Extracção de Recursos Bilingues a Partir de Corpora Paralelos

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

SDDI 2007

Embora os tradutores usem desde há séculos recursos para ajuda à tradução como sejam dicionários, foi com o advento dos computadores que esses recursos se tornaram ainda mais importantes. Com a utilização de computadores os tradutores começaram a ter acessível uma grande quantidade de informação: não só os habituais dicionários se tornaram electrónicos, mas também as traduções já realizadas (pelos próprios tradutores ou por terceiros) começaram a ser reutilizadas.

Da mesma forma, a investigação na área da tradução automática tem vindo a crescer. No entanto, esta evolução não tem sido suficiente para que a tradução automática, por sí só, seja realmente útil. Deste modo, os investigadores desta área têm produzido recursos úteis para os seus sistemas de tradução mas que também podem ser aproveitados pelos tradutores humanos.

Nesta apresentação vai ser abordada uma metodologia baseada em métodos estatísticos para a extracção de dicionários de tradução, terminologia bilingue e exemplos de tradução, úteis não só para a tradução automática (seja ela baseada em regras, baseada em estatística ou baseada em exemplos) mas também para a tradução assistida por computador.

A extracção de dicionários (probabilísticos) de tradução foi trabalho já documentado na minha tese de mestrado. Será abordado superficialmente o algoritmo usado para contextualizar todo o restante trabalho.

A extracção de exemplos e de terminologia usa como base os dicionários probabilísticos de tradução para obter relacionamento entre palavras. Posteriormente é usado um mecanismo de padrões para detectar terminologia bilingue, e extrair exemplos de tradução. Serão mostrados exemplos de padrões, bem como exemplo da terminologia extraída, e uma avaliação da mesma.

Em paralelo será também demonstrada a forma como o Cluster SeARCH tem vindo a ser utilizado para acelerar todo este processo.