 135–144.

[13] Alberto Simões e José João Almeida. Combinatory examples extraction for machine translation. Em Jan Tore Lønning e Stephan Oepen, editores, *11th Annual Conference of the European Association for Machine Translation*. 19–20 de Junho de 2006. p. 27–32. ISBN 82-7368-294-3. `http://alfarrabio.di.uminho.pt/~albie/publications/eamt06.pdf`.

[14] Alberto Simões e José João Almeida. NatServer: a client-server architecture for building parallel corpora applications. *Procesamiento del Lenguaje Natural*. 37:91–97. Setembro de 2006. `http://alfarrabio.di.uminho.pt/~albie/publications/sepln06.pdf`.

[15] Alberto Simões e José João Almeida. Parallel corpora based translation resources extraction. *Procesamiento del Lenguaje Natural*. (39):265–272. Setembro de 2007.

[16] Alberto Simões e José João Almeida. Biingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*. Setembro de 2007.

[17] Alberto Simões e José João Almeida. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*. 31:217–224. September de 2003. `http://alfarrabio.di.uminho.pt/~albie/publications/sepln2003.pdf`.

[18] Alberto Simões, Rúben Fonseca e José João Almeida. Makefile::Parallel dependency specification language. Em Anne-Marie Kermarrec, Luc Bougé e Thierry Priol, editores, *Euro-Par 2007*. Agosto de 2007. p. 33–41.

[19] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş e Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Em *5th International Conference on Language Resources and Evaluation (LREC'2006)*. 24–26 de Maio de 2006.

[20] Felipe Sánchez-Martínez e Mikel L. Forcada. Automatic induction of shallow-transfer rules for open-source machine translation. Em *TMI, The Eleventh Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*. 2007. p. 181–190.