Um Tradutor Estatístico entre Português e Espanhol

Wilker Ferreira Aziz, Ivandré Paraboni, Thiago Alexandre Salgueiro Pardo

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

wilker@grad.icmc.usp.br, ivandre@usp.br, taspardo@icmc.usp.br

1. Objetivos

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

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. Por córpus paralelo, entende-se um conjunto de pares de textos que são traduções mútuas. A partir de tal córpus obtém-se o modelo de tradução, que consiste em um conjunto de medidas utilizadas para decidir como uma palavra ou um conjunto delas (*phrase*) é traduzido da língua-fonte para a língua-alvo. Para se obter o modelo de tradução, é necessário que o córpus possua algum nível de correspondência, como os alinhamentos sentencial e lexical, técnicas a partir das quais se obtêm as correspondências entre sentenças e palavras que são traduções mútuas. A partir do modelo de tradução e de um modelo de língua, obtido através de um córpus de textos da língua-alvo, processa-se a decodificação, que realiza a tradução em si. O modelo de língua avalia, através de probabilidades, se uma sentença está bem formada.

Para este trabalho decidiu-se utilizar técnicas automáticas de obtenção do alinhamento, sendo tais técnicas independentes de língua, para que não fosse necessário incorporar conhecimento lingüístico à tradução. Escolheram-se as ferramentas TCAalign (Caseli, 2003) para a obtenção do alinhamento sentencial e a ferramenta GIZA++ (Och e Ney, 2003) que obtém, a partir do alinhamento sentencial, o alinhamento lexical e o modelo de tradução a partir desse último. A GIZA++ implementa os modelos IBM (Knight, 1999) de tradução, que descrevem a tradução como um conjunto de probabilidades extraídas do córpus. Por fim, a produção de um tradutor envolve a obtenção do modelo de língua e a decodificação dos dados obtidos. Para tal finalidade, decidiu-se utilizar a ferramenta ReWrite Decoder desenvolvida no ISI/USC.

3. Resultados

Este trabalho produziu um córpus português-espanhol alinhado sentencialmente e revisado, contando com 908.533 palavras. Além disso, estudou-se o paradigma estatístico de tradução, os modelos IBM e o uso de ferramentas de apoio à TA.

4. Próximos passos

Como próximos passos, têm-se o treinamento do modelo de tradução a partir do córpus produzido, a obtenção do tradutor via ReWrite e a avaliação dos resultados da tradução.

5. Conclusões Parciais

A revisão do alinhamento sentencial corrigiu 9,43% dos alinhamentos, que representa um valor absoluto de 1.668 pares de sentenças que incluiriam erros no modelo de tradução.

6. Agradecimentos

Este trabalho conta com o apoio FAPESP, CAPES e CNPq.

7. Referências Bibliográficas

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

Knight, K. (1999). *A Statistical MT Tutorial Workbook.*

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