PorTAl: Recursos e Ferramentas de Tradução Automática para o Português do Brasil

Thiago Lima Vieira, Helena de Medeiros Caseli

[1] Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brasil

{helenacaseli,thiago.lima.vieira}@dc.ufscar.br

Abstract. *This paper describes the machine translation (MT) site PorTAl developed aiming at integrating useful tools and resources for MT and the multilingual processing. Currently under development, the PorTAl will provide tools and resources for Brazilian Portuguese, English and Spanish (initially). In a near future we believe that the PorTAl will stimulate a progress in multilingual applications, particularly related to the Brazilian Portuguese.*

Resumo. *Este artigo descreve o portal de tradução automática (TA) PorTAl desenvolvido com o intuito de integrar ferramentas e recursos úteis para TA e o processamento multilíngue. O PorTAl, atualmente em desenvolvimento, envolverá a disponibilização de ferramentas e recursos para os idiomas português do Brasil, inglês e espanhol (inicialmente). A longo prazo, acredita-se que o PorTAl impulsionará um avanço nas aplicações de processamento multilíngue, principalmente no que diz respeito ao português do Brasil.*

1. Introdução

Desde sua concepção, há mais de setenta anos atrás, a Tradução Automática (TA) tem sido vista como uma das principais áreas do Processamento de Língua Natural (PLN) [Nirenburg et al. 1993]. Iniciada com metas ambiciosas de geração de uma tradução perfeita, para todos os idiomas e em quaisquer domínios, a TA é vista com muito mais cautela nos dias de hoje. As propostas iniciais de utilização de conhecimento linguístico profundo deram lugar a técnicas empíricas baseadas em exemplos e *corpora*. Na atualidade, o estado da arte na TA baseia-se em medidas estatísticas para se determinar qual é a melhor tradução (para uma língua alvo) dada uma sentença de entrada (em uma língua fonte).

Embora existam, atualmente, sistemas e ferramentas para a TA comerciais e gratuitos, disponíveis online, desenvolvidos seguindo várias abordagens (TA direta ou indireta, por transferência ou interlíngua) e vários paradigmas (TA baseada em regras, TA estatística, TA baseada em exemplos, etc.) ainda não é possível alcançar as ambiciosas metas de suas origens: produzir TA de boa qualidade em domínios irrestritos, por meio de sistemas completamente automáticos. Assim, apesar de muito beneficiada pelos avanços dos últimos anos, a TA ainda requer pesquisas para superar problemas linguístico-computacionais e tecnológicos sendo considerada por muitos pesquisadores uma área ainda carente de recursos e ferramentas [Pardo et al. 2009].

Nesse contexto, está inserido o Portal de Tradução Automática **PorTAl** no qual recursos e ferramentas úteis para a TA e outras aplicações multilíngues serão disponibilizados *online* para os idiomas português do Brasil, inglês e espanhol, inicialmente.

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 179–183,
Cuiabá, MT, Brazil, October 24–26, 2011.

2. O PorTAl

O Portal de TA (PorTAl) está sendo desenvolvido com o intuito de disponibilizar livremente na Web ferramentas e recursos úteis para a TA ou o processamento multilíngue, para usuários leigos ou especialistas da área de linguística e computação. Por meio do PorTAl, o usuário leigo poderá traduzir um texto ou fazer buscas em um léxico bilíngue, enquanto usuários especialistas poderão ir além, por exemplo, treinando um tradutor.

2.1. Recursos e Ferramentas

Entre os recursos inicialmente disponíveis para o PorTAl estão os *corpora* paralelos e os léxicos bilíngues para os pares português-inglês e português-espanhol. Um *corpus* paralelo é um conjunto de pares de textos em dois idiomas que são a tradução um do outro (textos paralelos). Tais pares de textos são usados como entrada para o treinamento de um modelo de tradução e outras ferramentas de processamento paralelo, como os alinhadores lexicais. A Tabela 1 exibe detalhes dos *corpora* inicialmente disponíveis para o PorTAl.

Tabela 1. *Corpora* inicialmente disponíveis para o PorTAl

Domínio	**português–inglês**		**português–espanhol**		**Total**
	português	inglês	português	espanhol	
Acadêmico	25.385	22.731	1.735	1.739	51.590
Literário	245.644	238.374	–	–	484.018
Jurídico	315.349	316.212	705.569	749.599	2.086.729
Jornalístico/Científico	1.010.196	1.070.286	503.715	546.570	3.130.767
TOTAL	**1.596.574**	**1.647.603**	**1.211.019**	**1.297.908**	**5.753.104**

Os léxicos bilíngues, podem ser entendidos como listas de pares de palavras e unidades multipalavras que são possíveis traduções. Até o momento foram gerados, automaticamente, léxicos bilíngues a partir dos *corpora* jornalístico-científico português-inglês (contendo cerca de 31.000 entradas em português e 24.000 em inglês) e português-espanhol (contendo cerca de 19.000 entradas em português e 21.000 em espanhol). Outros léxicos também serão gerados para outros domínios a partir dos *corpora* citados.

Entre as ferramentas acessíveis via PorTAl estão: os alinhadores sentencial TCAalign[1] [Caseli 2003] e lexicais LIHLA [Caseli et al. 2005] e GIZA++[2] [Och e Ney 2003]; os indutores de léxicos e regras do projeto ReTraTos[3] [Caseli et al. 2006]; e o *toolkit* de TA Moses[4] [Koehn et al. 2007]. Os alinhadores sentencial e lexical, respectivamente, determinam as correspondências de tradução entre sentenças e palavras (ou unidades multipalavra) fonte e alvo. Os indutores de regras e léxicos geram regras morfossintáticas de tradução e entradas para um léxico bilíngue, por meio de processos separados que têm como entrada um *corpus* paralelo alinhado lexicalmente e etiquetado morfossintaticamente. Por fim, o *toolkit* de TA Moses, a partir de um *corpus* alinhado sentencialmente, gera modelos estatísticos de tradução e língua que modelam a probabilidade de tradução fonte-alvo e de geração da sentença alvo, nesta ordem. O fluxo de processamento planejado para o PorTAl é ilustrado na Figura 1.

[1] `http://www.nilc.icmc.usp.br/nilc/projects/aligners.htm`
[2] `http://code.google.com/p/giza-pp/`
[3] `http://retratos.sourceforge.net/`
[4] `http://www.statmt.org/moses/`

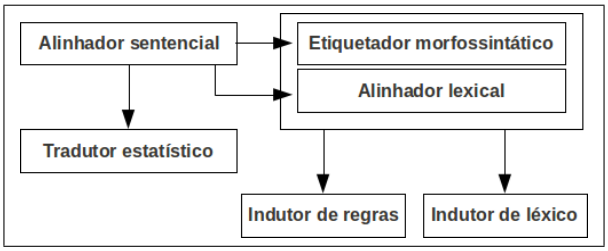

Figura 1. Fluxo de processamento das ferramentas do PorTAl

2.2. Desenvolvimento

Para o desenvolvimento do PorTAl, optou-se pela utilização de tecnologias capazes de torná-lo uma aplicação leve e, assim, deixar o servidor livre para o processamento das requisições das ferramentas de TA. Para tanto, selecionou-se tecnologias gratuitas como a linguagem de programação PHP, o *framework* de desenvolvimento Zend, o servidor Web Apache no Linux/Ubuntu Server, o banco de dados PostgreSQL e o *toolkit* Dojo.

A linguagem PHP e o servidor Web Apache foram selecionados por apresentarem um processamento mais sutil no servidor. Outras opções possíveis como Java para Web foram analisadas, porém descartadas pois sistemas com Java necessitam de servidores mais robustos em termos de *hardware*. Para agilizar o desenvolvimento em PHP optou-se por utilizar o *framework* gratuito Zend, que possui bibliotecas de rotinas comuns a aplicações Web, padrões de projetos, abstração do banco dados, entre outros. Com relação ao banco de dados (BD), optou-se pelo PostgreSQL que, além de ser gratuito como o MySQL, é considerado melhor em termos de performance (consultas em tabelas) e não corre o risco de ser extinto a curto prazo.[5]

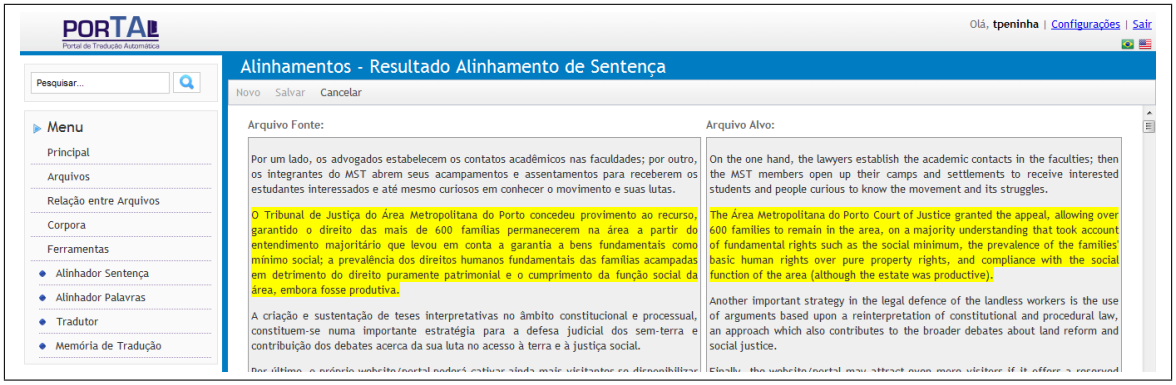

Figura 2. Demonstração do toolkit Dojo no PorTAl

Por fim, o *toolkit* Dojo está sendo utilizado para prover JavaScript e Ajax no PorTAl. O Dojo permite a criação de funcionalidades de forma mais simples, como se fosse uma camada sobre o JavaScript, além de possuir componentes de interface rica para Web (RIA, *Rich Internet Application*) que também podem ser explorados. O JavaScript e o Ajax surgem, nesse cenário, para permitir uma navegação mais amigável com a interface mais dinâmica e rápida. A Figura 2 demonstra como o Dojo pode ser utilizado, neste caso, para colorir as sentenças do alinhamento sentencial de textos paralelos.

[5]Desde a compra da *Sun Microsystem* (mantedora do MySQL) pela Oracle (proprietária do BD Oracle) a comunidade livre receia o fim do MySQL por não acreditar que a empresa manterá um BD gratuito para concorrer com o seu proprietário.

3. Trabalhos Relacionados

Entre os sistemas *online* de disponibilização de recursos e ferramentas de PLN similares ao PorTAl estão: o LacioWeb[6], o OntoLP[7], a Linguateca[8], o e-Termos[9], o PorSimples[10], e o LX-Center[11]. O LacioWeb [Aluísio et al. 2003] é um portal de *corpus* para disponibilização *online* e livre de *corpus* em português do Brasil, desenvolvido em PHP com servidor Web Apache. O OntoLP [Vieira et al. 2007] também é um portal desenvolvido em PHP que disponibiliza ontologias para diversas áreas. Além desses dois portais para um tipo específico de recurso, deve-se destacar o maior e mais variado repositório de recursos para o português: a Linguateca [Santos 2009]. Semelhante ao PorTAl, a Linguateca foi desenvolvida com PHP e JavaScript, e o servidor Web é o Apache.

Diferentemente dos trabalhos citados, nos quais o foco está na disponibilização de recursos prontos, o e-Termos [Oliveira 2009] é um sistema Web de gestão de terminologia que tem como ponto forte a construção colaborativa de bases terminológicas. Utiliza a linguagem PHP, JavaScript, o padrão Ajax, entre outros para desenvolver a interface de apresentação. Outras semelhanças com o PorTAl são o servidor Web Apache e sistema operacional Linux. O LX-Center [Branco e Silva 2004] é outro sistema Web que disponibiliza gratuitamente serviços para textos em português, como: lematizador verbal, etiquetadores de *part-of-speech*, *parsers* léxicos, concordanciador de *corpus*, entre outros. O LX-Center utiliza PHP e JavaScript. Por fim, o PorSimples [Aluísio e Gasperin 2010] é um sistema *online* cujo foco principal está na simplificação textual realizada, principalmente, por meio da sumarização e simplificação léxica e sintática dos textos escritos em português do Brasil. Em seu desenvolvimento foram utilizados PHP e JavaScript, com servidor Web Apache e sistema operacional Linux.

A partir desse breve relato é possível verificar que a maioria dos sistemas desenvolvidos até então apenas disponibiliza recursos e não oferece processamento interligado e *online* das ferramentas conforme esboço apresentado na Figura 1. Além disso, nenhum dos trabalhos citados está voltado especificamente para a TA e suas particularidades como o processamento paralelo (de textos que são tradução mútua) e multilíngue. Quanto às ferramentas para desenvolvimento, a maioria dos trabalhos citados utiliza PHP e Javascript, servidor Web Apache e sistema operacional Linux, todas opções adotadas para o PorTAl.

4. Considerações Finais

O PorTAl[12], apresentado neste artigo, visa a disponibilização de ferramentas e recursos *online* e de modo transparente para o usuário. Entre os benefícios de tal iniciativa podem ser citados: a padronização de recursos e ferramentas em várias pesquisas de TA ou outras áreas do PLN e, principalmente, o avanço nas pesquisas com o português do Brasil.

Agradecimentos

Os autores deste artigo agradecem à FAPESP pelo apoio financeiro (#2010/07517-0 e #2010/16239-4).

[6] `http://www.nilc.icmc.usp.br/lacioweb/`
[7] `http://www.inf.pucrs.br/~ontolp/index.php`
[8] `http://www.linguateca.pt/`
[9] `http://www.etermos.ufscar.br/`
[10] `http://caravelas.icmc.usp.br/wiki/index.php`
[11] `http://lxcenter.di.fc.ul.pt`
[12] `http://www.lalic.dc.ufscar.br/portal` – disponível a partir de dezembro de 2011.

Referências

Sandra M. Aluísio e Caroline Gasperin (2010). Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. *NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 1:46–53.

Sandra M. Aluísio, Gisele M. Pinheiro, Marcelo Finger, Maria das Graças V. Nunes e Stella E. O. Tagnin (2003). The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. *CORPUS LINGUISTICS*, páginas 14–21.

António Branco e João Silva (2004). Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. Em *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, páginas 507–510, Paris, ELRA.

Helena M. Caseli (2003). Alinhamento sentencial de textos paralelos português-inglês. Tese de mestrado, ICMC-USP.

Helena M. Caseli, Maria das Graças V. Nunes e Mikel L. Forcada (2005). Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. *Procesamiento del Lenguaje Natural*, 35:237–244.

Helena M. Caseli, Maria das Graças V. Nunes e Mikel L. Forcada (2006). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20:227–245.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin e Evan Herbst (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 19–51.

Sergei Nirenburg, Constantine Domashnev e Dean J. Grannes (1993). Two Approaches to Matching in Example-Based Machine Translation. Em *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, páginas 47–57, Leuven, Belgium.

Franz J. Och e Hermann Ney (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Leandro H. M. Oliveira (2009). e-Termos: Um ambiente colaborativo web de gestão terminológica. Tese de mestrado, ICMC-USP.

Thiago A. S. Pardo, Helena M. Caseli e Maria das Graças V. Nunes (2009). Mapeamento da Comunidade Brasileira de Processamento de Línguas Naturais. Em *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology*, páginas 1–21, São Carlos, SP, Brazil.

Diana Santos (2009). Caminhos percorridos no mapa da portuguesificação: A linguateca em perspectiva. *Linguamática*, 1(1):25–58.

Renata Vieira, Patrícia M. Pizzinato, Larissa A. de Freitas, Anderson Bestetti e Lucelene Lopes (2007). OntoLP – Portal de Ontologias.