Segmentação Bilingue com base na *Marker Hypothesis*⋆

Alberto Simões

Departamento de Informática, Universidade do Minho
Campus de Gualtar, 4710-057 Braga, PORTUGAL
ambs@di.uminho.pt

Resumo A existência de **exemplos de tradução** é imprescindível para tradução assistida por computador bem como para tradução automática baseada em dados (EBMT e SMT). No entanto, o uso de unidades de tradução de corpora paralelos directamente na tradução não é eficaz já que estas unidades são demasiado grandes, e portanto, torna-se pouco provável que uma mesma unidade de tradução tenha de ser traduzida mais do que uma vez.
Para colmatar este problema há necessidade de explorar outras metodologias para a divisão de unidades de tradução em segmentos paralelos mais pequenos. Uma das abordagens que tem vindo a ser utilizada é a segmentação baseada em marcadores (***Marker Hypothesis***). Este documento pretende documentar as experiências realizadas na utilização deste método para a segmentação de texto português (paralelo com o inglês).

Orientador: José João Almeida (jj@di.uminho.pt)
Palavras-chave: extracção de exemplos de tradução, tradução automática, corpora paralelos, processamento de linguagem natural
Grau a que se candidata: Doutor
Data de incício: Setembro de 2004
Data prevista para conclusão: Dezembro de 2007

1 Introdução

A tradução assistida por computador e a chamada tradução automática baseada em dados (onde se incluem a tradução automática baseada em exemplos (EBMT) e a tradução automática estatística (SBMT)) tiram partido de corpora paralelos no sentido de reutilizar traduções realizadas previamente.

Um tradutor, enquanto utilizador de uma ferramenta de tradução assistida, consegue gerir de forma mais ou menos controlada o tamanho das unidades das suas memórias de tradução, quando se extrai unidades de tradução de forma

⋆ Este trabalho é parte integrante do doutoramento que visa a extracção de recursos bilingues (dicionários de tradução, terminologia bilingue, exemplos de tradução) para a tradução assistida por computador. Optou-se por não apresentar todo o trabalho de doutoramento já que o mesmo se encontra para publicação [7].

automática isto não é possível. Basta analisar alguns dos corpora paralelos disponíveis, como sejam o EuroParl [5] ou o JRC-Acquis [10], para se verificar que as unidades de tradução são muito grandes (acima das 20 palavras).

Dado o tamanho destas unidades de tradução, a sua reutilização é difícil (pouco provável que um sistema precise de traduzir uma frase similar a uma outra de 25 palavras que está num dos corpora base). Para resolver este problema são habitualmente empregues técnicas de *chunking* para a divisão destas unidades de tradução em segmentos mais pequenos com uma maior taxa de reutilização.

No âmbito deste doutoramento te-se vindo a fazer algumas propostas para segmentação de texto, geração combinatória de exemplos [6] e extracção de terminologia bilingue. No entanto, não extrai segmentos a que se possam chamar "linguísticos" [9], pelo que se decidiu investir no estudo de um outro método. Esta abordagem dupla não se prendeu com a falta de qualidade do método original, mas pelo interesse dos dois tipos de resultados, já que é defendido [9] que os exemplos não "linguísticos" são mais ricos para a tradução automática, nomeadamente, tradução automática baseada em exemplos.

Assim, e com base em trabalho já realizado por outros grupos de investigação [11] decidiu-se analisar a *Marker Hypothesis*[1] [4] para a segmentação de texto bilingue em que uma das línguas é o português, e extracção de segmentos paralelos de tamanhos reduzidos para uso na tradução automática. Embora este método tenha vindo a ser utilizado durante alguns anos, desconhecemos alguma análise da sua aplicabilidade à língua portuguesa.

2 A *Marker Hypothesis*

A *Marker Hypothesis* defende que existe um conjunto fechado de palavras que são usadas na construção de frases e que servem como delimitadores de segmentos/sintagmas. Em [4] é também defendida uma visão psico-linguística destas marcas, em que se apresentam estudos sobre a facilidade de aprendizagem de uma língua de acordo com a existência ou não de marcas nessa linguagem.

As marcas de determinada língua incluem habitualmente as artigos, conjunções, preposições, pronomes, locuções, numerais e alguns advérbios. Por exemplo, na seguinte frase,

O João passou toda a tarde a brincar com os colegas.

as marcas são os artigos "*o*", "*a*" e "*os*", a preposição "*com*" e o pronome "*toda*":

O João passou toda a tarde a brincar com os colegas.

Se considerarmos todos os segmentos que começam por uma ou mais marcas, e terminam antes do próximo conjunto de marcas, obtemos os seguintes segmentos:

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

[1] Dada a falta de uma tradução "*oficial*" para a língua portuguesa optou-se pelo uso da sua forma original.

que embora não correspondam aos verdadeiros sintagmas da frase, correspondem a uma boa aproximação (um bom ponto de partida).

A lista de marcadores para a língua portuguesa foi construída com base na lista inglesa usada no projecto MaTrEx[1], à qual se juntaram outras marcas que foram vindo a ser reconhecidas em corpora. A tabela 1 mostra um excerto desta lista. É interessante reparar que a lista portuguesa é razoavelmente maior devido à flexão de género e número que não se verifica em inglês (um total de 398 marcas para a língua inglesa e de 596 marcas para a língua portuguesa).

Dada a lista de palavras, a segmentação em qualquer uma das línguas é realizada do seguinte modo:

- anotar todas as marcas existentes na frase;
- cada segmento começa com uma marca e aglutina todas as marcas adjacentes;
- cada segmento termina numa palavra que não é marca e que está imediatamente antes de uma marca (transição *não marca – marca*) ou no final da frase;
- em cada segmento é anotada a parte correspondente às marcas para que se possam agrupar por tipo de marca;

A tabela 2 mostra alguns segmentos (entre os mais comuns) extraídos a partir do corpus EuroParl PT:EN. A lista de segmentos diferentes extraídos do lado português tem 3 070 398 segmentos, enquanto que a lista para o lado inglês tem 3 103 797 segmentos.

3 Alinhamento de Segmentos

Enquanto que a segmentação monolingue usando a *Marker Hypothesis* tem um algoritmo bem definido, o mesmo não acontece para a extracção de segmentos bilingues alinhados. Embora se pudesse esperar que tendo uma frase e a sua tradução, o número de segmentos fosse o mesmo e a ordem dos mesmos fosse muito próxima, a verdade é que tal não acontece: não só as traduções são demasiado livres, como a diferença entre línguas leva a que a quantidade de segmentos em cada língua seja diferente. Um exemplo típico entre a língua portuguesa e inglesa é o uso de adjectivo:

(*o* casaco) (*de* peles) // (*the* fur coat)

Para ajudar no alinhamento de segmentos entre línguas foram usados dicionários probabilísticos de tradução (ver secção 3.1). Estes dicionários permitem-nos obter medidas de "*tradutibilidade*" entre dois segmentos, já que incluem probabilidades de tradução entre palavras de diferentes línguas.

O algoritmo compreende os seguinte passos:

- segmentação da unidade de tradução em cada uma das línguas usando o algoritmo da *Marker Hypothesis*;

Inglês	***Português***
most	maior; maioria
much	muito
my	meu; minha; meus; minhas
near	perto; próximo; quase
nearby	perto; próximo
neither	tão-pouco; também não
next	seguinte; próximo; próxima
nigh	próximo
no	não
nor	nem
now	agora; uma vez que; considerando que
of	de; por; em
off	de; fora
on	em; sobre; em cima de; de; relativa
once	desde que; uma vez que; se
one	um; uma
only	todavia; mas; contudo
onto	para; para cima de; em direcção a
or	ou; se não
other	outro; outra; outras; outros
our	nosso; nossa; nossos; nossas
ours	o nosso; a nossa; os nossos; as nossas
out	fora
over	sobre; em cima de; por cima de
owing to	devido a: por consequência de; por causa de
own	próprio; ser proprietário
past	por; para além disso; fora de
pending	durante; até
per	por; através de; por meio de; devido a acção de
plus	mais; a acrescentar a; a adicionar a
round	em torno de; à volta de
same	mesmo; mesma
several	vários
sort of	espécie de; género de; tipo de; de certo modo
since	desde; desde que; depois que
so	portanto; por isso
some	algum; alguns; alguma; algumas
subject to	sujeito a
such	este; esse; aquele; isto; aquilo
supposing	supondo; se; no caso de; dada a hipótese de
than	de; que; do que; que não
that	aquele; aquela; aquilo; esse; essa; isso; ...
the	o; a; os; as

Tabela 1. Excerto de marcadores EN:PT.

34 137	**da** comissão	13 566	**and** gentlemen
17 277	**do** conselho	11 466	**the** commission
16 891	**da** união europeia	9 182	**to** make
11 379	**em** matéria	8 712	**to** be
9 880	**de** trabalho	8 356	**to** do
9 850	**da** união	7 992	**of the** european union
9 479	**no** sentido	7 941	**of the** committee
5 332	**em primeiro** lugar	7 814	**to** say
3 245	**no que** diz respeito	7 574	**with** regard
2 214	**para o** desenvolvimento	3 749	**in the** european union

Tabela 2. Segmentos que mais ocorrem no EuroParl (PT e EN).

- cálculo de uma matriz de correspondências entre segmentos utilizando o dicionário probabilístico de tradução;
- selecção das células relevantes para o alinhamento dos segmentos;
- extracção das correspondências bilingues entre segmentos;
- "generalização" e acumulação dos resultados obtidos;

Este algoritmo é detalhado na secção secção 3.2 com um exemplo. Na secção 3.3 são discutidos os resultados obtidos.

3.1 Dicionários Probabilísticos de Tradução

Os dicionários probabilísticos de tradução (PTD^2) [8] são dicionários que associam a palavras de determinada língua $\mathcal{L}_\alpha$ um conjunto de palavras numa outra língua $\mathcal{L}_\beta$, juntamente com uma probabilidade de tradução:

$$PTD(\mathcal{L}_\alpha, \mathcal{L}_\beta) = w_{\mathcal{L}_\alpha} \rightharpoonup \left(occ(w_{\mathcal{L}_\alpha}) \times \left(w_{\mathcal{L}_\beta} \rightharpoonup \mathcal{P}\left(\mathcal{T}\left(w_{\mathcal{L}_\alpha}\right) = w_{\mathcal{Y}_\beta}\right)\right)\right)$$

A figura 1 mostra duas entradas de um PTD. É importante realçar que estes dicionários não podem ser vistos como dicionários de tradução típicos já que como são extraídos automaticamente, alguns dos relacionamentos não são verdadeiras traduções. De notar que estes dicionários têm o tamanho correspondente ao número de palavras existente no corpus, pelo que quanto maior for a diversidade léxica de um corpus, maior será também a diversidade do dicionário extraído.

3.2 Algoritmo de Alinhamento

A segmentação de uma unidade de tradução resulta num conjunto de segmentos para cada uma das línguas. O alinhamento destes segmentos pode ser visto como um caso particular do alinhamento de frases [3], e portanto com uma solução conhecida baseada em programação dinâmica. A abordagem proposta baseia-se no uso de dicionários probabilísticos de tradução para o alinhamento de

```
europe => {    count => 42853,
               trans => {      europa    => 0.9471,
                               europeus  => 0.0339,
                               europeu   => 0.0081,
                               europeia  => 0.0011,
                         },
          },
stupid => {    count => 180,
               trans => {      estúpido  => 0.1755,
                               estúpida  => 0.1099,
                               estúpidos => 0.0741,
                               avisada   => 0.0565,
                               direita   => 0.0558,
                               impasse   => 0.0448,
                          },
           },
```

Figura 1. Extracto de um Dicionário Probabilístico de Tradução.

segmentos, associando assim ao cada par de segmentos um valor probabilístico correspondente a uma valoração qualitativa do alinhamento.

O processo baseia-se na construção de uma matriz de alinhamento em que cada segmento da língua $\mathcal{L}_\alpha$ corresponde a uma linha da matriz, e cada segmento da língua $\mathcal{L}_\beta$ corresponde a uma coluna da matriz. Cada célula (i, j) é preenchida com uma medida da probabilidade do segmento $s_{\beta,j}$ ser tradução do segmento $s_{\alpha,i}$. Segue-se a selecção das células com maior probabilidade e que correspondem ao alinhamento pretendido. Embora a abordagem seja extremamente simples, não é igualmente simples a definição de uma fórmula para o cálculo da probabilidade de tradução entre dois segmentos.

Um dos problemas na construção desta fórmula é a baixa probabilidade de tradução que existe habitualmente entre marcas: dada a profusa flexão da língua portuguesa, as probabilidades associadas à tradução de um artigo da língua inglesa para a língua portuguesa são bastante baixas (considerando um caso óptimo de "*the*" traduzido por "*a*", "*o*", "*as*" e "*os*" teríamos 25% de probabilidade para cada uma destas traduções). Embora o mesmo vá acontecendo com o resto do léxico, as probabilidades não são tão baixas como as dos marcadores, pelo que dar mais peso às palavras que não são marcas deverá ser suficiente.

Por outro lado, se um segmento em determinada língua corresponder a vários segmentos na outra língua, só algumas palavras do primeiro segmento vão ter uma tradução válida em cada um dos segmentos da outra língua. Assim, a probabilidade de tradução não deve ser vista como "*a probabilidade de* s_α *e* s_β *serem traduções mútuas*" mas antes, considerando que $size(s_\alpha) > size(s_\beta)$, como "*a probabilidade de a tradução de* s_β *estar contida em* s_α."

[2] Probabilistic Translation Dictionaries

Data: Sejam s_α e s_β dois segmentos, na língua $\mathcal{L}_\alpha$ e $\mathcal{L}_\beta$ respectivamente, tal que $s_\alpha < s_\beta$ e, $\mathcal{D}_{\alpha,\beta}$ e $\mathcal{D}_{\beta,\alpha}$ os dicionários probabilísticos de tradução entre essas línguas.

$SomaProbs \longleftarrow 0$
for $w_\alpha \in marcas\,(s_\alpha)$ **do**
  $Trads_{w_\alpha} \longleftarrow \mathcal{T}_{\mathcal{D}_{\alpha,\beta}}\,(w_\alpha)$
  for $w_\beta \in Trads_{w_\alpha}$ **do**
    if $w_\beta \in marcas\,(s_\beta)$ **then**
      $SomaProbs \longleftarrow SomaProbs + \mathcal{P}\,(w_\beta \in Trads_{w_\alpha})$

$ProbMarcas \longleftarrow \frac{SomaProbs}{\text{size(marcas}(w_\alpha))}$

$SomaProbs \longleftarrow 0$
for $w_\alpha \in texto\,(s_\alpha)$ **do**
  $Trads_{w_\alpha} \longleftarrow \mathcal{T}_{\mathcal{D}_{\alpha,\beta}}\,(w_\alpha)$
  for $w_\beta \in Trads_{w_\alpha}$ **do**
    if $w_\beta \in texto\,(s_\beta)$ **then**
      $SomaProbs \longleftarrow SomaProbs + \mathcal{P}\,(w_\beta \in Trads_{w_\alpha})$

$ProbTexto \longleftarrow \frac{SomaProbs}{\text{size(texto}(w_\alpha))}$

$Prob = 0.1 \times ProbMarcas + 0.9 \times ProbTexto$

Algoritmo 1: Cálculo (simplificado) de probabilidade de tradução entre dois segmentos.

O algoritmo 1 mostra de forma simplificada o processo de cálculo da probabilidade de tradução entre dois segmentos tendo em conta os detalhes discutidos. Utilizando esta fórmula em cada combinação de dois segmentos é preenchida uma matriz de alinhamento (ver exemplo na tabela 3).

O passo seguinte é encontrar as células com maiores probabilidades e fazer o alinhamento entre segmentos. Do exemplo apresentado seriam extraídos os seguintes exemplos de tradução:

- *a presente decisão produz efeitos / this decision shall take effect*
- *em 16 de setembro de 1999 / on 16 september 1999*

Depois de ordenados e de calcular as contagens de ocorrências por tradução, chegamos a exemplos como os que se apresentam na tabela 4.

3.3 Análise de Resultados

Uma análise aos resultados mostra que exemplos com mais de 50 ocorrências (cerca de 2000 exemplos para o EuroParl) são traduções fiáveis, havendo apenas necessidade de filtrar algumas traduções de números e de pontuações, que embora correctas são de pouca utilidade.

	this decision shall take effect	*on* 16 september 1999
a presente decisão produz efeitos	23.18%	5.86%
em 16	0.00%	76.41%
de setembro	0.00%	85.60%
de 1999	0.00%	84.10%

Tabela 3. Matriz de alinhamento.

Ocorrências	***Português***	***Inglês***
22 653	senhor presidente	mr president
4 380	senhora presidente	madam president
2 293	da união europeia	of the european union
2 274	, em nome	, on behalf
1 980	espero	i hope
1 932	da comissão	of the committee
1 855	gostaria	i would like
1 848	na europa	in europe
1 691	o debate	the debate
1 530	do conselho	of the council
1 501	penso	i think
1 460	da comissão	of the commission
1 390	está encerrado	is closed
1 367	penso	i believe
1 335	da europa	of europe
1 235	gostaria	i should like
1 178	em segundo lugar	secondly
305	dos direitos do homem	of human rights
242	dos direitos da mulher	on women 's rights
192	a proposta da comissão	the commission 's proposal

Tabela 4. Exemplos mais ocorrentes no EuroParl.

Embora estas traduções sejam correctas não podem ser vistas como única fonte para a tradução automática. O problema da junção de segmentos numa tradução que tiveram origem em frases diferentes (problema conhecido como *Boundary Friction* [2]) não é de todo resolvido. Inclusive, é defendido que segmentos linguísticos como os obtidos por este método não têm resultados bons na tradução baseada em exemplos. Neste sentido, a avaliação deste recurso deve ser feita não como um recurso isolado mas como parte integrante de um sistema de tradução automática.

4 Conclusões

A *Marker Hypothesis* tem resultados igualmente interessantes na língua inglesa e portuguesa. A verdade é que existindo um maior número de marcadores e de

uso bastante mais intensivo para a língua portuguesa, a quantidade de segmentos extraídos é maior do que a quantidade de segmentos extraídos da língua inglesa. Esta desproporção leva a que o alinhamento entre segmentos não seja trivial. O uso de dicionários probabilísticos de tradução mostrou-se imprescindível para o alinhamento eficaz destes segmentos.

No entanto, e em paralelismo com o que acontece no alinhamento de corpora ao nível da frase, o alinhamento obtido nunca corresponde a um alinhamento ao nível do elemento mais pequeno. Ou seja, nem sempre é possível alinhar apenas um segmento com um segmento, obtendo-se frequentemente relações de $n:1, 1:n$ e até mesmo $n:m$.

Embora nem todos os exemplos extraídos usando esta metodologia possam ser considerados correctos para uso na tradução, os que ocorrem um maior número de vezes são correctos. A acumulação de exemplos extraídos de diferentes corpora levará a que a quantidade de exemplos de qualidade seja maior.

Como trabalho futuro pretende-se comparar a qualidade dos resultados obtidos com este método de segmentação com o apresentado em [6]. Está também em estudo a possibilidade de se cruzarem os algoritmos.

Agradecimentos

Alberto Simões tem uma bolsa da Fundação para a Computação Científica Nacional e o trabalho aqui relatado é parcialmente suportado pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC/339/1.3/C/NAC, co-financiado pelo POSC.

Referências

1. Stephen Armstrong, Marian Flanagan, Yvette Graham, Declan Groves, Bart Mellebeek, Sara Morrissey, Nicolas Stroppa, and Andy Way. MaTrEx: machine translation using examples. In *TC-STAR OpenLab Workshop on Speech Translation*, Trento, Italy, 2006.
2. Ralf D. Brown, Rebecca Hutchinson, Paul N. Bennett, Jaime G. Carbonell, and Peter Jansen. Reducing boundary friction using translation-fragment overlap. In *MT Summit IX*, New Orleans, 2003.
3. William A. Gale and Kenneth Ward Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.
4. Thomas R. G. Green. The necessity of syntax markers. two experiments with artificial languages. *Journal of Verbal Learning and Behaviour*, 18:481–496, 1979.
5. Philipp Koehn. EuroParl: a multilingual corpus for evaluation of machine translation. Draft, Unpublished, 2002.
6. Alberto Simões and J. João Almeida. Combinatory examples extraction for machine translation. In Jan Tore Lønning and Stephan Oepen, editors, *11th Annual Conference of the European Association for Machine Translation*, pages 27–32, Oslo, Norway, 19–20, June 2006.
7. Alberto Simões and José João Almeida. Parallel corpora based translation resources extraction. *Procesamiento del Lenguaje Natural*, (39):265–272, September 2007.

8. Alberto Manuel Brandão Simões. Parallel corpora word alignment and applications. Master's thesis, Escola de Engenharia - Universidade do Minho, 2004.
9. Harold Somers. Review article: Example based machine translation. *Machine Translation*, 14(2):113–157, 1999.
10. Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy, 24–26 May 2006.
11. Tony Veale and Andy Way. Gaijin: A template driven bootstrapping approach to EBMT. In *NeMNLP'97*, Sofia, Bulgaria, 1997.