VisualTCA: Uma Ferramenta Visual On-line para Alinhamento Sentencial de Textos Paralelos

Felipe Tassario Gomes, Thiago Alexandre Salgueiro Pardo, Helena de Medeiros Caseli

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
CP 668 – ICMC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.nilc.icmc.usp.br

felipe.gomes@gmail.com, {taspardo, helename}@icmc.usp.br

Resumo. *Apresenta-se, neste artigo, uma ferramenta visual on-line para alinhamento sentencial de textos paralelos, independentes de sua língua. Descrevem-se o funcionamento e as capacidades da ferramenta, assim como suas possíveis aplicações.*

1. Introdução

Textos paralelos são conjuntos de textos que englobam um texto original e a sua tradução em uma ou várias línguas. O processo de alinhamento de textos paralelos consiste em identificar a correspondência entre as partes de um texto (parágrafos, sentenças e/ou palavras) e sua tradução (Caseli, 2003).

O alinhamento é um importante processo na Lingüística e na Lingüística Computacional, principalmente em áreas que lidam e estudam as relações existentes entre diversas línguas. Ele é útil, por exemplo, para ajudar um tradutor profissional a verificar erros e fazer pequenas correções em textos já traduzidos por ele; ou para ajudá-lo a refinar, verificar a qualidade e corrigir traduções feitas por um tradutor automático. Na subárea da Lingüística de Córpus, em particular, o alinhamento pode ser a base para diversos estudos posteriores, como a correlação entre línguas, a evolução e o uso de termos estrangeiros em uma língua, etc. O alinhamento automático, cujos primeiros estudos datam por volta do início dos anos 90 (Véronis, 2000), ganhou grande importância nos últimos anos devido, principalmente, ao surgimento e avanço de várias aplicações que podem se beneficiar diretamente dele. Tais aplicações incluem memórias de tradução, tradução automática estatística, recuperação de informações em diferentes línguas e construção de dicionários bilíngües, entre outras. O alinhamento automático e visual, por sua vez, pode se mostrar útil nos vários cenários citados.

Apresenta-se, neste artigo, a VisualTCA, uma ferramenta visual on-line para alinhamento sentencial automático de textos paralelos. Baseada no método de alinhamento TCA (*Translation Corpus Aligner*) (Hofland, 1996), a ferramenta é independente de língua e encontra-se disponível para uso pela comunidade de pesquisa. Na Seção 2, a VisualTCA é introduzida e sua utilização é demonstrada. Na Seção 3, o sistema de alinhamento subjacente à ferramenta é brevemente descrito. Algumas considerações finais são feitas na Seção 4.

2. A Ferramenta

A ferramenta desenvolvida consiste em um alinhador sentencial automático on-line, cujos alinhamentos produzidos podem ser visualizados e estudados. Por ser on-line, a ferramenta não precisa ser instalada em um computador pessoal para ser utilizada.

Ao abrir a ferramenta, o usuário encontra duas caixas de texto em branco, como exibido na Figura 1. Nelas, ele deverá colocar o par de textos paralelos que deseja alinhar. Coloca-se na caixa à esquerda o texto na sua língua original (chamado texto-fonte) e, na caixa à direita, coloca-se a tradução (chamado texto-alvo).

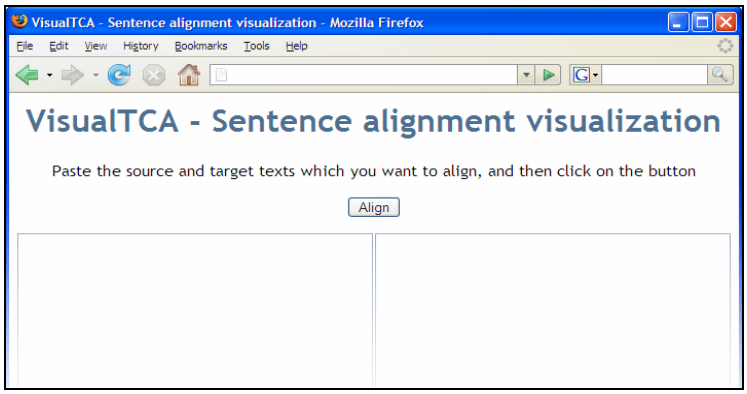

Figura 1. Tela inicial da ferramenta

Ao se acionar o alinhamento (clicando-se botão *Align*), os textos previamente especificados serão alinhados e exibidos na tela, lado a lado. Durante o processo de alinhamento, uma pequena janela irá aparecer no meio da tela. Ela irá informar ao usuário sobre o início do alinhamento e também sobre a categoria dos alinhamentos realizados pela ferramenta, tão logo eles forem produzidos. Ao término do processo, a janela informará ao usuário sobre a conclusão do trabalho e desaparecerá em alguns segundos.

A categoria do alinhamento diz respeito à forma como cada alinhamento se dá: um alinhamento direto indica que uma única sentença de um texto é considerada a tradução direta de uma única sentença no outro texto, sendo esta categoria a mais freqüente (referenciada por alinhamento 1-1); ocorrem também casos de omissões, como os alinhamentos (0-1) ou (1-0), em que uma sentença não corresponde a nenhuma na tradução, ou vice-versa; e casos de alinhamentos de contrações (por exemplo, 2-1) e expansões (1-2). Algumas categorias de alinhamentos são exemplificadas na Tabela 1.

Tabela 1. Categorias de alinhamento sentencial produzidos pela ferramenta

	Texto-fonte	Texto-alvo
1-1	Por terem acesso a fogão a gás, não buscam mais lenha na mata.	Having access to gas stoves, they no longer look for firewood in the forest.
1-0	As pesquisas avançam no laboratório.	*(frase não traduzida)*
0-1	*(frase não presente no texto fonte)*	It is the biggest city of the country.
2-1	Ela também possui ação analgésica de longa duração. Até o momento, quatro patentes foram requeridas.	It also has a long-lasting analgesic action, and up to now, four patents requests have been filed.
1-2	O pesquisador notou que os habitantes pescavam intensamente, e completavam a alimentação com frutas e milho.	The researcher noted that the inhabitants used to fish intensively. They would complete their diet with fruits and corn.

Com o processo concluído, os textos estarão completamente alinhados e exibidos na tela. Ao passar o ponteiro do mouse sobre as sentenças de qualquer um dos textos, o sistema destaca-as em uma cor de fundo diferente, juntamente com sua(s) sentença(s) correspondentes no outro texto. A Figura 2 ilustra esse processo.

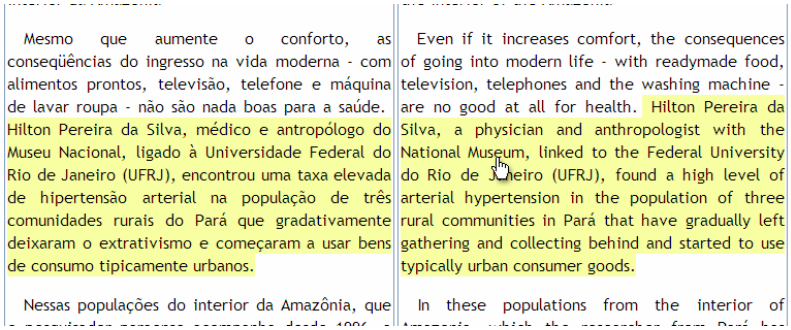

Figura 2. Visualizando os alinhamentos resultantes

Ao retirar o ponteiro do mouse de cima da sentença, ela volta a seu estado original, ou seja, sem cor de fundo. Caso se deseje que o destaque seja permanente nos textos, basta clicar na sentença desejada. Ela e seus alinhamentos correspondentes permanecerão destacados até que se clique nelas novamente. Dessa forma, uma pessoa pode visualizar como o alinhador automático realizou o alinhamento de cada sentença do par de textos.

Com a ferramenta, opcionalmente, é possível destacar todos os alinhamentos de uma mesma categoria simultaneamente. Isso é possível pelo uso de uma caixa de informações exibida na parte inferior da ferramenta. Essa caixa está ilustrada na Figura 3. Basta que se clique no pequeno quadrado colorido ao lado de cada categoria para que todas as sentenças pertencentes àquela categoria sejam destacadas com a cor correspondente; ao se clicar novamente no mesmo quadrado colorido, o destaque das sentenças daquela categoria de alinhamento será escondido, voltando ao seu estado original. A caixa também contém informações básicas sobre o par de textos alinhados (número de parágrafos, sentenças e palavras) e sobre a quantidade de cada categoria de alinhamento realizada.

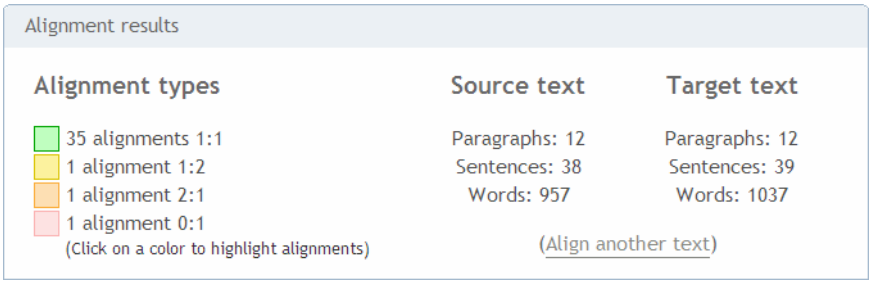

Figura 3. Caixa com informações sobre o resultado dos alinhamentos

Junto com a ferramenta, em sua tela inicial, é fornecido um par de textos de exemplo, com um texto em português e sua tradução em espanhol, caso o usuário deseje apenas conhecer rapidamente a ferramenta e/ou compreender como ela funciona. Também está disponível uma seção de ajuda, que conta com um texto explicativo com instruções passo a passo sobre sua utilização, juntamente com imagens ilustrando os pontos principais do processo.

O texto é submetido via web para a ferramenta e o resultado é exibido na tela do navegador de internet do usuário. A tarefa completa de alinhar os textos pode demorar de alguns segundos a alguns minutos, dependendo do tamanho dos textos. O método de alinhamento, subjacente à interface visual da ferramenta, apresenta complexidade linear em relação ao tamanho dos textos. Tal método é brevemente introduzido na seção seguinte.

3. Método de Alinhamento

O sistema de alinhamento automático utilizado pela VisualTCA é o TCAalign (Caseli, 2003), uma implementação do método TCA (*Translation Corpus Aligner*) (Hofland 1996), método este independente de língua que se utiliza de vários critérios para realizar o alinhamento, tais como correspondências baseadas em palavras cognatas, ocorrências de nomes próprios e caracteres especiais nos textos, pares de palavras-âncora (i.e., palavras cuja tradução de uma língua para outra é única, podendo servir de suporte opcional para o alinhamento), correlações entre comprimentos das sentenças, etc.

Este método foi escolhido para ser utilizado pela ferramenta por ser, dentre os vários métodos implementados no projeto de alinhamento textual PESA (*Portuguese-English Sentence Aligment*), o método que obteve os melhores resultados nas avaliações realizadas. Para os pares de língua português-espanhol e português-inglês, o TCA teve desempenho médio de aproximadamente 95%. Mais detalhes sobre o TCA e outros métodos de alinhamento podem ser encontrados em Caseli (2003).

4. Considerações finais

Apresentamos, nesse artigo, uma ferramenta visual on-line para o alinhamento sentencial automático de textos paralelos, a VisualTCA, que consideramos ser de interesse direto de pesquisadores da área de tradução, quer da Lingüística, quer da Lingüística Computacional. A ferramenta encontra-se hospedada e disponível na página do NILC, grupo de pesquisa em que foi desenvolvida.

Como trabalhos futuros, pretendemos estender a VisualTCA para também realizar o alinhamento lexical e permitir a edição humana, quer para corrigir alinhamentos inapropriados, quer para interferir e auxiliar no processo de alinhamento.

Agradecimentos

À FAPESP e ao CNPq, pelo apoio à realização desta pesquisa.

Referências

Caseli, H.M. (2003). *Alinhamento sentencial de textos paralelos português-inglês*. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

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

Véronis, J. (2000). From the Rosetta stone to the information society: A survey of parallel text processing. In J. Véronis (ed.), *Parallel text processing: Alignment and use of translation corpora*, pp. 1-24. Kluwer Academic Publishers.