Um Tradutor Estatístico entre Português, Espanhol e Inglês

Wilker Ferreira Aziz, Ivandré Paraboni, Thiago Alexandre Salgueiro Pardo

Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

1. Objetivos

Este trabalho tem por objetivo o desenvolvimento e avaliação de um tradutor automático de abordagem estatística entre as línguas portuguesa, espanhola e inglesa.

2. Materiais e Métodos

A tradução automática (TA) estatística é baseada na obtenção do conhecimento necessário à tradução a partir de um conjunto de exemplos de traduções, o córpus paralelo. A partir de tal córpus obtém-se o modelo de tradução (TM), que consiste em um conjunto de medidas utilizadas para decidir como uma palavra ou um conjunto delas é traduzido da língua-fonte para a língua-alvo. Para o aprendizado é necessário identificar no córpus segmentos que são traduções mútuas. O alinhamento consiste na identificação de tais segmentos que podem ser obtidos em nível sentencial e lexical. A partir do TM e de um modelo de língua (LM), obtido através de um córpus da língua-alvo, faz-se a decodificação, que realiza a tradução em si.

Neste trabalho, escolheu-se as ferramentas TCAalign [1] para a obtenção do alinhamento sentencial, a ferramenta GIZA++ [2] para a obtenção do TM, o pacote CMU-Cambridge para a obtenção do LM e a ferramenta ReWrite Decoder desenvolvida no ISI/USC para a decodificação.

3. Resultados

Produziram-se corpora paralelos entre as línguas portuguesa (PB), espanhola (ES) e inglesa (IA) alinhados sentencialmente e revisados. Utilizando as ferramentas citadas, produziram-se protótipos de tradutores entre português e espanhol e entre português e inglês. Sobre um conjunto de teste, realizou-se a avaliação comparativa com outros tradutores com a métrica BLEU [3], como mostra a Tabela 1. Quanto mais próxima de 1, melhor a tradução.

Tabela 1. Score BLEU

Tradutor	PB-ES	ES-PB	Tradutor	PB-IA	IA-PB
Apertium	0.6248	0.5955	Google	0.3157	0.3941
Z*	0.5673	0.5734	Z*	0.1892	0.2228

*Protótipo desenvolvido por este projeto.

4. Conclusões

Os resultados da avaliação mostram que a tradução entre as línguas portuguesa e espanhola são mais simples, provavelmente devido à proximidade que apresentam. Da comparação com o sistema Apertium, baseado em conhecimento lingüístico, conclui-se que o paradigma estatístico apresenta grande potencial, já que um córpus maior provavelmente produziria melhores resultados. O resultado para português-inglês sugere a necessidade de um córpus maior e de melhores estratégias de tradução.

5. Agradecimentos

Conta-se com apoio FAPESP, CAPES e CNPq.

6. Referências Bibliográficas

[1] Caseli, H.M. Alinhamento sentencial de textos paralelos português-inglês. Dissertação de Mestrado. ICMC-USP, Abril. (2003)

[2] Och, F.J. and Ney, H. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol. 29, N. 1, pp. 19-51. (2003)

[3] Papineni, K.; Roukos, S.; Ward, T. and Zhu, W. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of he 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318. (2002)

[4] Aziz, W. F.; Pardo, T. A. S.; Paraboni, I. An Experiment in Spanish-Portuguese Statistical Machine Translation. SBIA 2008.