O Projeto PESA:

Alinhamento Sentencial de Textos Paralelos Português-Inglês

Helena de Medeiros Caseli
Maria das Graças Volpe Nunes
Núcleo Interinstitucional de Lingüística Computacional – NILC
ICMC-USP - São Carlos
{helename, mdgvnune@icmc.sc.usp.br}

Resumo

O alinhamento sentencial de textos paralelos é uma subárea de PLN (Processamento de Linguagem Natural) que vem despertando o interesse da comunidade científica devido, principalmente, ao grande número de aplicações para as quais pode ser útil. Nesse contexto está inserido o PESA (*Portuguese-English Sentence Alignment*), um projeto que visa estudar, implementar e avaliar diferentes técnicas de alinhamento sentencial de textos paralelos escritos em português brasileiro e em inglês. Trata-se do primeiro projeto dessa natureza a envolver o português brasileiro. Além de apresentar as características do PESA e suas etapas, este artigo demonstra a relevância desse projeto enfatizando as contribuições por ele geradas.

1 Introdução

O alinhamento de textos paralelos é uma das áreas de Processamento de Linguagem Natural (PLN) que tem recebido mais atenção nos últimos anos, devido, principalmente, ao grande número de aplicações para as quais pode ser útil. Entre elas podemos citar: a tradução automática, o aprendizado de idiomas, a construção de léxicos bilíngües e a extração de terminologia.

Outro motivo para esse interesse crescente na área de alinhamento é o aumento do número de textos paralelos – textos acompanhados de sua tradução em uma ou mais línguas – com o advento da Internet. Os textos paralelos diferem daqueles denominados textos comparáveis por serem, estes últimos, textos escritos sobre o mesmo domínio, em línguas diferentes, mas que não são, necessariamente, traduções mútuas.

O processo de alinhamento de dois (ou mais) textos paralelos baseia-se na tarefa de encontrar as correspondências entre o texto original (texto fonte) e sua tradução (texto alvo).

Essas correspondências podem existir em diferentes níveis de resolução: do nível do texto completo até níveis menores como parágrafos, sentenças, palavras e caracteres. Os mais estudados são os níveis sentencial e lexical nos quais alinham-se sentenças e palavras, respectivamente.

Nesse contexto, o alinhamento sentencial de textos paralelos foi escolhido como objeto de estudo no projeto PESA (*Portuguese-English Sentence Alignment*)[1]. O PESA visa estudar, implementar e avaliar diversas técnicas (ou métodos) de alinhamento sentencial de textos paralelos. Os métodos escolhidos para a implementação foram divididos em três grupos de acordo com os critérios utilizados no alinhamento: empíricos, lingüísticos ou híbridos.

Os métodos empíricos são aqueles que não utilizam nenhum tipo de conhecimento a respeito das línguas envolvidas no processo de alinhamento. Os lingüísticos, por sua vez, utilizam informações específicas sobre as línguas em questão. Já os híbridos mesclam as duas abordagens anteriores com alguns critérios de alinhamento que utilizam informações lingüísticas e outros que não.

Para a avaliação desses métodos serão utilizados textos paralelos que são resumos e *abstracts* de trabalhos na área de computação desenvolvidos no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo em São Carlos (ICMC-USP-SC). Esses textos, 65 no total, serviram de base para a geração dos *corpora* de teste e de referência usados, respectivamente, para testar e avaliar os métodos de alinhamento. Esses *corpora* fazem parte dos recursos lingüísticos construídos para o projeto PESA e são apresentados na próxima seção juntamente com detalhes do projeto e suas fases já concluídas (Subseção 2.1) e ainda em andamento (Subseção 2.2). Também são apresentadas as principais contribuições do PESA em uma breve conclusão na Seção 3.

2 O projeto PESA

A escolha do alinhamento sentencial para estudo nesse projeto baseou-se no fato de esse possuir uma maior precisão em relação ao alinhamento lexical (o outro nível mais estudado) como apontado em (Véronis e Langlais, 2000). Além da precisão média acima de 95% para esse tipo de alinhamento, outro fator importante na escolha do nível de resolução sentencial foi o fato de o alinhamento lexical apresentar dificuldades especiais de tratamento consideradas sofisticadas para um primeiro trabalho nessa área.

O PESA encontra-se em desenvolvimento e, a partir de um levantamento bibliográfico

[1] Em http://nilc.icmc.sc.usp.br "Projects/PESA".

dos métodos mais referenciados na literatura em cada uma das classes de métodos já citadas - empíricos, lingüísticos e híbridos, optou-se pela implementação dos métodos de alinhamento sentencial de textos paralelos mostrados na Tabela 1.

Tabela 1 – Métodos de alinhamento sentencial selecionados para implementação no projeto PESA.

Método	Tipo	Sigla
Gale e Church (1991, 1993)	Empírico	GC
Melamed (2000)	Empírico	SIMR/GSA
Piperidis, Papageorgiou e Boutsis (2000)	Lingüístico	Piperidis *et al*
Melamed[2] (2000)	Híbrido	GSA+
Hofland (1996)	Híbrido	TCA

A escolha desses métodos foi feita, principalmente, com base na precisão apresentada por eles em outras línguas. Tentou-se também selecionar métodos que possuíssem critérios de alinhamento diferentes entre si, possibilitando, assim, uma análise mais ampla desse fator.

De modo geral, o PESA pode ser dividido em cinco etapas (ou fases): levantamento bibliográfico, construção dos recursos lingüísticos, implementação, teste e avaliação dos métodos de alinhamento sentencial de textos paralelos. As três últimas referem-se aos métodos da Tabela 1 e são as únicas que ainda não foram concluídas.

A primeira etapa resultou na seleção dos métodos de alinhamento sentencial de textos paralelos já citados. A segunda, também já finalizada, produziu os recursos lingüísticos necessários para o PESA: uma lista de palavras âncoras e os *corpora* de teste e de referência. A lista de palavras âncoras é formada por uma entrada na língua fonte e suas correspondentes traduções na língua alvo e é indispensável nos métodos híbridos escolhidos. Os *corpora* de teste são alinhados pelos métodos de alinhamento sentencial de textos paralelos e os resultantes, os *corpora* alinhados, são comparados com os *corpora* de referência. Esses últimos são os mesmos *corpora* de teste, porém com indicações de alinhamento entre as sentenças dos textos fonte e alvo inseridas por um especialista humano em um processo semi-automático. Mais detalhes sobre os recursos lingüísticos do PESA são apresentados na Subseção 2.1.

As outras três fases restantes no projeto PESA – implementação, teste e avaliação – são apresentadas na Subseção 2.2. A primeira engloba a implementação dos métodos selecionados, na ordem apresentada na Tabela 1, ou seja, de acordo com a classe a qual eles pertencem. Depois de implementado cada método será testado com a submissão dos *corpora* de teste e os resultados retornados, os textos paralelos de entrada alinhados, serão comparados com os textos dos *corpora* de referência na fase de avaliação.

[2] Este é o mesmo método empírico, porém, com a adição de um recurso lingüístico correspondente à lista de palavras âncoras.

2.1 Recursos Lingüísticos

Antes da implementação dos métodos de alinhamento sentencial de textos paralelos, é necessário construir os recursos lingüísticos utilizados nas fases de implementação, teste e avaliação. Tais recursos são: a lista de palavras âncoras, os *corpora* de teste e os *corpora* de referência.

A lista de palavras âncoras é uma lista de palavras na qual uma entrada na língua fonte (o português brasileiro) possui uma ou mais traduções na língua alvo (o inglês). Esse recurso lingüístico é utilizado pelos métodos híbridos como um dos critérios para alinhar as sentenças. A construção dessa lista baseou-se na análise das palavras mais freqüentes nos textos de três *corpora* da área de computação:

- O *corpusDT* formado por 52 textos científicos (dissertações e teses) da área de Ciências da Computação escritos em português brasileiro. Esse *corpus* é um dos resultados do projeto SciPo[3], desenvolvido no NILC (Feltrim, Nunes e Aluísio, 2001).
- O *corpus* cmp-lg (*Computation and Language*) formado por 183 artigos científicos escritos em inglês apresentados nas conferências da *Association for Computational Linguistics* (ACL) e preparado pela corporação MITRE[4].
- O *corpus* HCI composto por 102 introduções de trabalhos específicos da área de HCI, escritos em inglês. Esse *corpus* também está disponível no NILC (Silva, 1999).

As listas com as palavras mais freqüentes geradas com o auxílio da ferramenta computacional de processamento de textos WordSmith[5] foram analisadas e uma lista final com cerca de 250 entradas foi gerada para o par de línguas português-inglês no domínio da computação.

Os outros recursos lingüísticos, os *corpora*, são variações de um conjunto de 65 pares de resumos e *abstracts* (textos paralelos) de trabalhos na área de computação desenvolvidos no ICMC-USP-SC. Esse conjunto inicial, na verdade, foi dividido em dois: o *corpus* autêntico e o *corpus* pré-editado. O primeiro é formado pelos 65 textos paralelos na forma em que foram originalmente redigidos, sem nenhuma alteração quanto à sua forma ou conteúdo. O pré-editado, por sua vez, também é formado pelos mesmos 65 pares de textos, porém com correções, alterações e marcações feitas por um tradutor humano para a eliminação de ambigüidades, equívocos e erros de gramática e/ou tradução para o inglês (Martins, Caseli e Nunes, 2001).

A divisão do *corpus* em autêntico e pré-editado foi feita com o intuito de analisar o impacto da qualidade do *corpus* no desempenho dos métodos de alinhamento. O interesse nessa

[3] Em http://www.nilc.icmc.sc.usp.br "Projects/SciPo".
[4] Em http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/cmp_lg.html.
[5] Em http://www.liv.ac.uk/~ms2928/wordsmith.

análise vem do fato de que, segundo a literatura consultada, o desempenho desses métodos é melhor em *corpus* sem ruídos (ou limpos), ou seja, sem erros gramaticais ou de tradução.

A partir dos *corpora* autêntico e pré-editado foram construídos os de teste e de referência. Os primeiros foram construídos com a inserção de marcações de fronteiras de parágrafos (e) e sentenças (<s> e </s>), de acordo com o padrão XML[6] (*Extensible Markup Language*), nos textos presentes nos *corpora* originais. Essas marcações foram inseridas automaticamente com o auxílio da TagAlign (Caseli, Feltrim e Nunes, no prelo), uma ferramenta de pré-processamento de textos desenvolvida no NILC. Os *corpora* gerados nesse processo receberam as siglas de CAT para o *corpus* autêntico de teste e CPT para o *corpus* pré-editado de teste.

Além desses, outros dois *corpora* foram produzidos para a fase de teste: o *corpus* autêntico de teste etiquetado morfologicamente (CATE) e o *corpus* pré-editado de teste etiquetado morfologicamente (CPTE); ambos resultantes de um processo de etiquetação morfológica feito com o TreeTagger (Schmid, 1994) para o português e para o inglês.

Os *corpora* de referência foram construídos a partir dos *corpora* de teste (CAT e CPT) e são resultado de um processo semi-automático de marcação de correspondências entre as sentenças dos textos fonte e alvo. Essas marcações foram inseridas com o auxílio da TagAlign e os *corpora* gerados receberam as siglas CAR – *corpus* autêntico de referência – e CPR – *corpus* pré-editado de referência. Por serem considerados corretos, esses *corpora* serão utilizados como parâmetro na comparação com os resultados dos métodos de alinhamento sentencial na fase de avaliação.

Uma descrição detalhada do processo de construção desses recursos lingüísticos não faz parte do escopo deste artigo, mas pode ser obtida em (Caseli e Nunes, no prelo). Por hora, a apresentação resumida dos mesmos é suficiente para o entendimento do projeto. A Tabela 2 traz um quadro com todos os recursos lingüísticos e as siglas atribuídas a eles.

Tabela 2 – Recursos lingüísticos do projeto PESA.

Recurso lingüístico	**Sigla**
Corpus autêntico de teste	CAT
Corpus pré-editado de teste	CPT
Corpus autêntico de teste etiquetado morfologicamente	CATE
Corpus pré-editado de teste etiquetado morfologicamente	CPTE
Corpus autêntico de referência	CAR
Corpus pré-editado de referência	CPR
Lista de palavras âncoras	LPA

[6] Em http://www.w3.org/XML.

Após a construção dos recursos lingüísticos as três fases restantes do PESA – implementação, teste e avaliação – poderão ser efetuadas. Elas são apresentadas na próxima seção.

2.2 Implementação, Teste e Avaliação

As três etapas de implementação, teste e avaliação dos métodos de alinhamento sentencial de textos paralelos serão executadas sequencialmente para cada método das classes previamente apresentadas, ou seja, os empíricos – GC e SIMR/GSA – o lingüístico – Piperidis *et al* – e os híbridos – GSA+ e TCA. Dessa forma, pretende-se analisar separadamente os métodos de mesma classe e, posteriormente, todos de uma maneira geral.

Todos os métodos selecionados para implementação no PESA utilizam os recursos lingüísticos apresentados na subseção anterior em suas fases de implementação, teste e avaliação, em maior ou menor grau, como mostrado na Tabela 3.

Tabela 3 – Recursos lingüísticos utilizados pelos métodos de alinhamento sentencial de textos paralelos.

Método	Tipo	Implementação	Teste	Avaliação
GC	Empírico	-	CAT e CPT	CAR e CPR
SIMR/GSA	Empírico	-	CAT e CPT	CAR e CPR
Piperidis *et al*	Lingüístico	-	CATE e CPTE	CAR e CPR
GSA+	Híbrido	LPA	CAT e CPT	CAR e CPR
TCA	Híbrido	LPA	CAT e CPT	CAR e CPR

Na fase de implementação, apenas os métodos híbridos utilizam o recurso lingüístico representado pela lista de palavras âncoras como um dos critérios de alinhamento. De modo geral, esses métodos procuram cada palavra presente nas sentenças fonte e alvo sob análise na lista e consideram o par (palavra_fonte, palavra_alvo) encontrado como um ponto de correspondência candidato. Os dois métodos híbridos que serão implementados utilizam a lista de palavras âncoras de maneira similar à descrita.

Depois de implementados, os métodos serão testados com a submissão dos textos paralelos existentes nos *corpora* de teste. Os *corpora* autêntico e pré-editado serão fornecidos separadamente para que se possa analisar o impacto da qualidade dos textos no desempenho dos métodos. Os *corpora* alinhados resultantes serão, então, comparados com os textos presentes nos *corpora* de referência de forma automática utilizando o módulo de avaliação da ferramenta computacional TagAlign. Esse módulo, ainda em fase de projeto, terá a função de calcular as três métricas utilizadas nessa área para a avaliação dos métodos de alinhamento de textos paralelos:

$$recall = \frac{NúmerodeAlinhamentos\,\mathrm{Pr}opostos}{NúmerodeAlinhamentosde\,\mathrm{Re}ferência} \quad (1)$$

$$precision = \frac{NúmerodeAlinhamentosCorretos}{NúmerodeAlinhamentos\,\mathrm{Pr}opostos} \quad (2)$$

$$F = 2\frac{recall \times precision}{recall + precision} \quad (3)$$

Recall pode ser entendida como uma medida de completude: quanto maior *recall*, maior a capacidade do método em encontrar alinhamentos. *Precision*, por sua vez, mede a consistência: quanto maior *precision*, maior o número de alinhamentos corretos dentre os encontrados. Já *F-measure* mede a distância entre *recall* e *precision*, e quanto maior *F-measure*, mais próximos são esses valores, portanto, maior a capacidade de encontrar alinhamentos sendo eles corretos.

Esse módulo também implementará uma metodologia de avaliação de algoritmos muito empregada em Aprendizado de Máquina (AM) e que utiliza o estimador *r-fold cross-validation* para estimar o erro ou a precisão de um algoritmo (Freedman, Pisani e Purves, 1998 *apud* Baranauskas, 2001). No *r-fold cross-validation*, os *n* exemplos[7] são divididos aleatoriamente em *r* conjuntos mutuamente exclusivos (*folds*) cada um com aproximadamente *n/r* exemplos. Um treinamento é efetuado com os exemplos contidos nos (*r*-1) *folds* e a hipótese induzida é testada no *fold* restante. Esse processo é executado *r* vezes e, em cada uma delas, um *fold* diferente é usado para teste. O erro é calculado como a média dos erros obtidos em cada um dos *r folds*.

Dessa forma, dado um algoritmo A (um método de alinhamento) e um conjunto de exemplos T (um *corpus* de teste), T é dividido em *r* partições. Para cada partição *i*, uma hipótese *hi* é induzida e um erro *err(hi)*, i = 1, 2, ..., r, é calculado. A partir daí calcula-se a média, a variância e o desvio padrão para todas as partições usando as equações (4), (5) e (6).

$$media(A) = \frac{1}{r}\sum_{i=1}^{r} err(hi) \quad (4)$$

$$\mathrm{var}(A) = \frac{1}{r}\left[\frac{1}{r-1}\sum_{i=1}^{r}(err(hi) - media(A))^2\right] \quad (5)$$

$$sd(A) = \sqrt{\mathrm{var}(A)} \quad (6)$$

Para decidir se um algoritmo A é melhor, ou mais robusto, do que um algoritmo B (com 95% de confiança) deve-se assumir o caso geral e determinar se a diferença entre os dois é

[7] Os exemplos, no caso da avaliação de métodos de alinhamento sentencial, são os textos dos *corpora* de teste. Cada um dos 65 bitextos de um dos *corpora* de teste (CAT, CPT, CATE ou CPTE) é um exemplo.

significativa ou não, assumindo uma distribuição normal (Weiss e Indurkhya, 1998 *apud* Baranauskas, 2001). Para isso calcula-se a média e o desvio padrão combinados de acordo com as equações (7) e (8), respectivamente, e a diferença absoluta, em desvios padrões, como mostrado na equação (9).

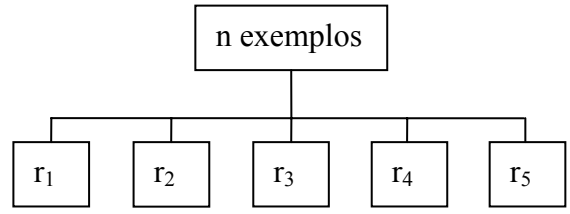
$$media(A-B) = media(A) - media(B) \tag{7}$$

$$sd(A-B) = \sqrt{\frac{sd(A)^2 + sd(B)^2}{2}} \tag{8}$$

$$ad(A-B) = \frac{media(A-B)}{sd(A-B)} \tag{9}$$

Assim, se ad(A – B) > 0 então B é melhor do que A. Além disso, se ad(A – B) ≥ 2 desvios padrões então B supera A com 95% de confiança. Porém, se ad(A – B) ≤ 0 então A é melhor do que B e se ad(A – B) ≤ -2 pode-se dizer que A supera B com 95% de confiança.

Dessa forma, o módulo de avaliação, usando *r-fold cross-validation*, dividirá os 65 pares de textos paralelos (exemplos) em, por exemplo, 5 *folds* com 13 exemplos cada. As 5 hipóteses geradas: h1, h2, h3, h4 e h5 serão utilizadas para determinar o erro médio do método de acordo com a equação (4). Uma ilustração desse processo é apresentada na Figura 1.



1. r2 + r3 + r4 + r5 => h1 err(h1) = h(r1) – h1
2. r1 + r3 + r4 + r5 => h2 err(h2) = h(r2) – h2
3. r1 + r2 + r4 + r5 => h3 err(h3) = h(r3) – h3
4. r1 + r2 + r3 + r5 => h4 err(h4) = h(r4) – h4
5. r1 + r2 + r3 + r4 => h5 err(h5) = h(r5) – h5

Figura 1 – Cálculo do erro de um método de alinhamento utilizando *5-fold cross-validation*.

O erro calculado poderá ser usado para calcular as outras medidas que permitem compará-lo com outros métodos.

3 Conclusões

Após uma descrição geral do projeto PESA apresentada neste artigo, pode-se perceber que o PESA é um projeto complexo que não visa apenas estudar, implementar e avaliar diversos métodos de alinhamento sentencial de textos paralelos, mas também produzir vários recursos como subprodutos. Entre esses recursos estão os *corpora* paralelos alinhados sentencialmente, a lista de palavras âncoras para o par de línguas português-inglês e as listas bilíngües de palavras consideradas pontos de correspondência candidatos geradas pelos métodos híbridos durante o processo de alinhamento.

Além dos recursos lingüísticos, úteis para outros projetos de processamento do português brasileiro no NILC e em instituições parceiras, há ainda recursos computacionais gerados para auxiliar tarefas específicas do projeto, como a ferramenta de pré-processamento de textos TagAlign com seus módulos de marcação, alinhamento e avaliação dos métodos.

Porém, o resultado prático mais importante desse projeto é o alinhador sentencial de textos paralelos para o português brasileiro e o inglês que tenha apresentado um bom resultado dentre todos os implementados. A existência de tal ferramenta computacional poderá favorecer projetos futuros gerando novos *corpora* alinhados ou mesmo servindo de base para a criação de um alinhador lexical de textos paralelos.

Nesse contexto, o PESA apresenta-se como uma fonte muito importante de recursos e conhecimento na área de alinhamento de textos paralelos e de Processamento de Linguagem Natural, de uma forma geral.

Referências Bibliográficas

Baranauskas, J.A. (2001). *Extração automática de conhecimento por múltiplos indutores*. Tese de Doutorado, ICMC-USP, São Carlos.

Caseli, H.M; Feltrim, V.D.; Nunes, M.G.V. (no prelo). *TagAlign: Uma ferramenta de pré-processamento de textos*. Série de Relatórios do NILC.

Caseli, H.M e Nunes, M.G.V. (no prelo). *A construção dos recursos lingüísticos para o projeto PESA*. Série de Relatórios do NILC.

Gale, W.A. e Church, K.W. (1991). A program for aligning sentences in bilingual corpora. In the *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.177-184, Berkley.

Gale, W.A. e Church, K.W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, vol. 19, no. 3, pp.75-102.

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC, NILC-TR-01-04.

Freedman, D.; Pisani, R.; Purves, R. (1998). *Statistics*, 3d ed. W.W.Norton & Company.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. In Hockey, S.; Ide, N.; Perissinotto, G. (eds.), *Research in Humanities Computing.* Oxford: Oxford University Press, pp.165-178.

Martins, M.S.; Caseli, H.M.; Nunes, M.G.V. (2001). *A construção de um corpus de textos paralelos português-inglês*. Série de Relatórios do NILC, NILC-TR-01-05.

Melamed, I.D. (2000). Pattern recognition for mapping bitext correspondence. In Véronis, J. (ed.), *Parallel text processing: Alignment and use of translation corpora*, Kluwer Academic Publishers, pp.25-47.

Piperidis, S.; Papageorgiou, H.; Boutsis, S. (2000). From sentences to words and clauses. In Véronis, J. (ed.), *Parallel text processing: Alignment and use of translation corpora*, Kluwer Academic Publishers, pp.117-138.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In the *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK.

Silva, M.H.B. (1999). *A abordagem de críticas para a construção de sistemas de aprendizado da escrita técnica*. Dissertação de Mestrado, ICMC-USP, São Carlos.

Weiss, S.M. e Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide*. San Francisco, CA: Morgan Kaufmann.

Véronis, J. e Langlais, P. (2000). Evaluation of the parallel text alignment systems: The ARCADE project. In Véronis, J. (ed.), *Parallel text processing: Alignment and use of translation corpora*, Kluwer Academic Publishers, pp.369-388.