Alinhamento de corpora paralelos

Alberto Manuel Simões
albie@alfarrabio.di.uminho.pt

Resumo *Este documento apresenta um conjunto de ferramentas denominado NATools para o alinhamento de corpora paralelos. É apresentado o processo de alinhamento tendo em conta os vários níveis intervenientes, desde o convencional alinhamento à frase, até ao alinhamento à palavra, com a criação dos respectivos dicionários de tradução.*

São apresentadas medidas em relação ao tempo usado para o alinhamento, bem como resultados obtidos. São discutidas técnicas para a detecção de traduções de termos multi-palavra usando o algoritmo de alinhamento à palavra.

Os dicionários de tradução obtidos irão ser explicados e as suas aplicações exploradas: navegação e consulta web dos dicionários produzidos e corpora usado; alinhamento ao segmento de palavra (ou tradução "por exemplo"); classificação automática da qualidade de um par de traduções.

1 Introdução

Este trabalho, realizado sobre o alinhamento de corpora paralelos faz parte do projecto TerminUM[2] apresentado anteriormente nestas actas.

O processo de alinhamento de corpora paralelos é um dos mais importantes na concepção e tratamento destes. De facto, sem um qualquer nível de alinhamento entre dois textos paralelos, de pouco nos serve o seu paralelismo, a não ser para estudos constrativos de léxico e frequências.

O alinhamento de corpora paralelos é normalmente classificado em alinhamento à frase, à palavra e alinhamento ao caracter (este último muito pouco utilizado a não ser para proceder a um dos alinhamentos referidos anteriormente).

Neste documento vamos abordar o alinhamento à frase, à palavra e também à sequência de palavras (ou alinhamento de *chunks*), usando o pacote (ou banca de trabalho) NATools.

O alinhamento à palavra que vamos descrever depende do pré-alinhamento à frase de corpus. Da mesma forma, o alinhamento à sequência de palavras é dependente do pré-alinhamento à palavra. Desta forma, podemos descrever o processo de alinhamento como:

1. Segmentação (ou *tokenização*) dos textos, dividindo-os em parágrafos, frases e palavras (ver secção 2);
2. Alinhamento à frase usando um de dois alinhadores, como discutido na secção 3;
3. Alinhamento à palavra usando uma variante do Twente-Aligner, explicado na secção 4; Esta secção inclui também discussão dos resultados obtidos pelo alinhador, assim como algumas técnicas para os melhorar;
4. Aplicações dos dicionários de tradução obtidos a partir do alinhamento à palavra, das quais se destaca: o processo de alinhamento à sequência de palavras; tradução estatística ou "por exemplo"; classificação de traduções; assim como outras.

Finalmente, terminamos com uma secção de conclusões e trabalho futuro que apontam as linhas que este trabalho pretende seguir.

2 Segmentação

Inicialmente, os textos são segmentados em parágrafos e frases, usando técnicas comuns. Para a detecção de parágrafos usa-se especialmente o conhecimento da estrutura dos documentos que estamos a processar. A detecção de frases tentam encontrar-se sinais de pontuação que as indiciem. Esta detecção da pontuação tem de ser suficientemente inteligente para detectar abreviaturas, acrónimos, e-mails, URLs, e todo o tipo de escrita menos convencional.

Da mesma forma, a detecção e posterior divisão em palavras também é delicada. Não só é preciso ter em atenção os caracteres que existem (ou não) na língua que estamos a processar, como os caracteres que podem (ou não) fazer parte das palavras. Como exemplo, é mais comum o hífen ser usado em Português numa única palavra do que em Inglês em que, quase de certeza, irá ser um ponto de divisão de palavras.

3 Alinhamento à frase

Para alinhar os corpora paralelos à frase utilizamos duas ferramentas distintas: *easy-align*[8] ou o Vanilla Aligner[1, 5].

Não se pretende comparar os alinhadores em termos de qualidade ou velocidade. De facto, o *easy-align* é muito mais robusto do que o Vanilla Aligner só que com o inconveniente de não ser Software livre. O Vanilla Aligner é baseado em código aberto, pelo que pode ser distribuído livremente. Uma discussão mais aprofundada sobre o alinhamento à frase pode ser encontrado em [3].

3.1 Vanilla Aligner

O algoritmo usado por Church & Gale é baseado no tamanho das frases de cada um dos corpus, tentando alinhar frases, ou pares de frases, com tamanhos similares.

O alinhamento resultante é sempre da forma n para m em que $(n, m) \in \{(0, 1), (1, 0), (1, 1), (1, 2), (2, 1)\}$, dado que para detectar pares da forma $(1, 3)$ ou $(3, 1)$ seria

necessário desligar a detecção de pares como $(0,1)$ ou $(1,0)$, ou o algoritmo não iria conseguir distinguir pares da forma $(3,1)$ de sequências como $(2,1)(1,0)$.

3.2 Easy-Align

O *easy-align* faz parte do IMS Workbench[8]. Usa um método de alinhamento chamado linguístico. Procura palavras com baixa distância de edição (número de caracteres a ser aumentado e removido de forma a que as palavras se tornem iguais) e usa-as como âncoras no processo de alinhamento.

Além destas palavras é possível passar-lhe um dicionário bilingue com traduções para serem usadas como âncoras.

4 Alinhamento à palavra

Para o alinhamento à palavra são usados pares de ficheiros, previamente alinhados à frase. O alinhador é uma versão melhorada do Twente-aligner[7, 6], desenvolvido por Djoerd Hiemstra, na qual foram alterados vários algoritmos e estruturas de dados.

4.1 Processo de alinhamento

O processo de alinhamento é baseado na correlação de ocorrências de cada termo. O seu fluxo de dados é apresentado na figura 1 e explicado de seguida.

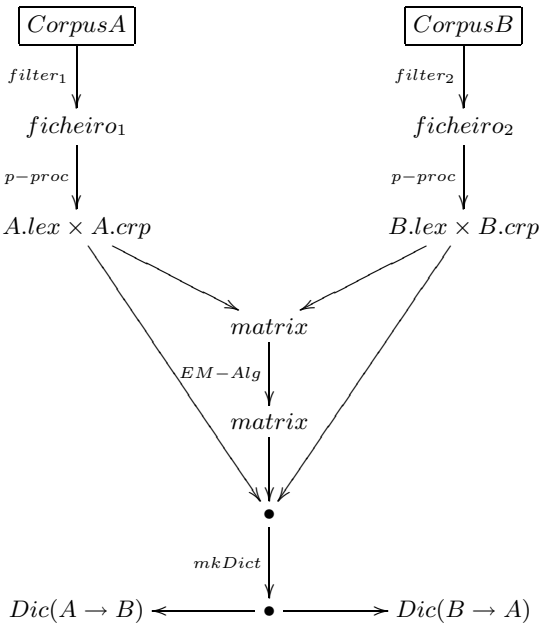

Figura 1: Fluxo de dados do alinhador à palavra do NATools

- o processo inicia com dois ficheiros alinhados à frase (em que cada frase é separada da outra por uma linha contendo apenas o caracter \$) a que iremos chamar *CorpusA* e *CorpusB*.
- o primeiro passo constituído por dois filtros a que chamamos $filter_1$ e $filter_2$ é o único dependente da língua dos corpora, razão que levou a dar-lhes nomes diferentes. O seu objectivo é segmentar as frases previamente alinhadas em palavras. No entanto, este passo veio a tornar-se útil para a realização de outras operações sobre o texto original como seja a lematização de verbos em Português (para aumentar a correlação entre as várias formas verbais portuguesas e, por exemplo, as pouquíssimas formas inglesas. Da mesma forma, para um corpus inglês pode ser útil remover os genitivos (possessivos) substituindo-os por um segment de caracteres especial (*token*);
- os ficheiros pré-processados irão ser analizados palavra por palavra. A cada palavra diferente encontrada é associando um identificador único nesse corpus (valor inteiro) e criado um índice de indirecção que permita rapidamente obter o identificador a partir da palavra e vice-versa. Dizemos que estes ficheiros armazenam o léxico dos corpora em questão e chamamos-lhes *A.lex* e *B.lex*. Além deste par de ficheiros, um outro par (*A.crp* e *A.crp*) é criado com cada um dos corpora codificados: em vez de conter a sequência de palavras que constituem o corpus, estes ficheiros contêm a sequência dos identificadores, o que transforma as complexas comparações entre palavras em simples comparações de inteiros;
- para o cálculo das co-ocorrências de palavras é usada uma matriz esparsa em que cada índice de linha corresponde a uma palavra no corpus de origem e cada índice de coluna corresponde a uma palavra no corpus destino. Desta forma, cada célula representa a relação $\mathcal{R}$ entre palavras de cada corpus ($w_\alpha \in \mathcal{A} \wedge w_\beta \in \mathcal{B}$) definida por:

  $$w_\alpha \mathcal{R} w_\beta \Leftrightarrow w_\alpha \in s^i_\alpha \subset \mathcal{A} \wedge w_\beta \in s^j_\beta \subset \mathcal{B} \wedge i = j$$

  em que s^n_γ é uma frase de cada um dos corpora e n é o número de ordem dessa frase. Sempre que um par de palavras co-ocorre o valor da célula correspondente é alterado convenientemente.
- o passo seguinte denominado na terminologia inglesa de EM-Algorithm (Entropy Maximization Algorithm-Algoritmo de maximização da entropia) tem como objectivo realçar na matriz as células de forte correlação, reduzindo o valor das outras. Para mais informações sobre este algoritmo aconselho a consulta de [9], uma bibliografia detalhada escrita por Ted Pederson.
- finalmente, a matriz é interpretada e os dicionários de tradução são criados. O sistema para a criação da matriz não é simétrico o que produz uma matriz assimétrica, implicando a criação de um dicionário de tradução da língua de origem para a língua de destino e um outro dicionário da língua de destino para a língua de origem.

O resultado deste alinhamento é um par de dicionários de tradução em que a cada palavra se faz corresponder um conjunto de possíveis traduções associadas à sua probabilidade de ser uma tradução correcta.

Podemos especificar matematicamente a estrutura deste dicionário como

$$w_\alpha \mapsto (\#occ \times w_\beta \mapsto P(\mathcal{T}(w_\alpha) = w_\beta))$$

do corpus de origem $\mathcal{C}_\alpha$ para o corpus de destino $\mathcal{C}_\beta$:

- $w_\alpha \in \mathcal{C}_\alpha \wedge w_\beta \in \mathcal{C}_\beta$;
- $\#occ$ é o número de ocorrências de w_α no corpus $\mathcal{C}_\alpha$;
- $P(\mathcal{T}(w_\alpha) = w_\beta)$ é a probabilidade de w_β ser uma tradução de w_α;

A tabela 1 mostra dois extractos dos dicionários resultantes do alinhamento da Bíblia.

Deus		**God**	
God	0.86	Deus	1.00
(null)	0.04		
God's	0.03		
He	0.01		
Yahweh	0.01		
him	0.01		
has	0.01		

gosta		**loves**	
loves	0.43	ama	0.67
detests	0.29	gosta	0.08
likes	0.29	amas	0.05
		estima	0.03
		conquista	0.02
		acabará	0.02
		curta	0.02

Tabela 1: Extracto do dicionário de alinhamento de Bíblia

Nestes extractos é de salientar:

- a palavra nula `(null)` aparece com alguma probabilidade como tradução possível de "Deus". Isto deve-se a frases em que o sujeito se subentenda;
- "God's" aparece como uma palavra. No entanto, o pré-processamento do corpus poderia substituir este género de construções por "God _GENITIVO_", o que iria aumentar a probabilidade da tradução correcta;
- no segundo exemplo, é curioso o uso de "loves" e "detests", ambos como sinónimos de "gosta". Na verdade, o "detests" deveria ser traduzido pelo "não gosta", mas o algoritmo de tratamento da matriz leva a que a correlação de "não" com "gosta" desapareça;

4.2 Análise de Resultados

As estruturas de dados e seus algoritmos associados foram alterados de forma a melhorar a eficiência a vários níveis. A tabela 4.2 mostra a comparação entre os tempos de processamento dos passos mais demorados no sistema de alinhamento.

Embora o tempo de processamento tenha diminuído para corpora grandes continuamos com tempos razoáveis como se pode verificar na tabela 3 para cinco corpora de tamanhos bastante diferentes. Fica aqui uma breve descrição dos mesmos:

TS Tom Sawyer de Mark Twain (PT–EN);

HP Harry Potter e a Pedra Filosofal (PT–EN);

	Twente	**NATools**
Análise do corpus	180 seg	4 seg
Inicial da matriz	390 seg	21 seg
Método iterativo	2128 seg	270 seg

Tabela 2: Comparação dos tempos de alguns dos passos do sistema de alinhamento para um corpus de 800 mil palavras (EN-PT)

IPC Corpus recolhido automaticamente da Internet `http://www.ipc.pt` (PT–EN);

Bib Bíblia (PT–EN);

EP Metade de um corpus de 6 milhões de palavras criado por Andrius Utka do Centro de Linguística de Corpus da Universidade de Birmingham com base em documentos do Parlamento Europeu;

	TS	**HP**	**IPC**	**Bib**	**EP**
mil palavras	77	94	118	805	3 500
Análise (seg)	0.5	1	1	5	67
Matriz (seg)	6	8	4	57	893
EM-Algorithm (seg)	42	73	44	468	5 523

Tabela 3: Times comparison for five different corpora

Torna-se aqui importante realçar o facto pelo qual apenas metade do corpus do Parlamento Europeu. Na verdade, na maquina usada (Pentium IV a 1.5Ghz, 512 Mb de RAM) a matriz de co-ocorrências de todo o corpus ocupa mais de 500 Mbytes. Não foi possível testar o alinhamento numa máquina com mais memória.

Este facto levou ao desenvolvimento de um método de adição de dicionários de tradução. Este método é apresentado na secção 4.4.

4.3 Alinhamento de tuplos

A diminuição drástica do tempo de alinhamento torna possível o alinhamento de corpora maior mas também a execução de novas experiências. Este é um caso de experiência bem sucedida: o alinhamento de tuplos de palavras.

No primeiro caso realizou-se o alinhamento de pares. Aplicou-se a cada corpora um pré-processador que junta palavras da seguinte forma:

```
Era uma vez um macaco .
```

seria substituído por:

```
BEGIN_Era Era_uma uma_vez
vez_um um_macaco macaco_. ._END
```

O alinhamento do corpus depois de processado aumenta o número de palavras diferentes para mais do dobro (pelo que o método pode não ser aplicável a corpora grandes). No entanto, este tipo de alinhamento mostrou-se bastante interessante para a extracção de termos multi-palavra (de tamanho dois).

A tabela 4 mostra dois extractos dos dicionários gerados ao alinhar a Bíblia em pares de palavras.

Foram realizadas algumas experiências com tuplos de tamanho maior que resultaram em matrizes de correlação enormes, e resultados práticos desanimadores.

Jesus Cristo		Christ Jesus	
Christ Jesus	0.67	Jesus Cristo	0.94
Jesus Christ	0.26	(null)	0.04
(null)	0.03	Cristo ,	0.01
Messiah ,	0.01		
Christ who	0.01		
the Messiah	0.01		

um pouco		a little	
a little	0.68	um pouco	0.54
(null)	0.19	(null)	0.27
a while	0.03	Pouco depois	0.06
me a	0.03	e ,	0.03
your company	0.02	uma criança	0.03
BEGIN Then	0.01	BEGIN Daqui	0.02

Tabela 4: Dicionários resultantes do alinhamento de pares de palavras

4.4 Alinhamento por fases

Como foi referido anteriormente, o facto de o alinhamento de corpora grandes não ser possível de realizar em máquinas com memória limitada, torna-se importante a soma de dicionários de tradução. Com este objectivo desenvolveu-se uma aplicação para somar dicionários.

A soma de dicionários é trivial no que respeita à soma do número de ocorrências de cada palavra mas a soma das probabilidades não pode ser realizada de forma tão simples.

A fórmula em questão deve:

- privilegiar as probabilidades do dicionário com maior número de ocorrências;
- privilegiar as probabilidades do dicionário em relação ao número de ocorrências em relação ao tamanho do corpus;

Dados estes objectivos, a fórmula usada foi:

$$\frac{\mathcal{P}_1(w_\alpha, w_\beta) \times \frac{\#_1(w_\alpha)}{\mathcal{S}_1} + \mathcal{P}_2(w_\alpha, w_\beta) \times \frac{\#_2(w_\alpha)}{\mathcal{S}_2}}{\frac{\#_1(w_\alpha)}{\mathcal{S}_1} + \frac{\#_2(w_\alpha)}{\mathcal{S}_2}}$$

onde:

- $\mathcal{P}_n(w_\alpha, w_\beta) = \mathcal{P}\left(\mathcal{T}\left(w_\alpha\right) = w_\beta\right)$ do dicionário n;
- $\#_n(w_\alpha)$ é o número de ocorrências de w_α no corpus de origem do dicionário n;
- $\mathcal{S}_n$ é o número total de palavras do corpus que deu origem ao dicionário, ou seja, para um dicionário $\mathcal{D}_n$ temos $\mathcal{S}_n = \sum_{w \in \mathcal{D}_n} \#_n(w)$.

O processo de alinhamento prevê verificar e número de palavras de cada corpus e de os dividir automaticamente para que o processo de alinhamento possa ser feito por fases.

Os corpora são analisados em relação ao número de palavras e caso necessário são divididos em porções. Cada porção é analisada criando um ficheiro de corpus para cada porção $(A_1.crp, B_1.crp), (A_2.crp, B_2.crp), \ldots, (A_n.crp, B_n.crp)$ mas apenas um ficheiro de léxico para cada corpus $(A.lex, B.lex)$. Isto deve-se ao facto de que se assim não for feito cada fatia alinhada iria ter identificadores das palavras diferentes o que levava à possibilidade de aparecimento de palavras repetidas nos dicionários (umas em minúsculas e outras em maiúsculas). Este processo está a ser revisto para facilitar a junção de dicionários não provenientes de alinhamento com os mesmos ficheiros de léxico.

O processo restante até à geração dos dicionários é igual à já apresentada. Finalmente, os dicionários gerados são somados para a criação de um único dicionário por sequência de línguas.

5 Aplicações

Muitas são as aplicações dos dicionários de tradução:

1. navegação sobre uma rede de palavras e respectivas traduções, com consulta directa no corpus que lhes deu origem;
2. classificação de traduções mediante os dicionários extraídos. Esta classificação pode ser usada para classificar memórias de tradução de acordo com a sua qualidade, assim como para que se possa apresentar extractos de corpora ordenados em relação à qualidade de tradução;
3. alinhamento ao segmento de palavras, que também dá origem à tradução por exemplo. Ou seja, realizar um tipo de tradução semelhante ao usado nos sistemas de memória de tradução, mas com segmentos mais curtos;

5.1 Navegação WEB

Os dicionários criados pelos métodos de alinhamento são grandes e difíceis de consultar. Para facilitar o seu estudo desenvolveu-se um conjunto de aplicações WEB para navegar sobre estes dicionários.

A figura 2 mostra o dicionário a ser consultado, de Inglês para Português, em relação à palavra "owl".

O sistema de navegação mostra dois níveis de tradução: para a palavra em causa mostra as suas possíveis traduções e, para cada uma destas, as traduções que aceita. No caso de a palavra que está a ser consultada (w_α) aparecer como tradução possível de uma das suas traduções (w_β), esta (w_β) aparecerá com uma cor diferente.

Da mesma forma, um sistema de cores é usado para salientar o valor da probabilidade de tradução: verde para probabilidades superiores a 70%, amarelo para as de 30% a 70% e para as restantes, a cor vermelha.

Esta consulta permite aceder às ocorrências da palavra no corpus como é mostrado na figura 3. Torna-se possível ao utilizador verificar quais os pares de frases que provavelmente terão dado resultados menos esperados.

5.2 Classificação de traduções

Dado que temos um dicionário de tradução com probabilidades de tradução entre duas palavras podemos usá-lo para tentar obter a probabilidade de tradução entre duas sequências de palavras.

Este método, a que chamamos "classificação de traduções" usa os dicionários calculados com um qualquer

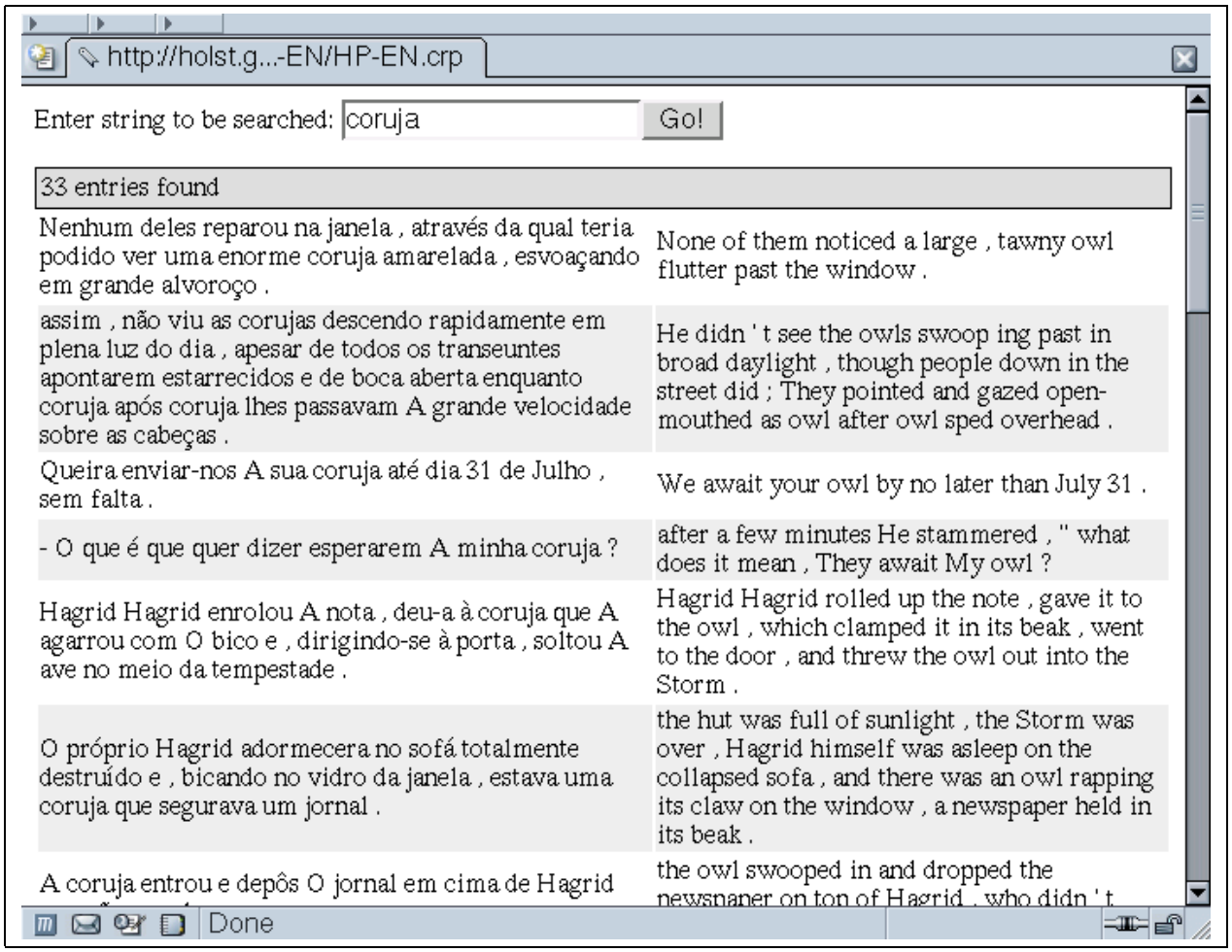

Figura 3: Consulta do corpora paralelo do Harry Potter

par de corpus nas línguas que queremos avaliar para classificar as traduções. É assim possível avaliar frases de um corpus paralelo sem o alinhar previamente à palavra.

O algoritmo usado inicia com um par de frases a ser avaliadas ($s_\alpha = w_{1,\alpha}w_{2,\alpha}\dots w_{n,\alpha}$ e $s_\beta = w_{1,\beta}w_{2,\beta}\dots w_{m,\beta}$). O método é calculado bidireccionalmente e é calculada uma média dos valores obtidos. Supondo que estamos a calcular de s_α para s_β, para cada palavra $w_{i,\alpha}$ são encontradas as suas possíveis traduções e, para cada uma delas (iniciando na com maior probabilidade de tradução) é verificado se esta se encontra em s_β. Caso exista, esse valor é somado. Caso não exista, é somado o valor 0. A soma é dividida finalmente pelo número de palavras da frase original (n).

A tabela 5 mostra os resultados deste método de avaliação para dois pares de frases.

Este valores podem ser úteis para:

- classificações de entradas numa memória de tradução;
- ordenamento de frases de um corpus paralelo de forma a que quando consultado sejam mostradas em primeiro lugar os pares com uma boa classificação;
- verificar, dados dois ficheiros, até que ponto podem ser considerados traduções um do outro;

Português	English	P(x)
Paulo, Apóstolo de Jesus Cristo por vontade e chamamento de Deus, e o irmão Sóstenes	From Paul, called to be an apostle of Christ Jesus by the will of God, and from Sosthenes, our brother	0.88
Pois em Jesus é que recebestes todas as riquezas, tanto da palavra como do conhecimento	For you have been fully enriched in him with words as well as with knowledge	0.18

Tabela 5: Exemplo de avaliação de traduções

5.3 Alinhamento ao segmento de palavras

Dado um segmento de palavras o seu alinhamento está a ser realizado com base no corpus paralelo existente e no algoritmo de avaliação de traduções apresentado na secção anterior. Todas as ocorrências do segmento de palavras é procurado no corpus. As frases em que o segmento ocorre e a sua respectiva tradução irão ser analisadas para tentar encontrar a tradução do segmento de palavras original.

Dado o tamanho do segmento original (n), é usado

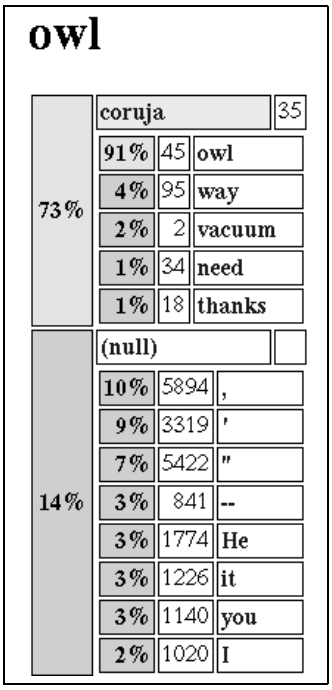

Figura 2: Navegação sobre o corpus do Harry Potter

um algoritmo de janela deslizante (três vezes, com $n-1$, n e $n+1$) para encontrar qual a janela com maior probabilidade de ser uma tradução do segmento, usando o método apresentado na secção anterior.

As três melhores janelas (com tamanhos diferentes) serão de novo avaliadas e escolhida a melhor tradução do segmento original.

A figura 4 mostra uma pequena linha de comando a alinhar segmentos de palavras sobre o corpus do parlamento europeu (EN-FR).

```
==> difficult situation

Using 6 occurrences (0.732864 seconds)
  situation difficile       - 0.8025
  situation très difficile  - 0.8025
  situation aussi difficile - 0.8025

==> sentenced to death

Using 1 occurrences (0.214145 seconds)
  condamné à mort - 0.443333333333333

==> final version

Using 7 occurrences (0.843922 seconds)
  version définitive - 0.5075
  définitive         - 0.09
  définitif          - 0.0875
```

Figura 4: Alinhamento de segmentos de palavras

Este método é usado não só para alinhamento ao segmento de palavras mas também para o que chamamos de "tradução por exemplo" ou "tradução estatística". É possível dotar um tradutor de um sistema que, para pequenas sequências de palavras as procure nos diversos corpora que tem disponíveis e proponha as respectivas traduções. À sequência é aumentada uma palavra (à direita) e removida a primeira (da esquerda) vontando-se a consultar o corpus. Esta sequência pode dar resultados interessantes.

No entanto, o método de alinhamento ao segmento é, por vezes, bastante demorado devido ao grande número de ocorrência do segmento em causa. Para resolver este problema são pré-calculados índices de qualidade de tradução para todas as frases do corpus. Assim, o método de alinhamento pode usar apenas as n melhores traduções.

6 Conclusões e trabalho futuro

Embora o trabalho no desenvolvimento do alinhador à palavra tenha sido feito na sua maioria por Djoerd Hiemstra, o aumento de eficiência que se conseguiu tornou possível novas metas, novas experiências. É de salientar que não fosse o programa ter como direito de cópia a licença GPL[4] (e portanto, ser software livre) não seria possível ser melhorado e distribuído por outra pessoa, e até com outro nome.

Embora existam outras ferramentas para alinhamento à frase (como o *easy-align*) e outras ferramentas para alinhamento à palavra, poucas são as que estão disponíveis livremente para serem usadas por qualquer pessoa.

O alinhamento de corpora paralelos é crucial para o desenvolvimento de ferramentas de tradução, não só como gerador de terminologia bilingue mas também como extractor de determinadas construções e respectivas traduções por exemplo.

Futuramente pretende-se utilizar de forma maciça o alinhamento à palavra a grandes corpora e das mais variadas fontes para proceder à análise dos respectivos dicionários antes e depois de serem somados. Para este exercício pretende-se alinhar:

- livros, romances, aventuras disponíveis em mais do que uma língua (como os vários Harry Potter, Tom Sawyer, Moby Dick e muitos outros);
- textos recolhidos de instituições que publicam documentação em mais do que uma língua (como o Parlamento Europeu);
- páginas extraídas de vários *sites* multilingues disponíveis na Internet;
- legendas de filmes (disponíveis pela Internet fora, embora por vezes com pouca qualidade);
- ficheiros de tradução de software (i18n);
- manuais variados

Referências

[1] Pernilla Danielsson and Daniel Ridings. Practical presentation of a "vanilla" aligner. In *TELRI Workshop in alignment and exploitation of texts*, February 1997.

[2] José João Dias de Almeida. Apresentação do projecto terminum. In *CP3A — Corpus Paralelos, Aplicações e Algoritmos Associados*, 2003.

[3] Helena de Medeiros Caseli. Alinhamento sentencial de textos paralelos português-inglês. Master's thesis, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, Fevereiro 2003.

[4] Free Software Foundation, Inc. GNU General Public License, June 1991.

[5] 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.

[6] Djoerd Hiemstra. Multilingual domain modeling in twenty-one: automatic creation of a bi-directional lexicon from a parallel corpus. Technical report, University of Twente, Parlevink Group, 1998.

[7] Djoerd Hiemstra. Using statistical methods to create a bilingual dictionary. Master's thesis, Department of Computer Science, University of Twente, August 1996.

[8] Oliver Christ & Bruno M. Schulze & Anja Hofmann & Esther König. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stutgart, March 8, 1999 (CQP V2.2).

[9] Ted Pederson. The EM Algorithm: Selected readings. Unpublished notes to accompany the panel discussion on the EM-Algorithm.