Investigação e Desenvolvimento de uma Ferramenta de Pós-Edição de Traduções Automáticas

Felipe Tassario Gomes, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC) Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

felipc@grad.icmc.usp.br, taspardo@icmc.usp.br

1. Introdução

A Tradução Automática (TA) consiste em traduzir automaticamente um texto escrito em uma língua natural (chamada de línguafonte) para uma outra língua natural (línguaalvo), traduzindo todo seu conteúdo de maneira a gerar um resultado inteligível e que preserve suas características, como estilo, coesão e significado [1].

Apesar de ter surgido ainda na década de 40, sendo a primeira aplicação não numérica da Computação, a TA continua como uma tarefa difícil de ser implementada e aperfeiçoada, devido à necessidade de grande quantidade de conhecimento humano e de mundo para uma tradução perfeita.

Diante da dificuldade em melhorar a qualidade da TA em análises que envolvem conhecimento humano, a utilização de algumas etapas de revisão manuais pode beneficiar a qualidade da tradução. Uma deficiência da TA ocorre quando a tradução de uma frase não pode ser feita sem conhecimento humano implícito sobre a frase, como contexto histórico ou local. Outro erro comum ocorre quando tal conhecimento não pode ser inserido no tradutor automático, como o uso de termos específicos à área de conhecimento do texto a ser traduzido, e que não estão presentes num tradutor automático de propósito geral. Além disso, erros similares são recorrentes no uso de um tradutor automático, como erros específicos de concordância ou adequação verbal.

2. Objetivos

Neste projeto de iniciação científica foi criada uma ferramenta de revisão de textos traduzidos automaticamente. Tal ferramenta possibilita um tradutor humano a analisar de forma eficiente um texto traduzido automaticamente e a realizar alterações no texto para obter uma tradução correta semântica e gramaticalmente.

A ferramenta desenvolvida trata-se de um ambiente online de pós-edição textual que permite ao revisor fazer as modificações manuais necessárias no texto, e que possui a capacidade de armazenar as modificações realizadas, com o objetivo de sugerir e repetir tais modificações automaticamente nas próximas vezes que o erro em questão for encontrado no texto. A ferramenta também oferece outras facilidades, como o alinhamento sentencial entre as sentenças do texto original e do traduzido, um corretor ortográfico integrado, e a visualização das modificações realizadas pelo tradutor para chegar no estado final do texto revisado.

3. Desenvolvimento

Para a implementação da ferramenta, foram realizados estudos sobre TA e sobre as técnicas e motivações [2] relevantes a ferramentas de pré e pós-edição textuais [3]. Com base nesses estudos, foram desenvolvidos diversos módulos com as funcionalidades desejadas na ferramenta.

Desenvolveu-se um módulo de memória de correção, baseado nas ferramentas de memória de tradução. Este módulo é capaz de armazenar as correções feitas pelo usuário e reproduzi-las automaticamente em situações recorrentes.

Também foi desenvolvido um módulo de detecção de diferenças textuais [4], utilizado pela memória de correção para detectar as modificações palavra-a-palavra realizadas, além de ser utilizado pelo ambiente online para exibir ao revisor as edições já realizadas no texto. Além disso, foram implementados e integrados à interface o módulo de alinhamento sentencial e o corretor ortográfico, bem como um glossário automático capaz de corrigir palavras traduzidas incorretamente pelo tradutor automático.

A ferramenta e seus módulos foram desenvolvidos utilizando tecnologias web e apresentadas como um sistema online, que facilita o alcance da ferramenta entre os usuários, por não necessitar de nenhuma instalação ou software especial para sua utilização. O ambiente está disponível através do site do NILC, e seu código fonte disponibilizado através de uma licença de software livre.

4. Resultados

A avaliação da ferramenta foi realizada com base nas metodologias de avaliação existentes na literatura [5]: foram dados pares de texto (texto original e uma tradução automática) para diversos usuários, e foi analisado como a ferramenta os ajudou a detectar tais erros e a corrigi-los. As medidas tiradas foram a precisão em corrigir os erros detectados, a cobertura em detectar todos os erros presentes, e a medida-F como uma medida única que combina a precisão e a cobertura. Os resultados são apresentados na Tabela 1, com os valores médios obtidos pelos usuários, para dois sistemas de tradução automática diferentes.

Tabela 1: Resultados da avaliação

Tradutor	Precisão	Cobertura	Medida-F
Babelfish	0.814	0.863	0.837
Google	0.887	0.923	0.904

5. Conclusões

O ambiente desenvolvido demonstrou-se como uma boa ferramenta para a fase de correção e pós-edição de um texto traduzido automaticamente, com facilidades para auxiliar um usuário tanto em analisar os textos e fazer as correções que julgue necessárias, bem como oferecer a possibilidade de aprender as alterações feitas pelo usuário para poder sugeri-las automaticamente em uma nova edição.

6. Agradecimentos

Às agências de fomento à pesquisa FAPESP, CAPES e CNPq.

Referências Bibliográficas

- [1] Specia, L. e Rino, L.H.M. (2002). Introdução aos Métodos e Paradigmas de Tradução Automática. Série de Relatórios Técnicos do NILC, NILC-TR-02-04. São Carlos, Março, 23p.
- [2] Reeder, F; Loehr, D;. (1998). Finding the Right Words: An Analysis of Not-Translated Words in Machine Translation. In *Machine Translation and Information Soup*, pp. 256-263.
- [3] Simard, M.; Goutte, C.; Isabelle, P. (2007) Statistical Phrase-Based Post-Editing. In the Proceedings of Human Language Technology Conference, NAACL-HLT.
- [4] Hunt, J.W, Mcllory, M.D. (1976). An algorithm for differential file comparison. In *Computer Science Technical Report*.
- [5] Font Llitjos, A.; Carbonell, J. (2004). The Translation Correction Tool: English-Spanish User Studies. In the *Proceedings of REC-2004: Fourth International Conference on Language Resources and Evaluation*, pp. 347-350.