Alinhamento Sentencial e Lexical de Córpus Paralelos: Recursos para a Tradução Automática

Helena de Medeiros Caseli[1], Maria das Graças Volpe Nunes[1]

[1]Núcleo Interinstitucional de Lingüística Computacional – NILC-ICMC-USP
Caixa Postal 668P – 13.560-970 – São Carlos – SP – Brasil

{helename, gracan}@icmc.usp.br

Abstract. *Parallel texts – texts in one language and their translation in other – and aligned parallel texts – with identification of translation correspondences – are becoming more and more important for many NLP applications, mainly, machine translation. In this paper we describe some experiments carried out on sentence and lexical alignment of Portuguese-English parallel texts from differents genres: scientific, law and journalistic. The linguistic and computational resources and the knowledge derived from these experiments are very important for future work in machine translation field.*

Keywords. *Parallel corpora; sentence alignment of parallel texts; lexical alignment of parallel texts; machine translation.*

Resumo. *Textos paralelos – textos em uma língua e sua tradução para outra – e textos paralelos alinhados – com marcas que identificam as correspondências – estão se tornando cada vez mais importantes em muitas aplicações de PLN, principalmente, na tradução automática. Neste artigo são apresentados alguns experimentos realizados no alinhamento sentencial e lexical de textos paralelos português-inglês de gêneros diferentes: científico, jurídico e jornalístico. Os recursos lingüísticos e computacionais e o conhecimento derivado desses experimentos são de grande importância para projetos futuros na área de tradução automática.*

Palavras-chave. *Córpus paralelos; alinhamento sentencial de textos paralelos; alinhamento lexical de textos paralelos; tradução automática.*

1. Introdução

Nos últimos anos, a utilização de textos paralelos e textos paralelos alinhados tem se tornado cada vez mais freqüente em inúmeras aplicações de Processamento de Língua Natural (PLN). Os textos paralelos, segundo a terminologia estabelecida pela comunidade de lingüística computacional, são textos acompanhados de sua tradução em uma ou várias línguas. Se esses textos possuírem marcas que identificam os pontos de correspondência entre o texto original (texto fonte) e sua tradução (texto alvo) eles são considerados alinhados.

Métodos automáticos de alinhamento de textos paralelos podem ser usados para encontrar os pontos de correspondências entre os textos fonte e alvo. Tais pontos de

correspondência podem ser determinados em diferentes níveis que vão desde os textos completos até suas partes constituintes como: capítulos, seções, parágrafos, sentenças, palavras e até mesmo caracteres. Os níveis de alinhamento mais estudados na atualidade são o de sentenças e o de unidades lexicais (palavras ou multipalavras), também chamados de alinhamento sentencial e lexical, respectivamente.

O processo automático de alinhamento de textos paralelos, resumidamente, pode ser entendido como a "busca", no texto alvo, de uma ou mais sentenças (ou unidades lexicais) que correspondam à tradução de uma dada sentença (ou unidade lexical) no texto fonte. Na Figura 1 tem-se um exemplo de um par de textos paralelos alinhados sentencial e lexicalmente no qual o texto fonte (em português) é apresentado à esquerda e o texto alvo (em inglês), à direita. Nesta figura, as sentenças estão separadas de acordo com o alinhamento sentencial (indicado pelas setas) e os números sobrescritos a cada palavra indicam o alinhamento lexical. No alinhamento lexical, o caractere "-" indica que a palavra fonte (ou alvo) não está alinhada com nenhuma palavra alvo (ou fonte), ou seja, o alinhamento lexical é do tipo 1:0 (ou 0:1).

Texto Fonte (em português)	**Texto Alvo (em inglês)**
Este[1] trabalho[2] apresenta[3] requisitos[4] funcionais[5] identificados[6] no[7] processo[8] de[9] Engenharia[9] Reversa[10] de[-] Software[11] que[12] possam[13] ser[14] suportados[15] por[16] um[17] Sistema[18] Hipertexto[19].	This[1] paper[2] discusses[3] the functional[5] requirements[4] identified[6] in[7] the[7] software[11] reverse[10] engineering[9] process[8] which[12] can[13] be[14] supported[15] by[16] a[17] hypertext[19] system[18].
Por[1] meio[2] da[3] modelagem[4] conceitual[5] e[6] navegacional[7] do[-] domínio[-] de[8] informações[9] relativas[10] ao[11] método[12] de[13] engenharia[13] reversa[14] Fusion-RE/I[15], foram[16] estabelecidos[16] os[17] requisitos[18] funcionais[19] de[20] um[21] aplicativo[22] hipermídia[23] de[24] suporte[25]ao[26] método[27], de[28] forma[28] a[28] nortear[29] o[30] engenheiro[31] de[-] software[32] responsável[33] pelo[34] processo[35] de[36] engenharia[36] reversa[37] e[38] possibilitar[39] o[40] acompanhamento[40] da[41] evolução[42] desse[43] processo[44].	By[1] means[2] of[3] a[3] conceptual[5] and[6] navigational[7] modeling[4] of[8] information[9] related[10] to[11] the[11] reverse[14] engineering[13] method[12] Fusion-RE/I[15], we[16] established[16] the[17] functional[19] requirements[18] of[20] a[21] hypermedia[23] application[22] to[24] support[25] the[26] method[27].
	Our[28] purpose[28] is[28] to[29] offer[29] guidelines[29] to[29] the[30] software[32] engineer[31] in[33] charge[33] of[34] the[34] reverse[37] engineering[36] process[35] and[38] to[39] make[39] possible[39] to[40] follow[40] the[41] evolution[42] of[43] this[43] process[44].

Figura 1 – Par de textos paralelos alinhados sentencial e lexicalmente.

Nesse contexto, este artigo apresenta o processo de construção de córpus paralelos para a realização de experimentos com métodos automáticos de alinhamento sentencial e lexical de textos paralelos. Assim, a próxima seção (Seção 2) traz uma descrição dos córpus paralelos construídos para os experimentos apresentados aqui; a Seção 3 apresenta os resultados desses experimentos e a última Seção (4), traz algumas conclusões deste artigo e propostas de trabalhos futuros.

2. Córpus Paralelos

Os córpus paralelos construídos para os experimentos com métodos de alinhamento sentencial e lexical são compostos por textos de três gêneros: científico, jurídico e jornalístico. O córpus científico possui 65 pares de resumos e *abstracts* de trabalhos acadêmicos da área de computação desenvolvidos no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP) em São Carlos[1], apresentados na forma de artigos

publicados em revistas especializadas, monografias de qualificação de mestrado, dissertações de mestrado e teses de doutorado e pertencentes a subdomínios variados da Computação.

Esse conjunto inicial, na verdade, foi dividido em dois: o córpus autêntico e o córpus pré-editado. O primeiro é formado pelos 130 textos na forma em que foram originalmente redigidos, sem nenhuma alteração em sua forma ou em seu conteúdo. O pré-editado, por sua vez, também possui os mesmos 130 textos, porém com correções e alterações feitas por um tradutor humano para a eliminação de ambigüidades, equívocos e erros de gramática e/ou tradução para o inglês[2]. Essa subdivisão do córpus em autêntico e pré-editado foi realizada para analisar o impacto da presença de ruídos (erros) no desempenho dos métodos de alinhamento sentencial de textos paralelos. O córpus científico autêntico possui 855 sentenças e 21.432 palavras enquanto o pré-editado contém 849 sentenças e 21.492 palavras. Sendo que, em ambos os córpus, há uma média de 7 sentenças/texto e 25 palavras/sentença.

O córpus jurídico, por sua vez, é composto por 4 pares de textos paralelos extraídos da documentação oficial da ALCA (Área de Livre Comércio das Américas)[3] num total de 725 sentenças e 22.069 palavras. Cada texto tem, em média, 91 sentenças e cada sentença, 30 palavras. Finalmente, o córpus jornalístico é composto por 8 pares de artigos do jornal "*The New York Times*", disponíveis na *web* em inglês e em português[4]. O córpus jornalístico possui 492 sentenças e 11.767 palavras, sendo que cada texto tem, em média, 30 sentenças e cada sentença, em média, 25 palavras.

A Tabela 1 detalha o número de palavras em cada córpus citado e em cada um dos idiomas envolvidos (português e inglês).

Tabela 1. Número de palavras (em português e inglês) nos córpus

Número de Palavras	Córpus Científico		Córpus Jurídico	Córpus Jornalístico
	Autêntico	Pré-editado		
Português	11.349	11.306	11.217	6.008
Inglês	10.083	10.186	10.852	5.759
Total	**21.432**	**21.492**	**22.069**	**11.767**

Além desses quatro córpus paralelos iniciais, uma versão corretamente alinhada sentencial e lexicalmente também foi gerada por um especialista humano para servir como referência (córpus de referência) na comparação dos alinhamentos produzidos pelos métodos automáticos.

3. Alinhamento de Textos Paralelos

Os textos paralelos dos córpus apresentados na Seção 2 foram alinhados pelos métodos automáticos de alinhamento sentencial e lexical e comparados com os textos alinhados do córpus de referência utilizando duas métricas: precisão e cobertura. A precisão mede o número de alinhamentos corretos em relação ao número de alinhamentos sugeridos pelo método, enquanto que a cobertura indica o número de alinhamentos corretos em relação ao número de alinhamentos existentes no córpus de referência. Ambas as métricas variam entre 0 e 1, sendo que um valor próximo do 1 indica um melhor o alinhamento.

3.1. Alinhamento Sentencial de Textos Paralelos

Os cinco métodos de alinhamento sentencial avaliados foram: GC (Gale & Church, 1991), GMA e GSA+ (Melamed, 2000), Piperidis et alli (2000) e TCA (Hofland, 1996).

O método GC (iniciais dos autores) baseia-se em um modelo estatístico simples de tamanhos de sentenças (em caracteres). A idéia principal por trás desse método é que o tamanho da tradução é proporcional ao tamanho da sentença fonte e, portanto, sentenças grandes tendem a ter traduções grandes enquanto que sentenças pequenas tendem a ter traduções pequenas. O GC é o método mais referenciado na literatura e apresenta uma precisão muito boa considerando-se sua simplicidade.

Os métodos GMA e GSA+ usam reconhecimento de padrão para encontrar alinhamentos entre as sentenças. Dois algoritmos são utilizados nesse processo: o SIMR (*Smooth Injective Map Recognizer*) e o GSA (*Geometric Segment Alignment*). O SIMR produz pontos de correspondência (alinhamentos entre palavras) e o GSA alinha os segmentos (sentenças) baseando-se nesses pontos de correspondência e em informações sobre as fronteiras dos segmentos. A diferença entre os métodos GMA e GSA+ é que, no primeiro, o SIMR considera apenas palavras cognatas para determinar os pontos de correspondência e, no segundo, uma lista de palavras âncoras[5] também é utilizada.

O método de Piperidis et alli (2000) baseia-se no ponto crítico da tradução: a preservação do significado. Nesse sentido, o critério de alinhamento desse método está relacionado à quantidade de palavras de classe aberta (verbos, substantivos, adjetivos e advérbios): duas sentenças só são alinhadas se possuírem quantidades similares de palavras de classe aberta.

Por fim, o TCA (*Translation Corpus Aligner*) utiliza vários critérios para encontrar a melhor correspondência entre as sentenças fonte e alvo: lista de palavras âncoras, palavras com iniciais maiúsculas (candidatas a nomes próprios), caracteres especiais (? e !), palavras cognatas e tamanho das sentenças.

Os valores de precisão e cobertura para cada um dos cinco métodos nos quatro córpus de teste (descritos na Seção 2) são apresentados nas Tabela 2. É importante citar, aqui, que o córpus jornalístico só foi utilizado na avaliação dos três métodos de melhor desempenho nos experimentos com os córpus científico e jurídico.

Tabela 2. Precisão e Cobertura dos Métodos de Alinhamento Sentencial

Córpus	Métrica	GC	GMA	GSA+	Piperidis et al	TCA
Científico Autêntico	**Precisão**	0,9125	0,9485	0,9507	0,8589	0,9017
	Cobertura	0,9012	0,9556	0,9531	0,8716	0,9062
Científico Pré-editado	**Precisão**	0,9759	0,9904	0,9904	0,9784	0,9420
	Cobertura	0,9736	0,9928	0,9928	0,9784	0,9375
Jurídico	**Precisão**	0,9917	0,9876	0,9876	0,9833	1,0000
	Cobertura	0,9890	0,8788	0,8788	0,9725	1,0000
Jornalístico	**Precisão**	-	0,8788	0,8832	-	0,9190
	Cobertura	-	0,8571	0,8571	-	0,9507

Em relação aos dados da Tabela 2, pode-se concluir que a maioria dos métodos de alinhamento sentencial avaliados obteve uma precisão de acordo com a relatada na literatura

para outras línguas (acima de 95%). Sendo que os menores valores foram obtidos nos córpus científico autêntico e jornalístico.

A hipótese de que a presença de ruídos atrapalha o alinhamento foi confirmada como pode ser observado pelos valores de precisão que, em todos os métodos, foram maiores no córpus científico sem ruídos (pré-editado) – em média 97,54% – do que no córpus com ruídos (autêntico) – em média 91,45% – confirmando-se o que já havia sido relatado na literatura (Gaussier et al., 2000).

3.2. Alinhamento Lexical de Textos Paralelos

Com relação ao alinhamento lexical de textos paralelos, os métodos avaliados foram: SIMR (Melamed, 2000) e LWA (Ahrenberg et al., 2002). O SIMR é o mesmo método usado no alinhamento sentencial (GMA) apresentado na Seção 3.1. Este método considera apenas palavras cognatas no alinhamento lexical e não alinha unidades multipalavras.

O LWA (*Linköping Word Aligner*), por sua vez, baseia-se na informação de co-ocorrência das palavras e alguns módulos lingüísticos para encontrar as correspondências entre as unidades léxicas (palavras e multipalavras) fonte e alvo. Três módulos lingüísticos foram implementados responsáveis pela categorização (etiquetação de *part-of-speech*) das palavras, marcação de unidades multipalavras a partir de listas de unidades multipalavras definidas previamente e demarcação de uma região (uma janela de *n* palavras à esquerda e *n* palavras à direita) dentro da qual as correspondências são procuradas.

Na avaliação dos métodos de alinhamento lexical, o córpus científico autêntico não foi utilizado, uma vez que a razão para a qual ele foi criado – testar a hipótese de que a presença de ruídos afeta o alinhamento – já havia sido analisada nos experimentos com o alinhamento sentencial. Assim, a Tabela 3 resume os valores de precisão e cobertura para os dois métodos de alinhamento lexical avaliados nos três córpus de teste.

Tabela 3. Precisão e Cobertura dos Métodos de Alinhamento Lexical

Córpus	Métrica	SIMR	LWA
Científico Pré-editado	**Precisão**	0,9383	0,5888
	Cobertura	0,1832	0,6514
Jurídico	**Precisão**	0,9561	0,6215
	Cobertura	0,2000	0,5983
Jornalístico	**Precisão**	0,9101	0,5184
	Cobertura	0,1679	0,5938

De acordo com os valores da Tabela 3 é possível notar que o SIMR apresentou uma grande diferença entre precisão (em média 93,48%) e cobertura (em média 18,37%); enquanto o LWA se manteve mais estável (57,62% de precisão e 61,6% de cobertura, em média). Além disso, assim como no alinhamento sentencial, no lexical a precisão no córpus jornalístico foi menor (91,01% para o SIMR e 51,84% para o LWA) do que nos outros.

4. Conclusões e Trabalhos Futuros

Este artigo apresentou os resultados de experimentos realizados com métodos de alinhamento sentencial e lexical para córpus paralelos português-inglês, provenientes de gêneros diferentes. Os recursos lingüísticos e computacionais, bem como o conhecimento

adquirido nesses experimentos, são de grande importância para diversas aplicações de PLN, principalmente, no que diz respeito a *Example-Based Machine Translation* (EBMT).

Nesse sentido, algumas propostas de trabalhos futuros são a construção de outros córpus paralelos e paralelos alinhados nos idiomas português, inglês e espanhol (possivelmente de outros gêneros e domínios); e a utilização dos textos paralelos alinhados na extração de regras de tradução, de modo automático.

---

[1] A identificação e aquisição dos textos foram feitas por Feltrim et alli (2001).

[2] Detalhes sobre o processo de coleta e pré-edição dos textos dos córpus autêntico e pré-editado podem ser obtidos em (Martins et al., 2001).

[3] Disponíveis no site oficial da ALCA (http://www.ftaa-alca.org/alca_e.asp).

[4] Em inglês: http://www.nytimes.com. Em português:http://ultimosegundo.ig.com.br/useg/nytimes.

[5] Uma lista de palavras que são a tradução uma da outra, ou seja, cada entrada dessa lista é uma palavra na língua fonte com uma ou mais palavras na língua alvo que correspondem a sua tradução.

Referências

AHRENBERG, L.; ANDERSSON, M.; MERKEL, M. A system for incremental and interactive word linking. In: *Third International Conference on Language Resources and Evaluation* (LREC 2002), Las Palmas, p.485-490, 2002.

FELTRIM, V.D.; NUNES, M.G.V.; ALUÍSIO, S.M. *Um corpus de textos científicos em português para a análise da estrutura esquemática*. Série de Relatórios do NILC. NILC-TR-01-4. 2001.

GALE, W.A.; CHURCH, K.W. A program for aligning sentences in bilingual corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (ACL), Berkley, p.177-184, 1991.

GAUSSIER, E.; HULL, D.; AÏT-MOKTHAR, S. Term alignment in use: Machine-aided human translation. In: VÉRONIS, J. (ed.). *Parallel text processing*. s.l.: Kluwer Academic Publishers, p.253-74, 2000.

HOFLAND, K. A program for aligning English and Norwegian sentences. In HOCKEY, S., IDE, N., PERISSINOTTO, G. (eds.). *Research in Humanities Computing*, Oxford, Oxford University Press, p.165-178, 1996.

MARTINS, M.S; CASELI, H.M.; NUNES, M.G.V. *A construção de um corpus de textos paralelos inglês-português*. Série de Relatórios do NILC. NILC-TR-01-5. 2001.

MELAMED, I.D. Pattern recognition for mapping bitext correspondence. In VÉRONIS, J. (ed.). *Parallel text processing*. s.l.: Kluwer Academic Publishers, p.25-47, 2000.

PIPERIDIS, S., PAPAGEORGIOU, H., BOUTSIS, S. From sentences to words and clauses. In VÉRONIS, J. (ed.). *Parallel text processing*. s.l.: Kluwer Academic Publishers, p.117-138, 2000.