ALINHAMENTO SENTENCIAL E LEXICAL DE CÓRPUS PARALELOS: RECURSOS PARA A TRADUÇÃO AUTOMÁTICA

Helena de Medeiros CASELI (PG – USP)
Maria das Graças Volpe NUNES (USP)

Nos últimos anos, a utilização de textos paralelos – textos acompanhados de sua tradução em uma ou várias línguas – e textos paralelos alinhados – com marcas que identificam os pontos de correspondência entre o texto original e sua tradução – tem se tornado cada vez mais freqüente em inúmeras aplicações de Processamento de Língua Natural (PLN). Métodos automáticos de alinhamento de textos paralelos podem ser usados para gerar os córpus paralelos alinhados que são utilizados, principalmente, em aplicações de (i) tradução automática, (ii) recuperação de informações por meio da troca de dados entre línguas diferentes e (iii) aprendizado de idiomas. Os dois níveis de alinhamento mais estudados na literatura são o alinhamento sentencial e o lexical, nos quais são determinadas as correspondências entre as sentenças e as unidades lexicais (palavras ou multipalavras) do texto original (texto fonte) e de sua tradução (texto alvo). Este artigo apresenta o processo de construção de córpus paralelos e de alinhamento sentencial e lexical dos mesmos. Os córpus paralelos são compostos por textos de gêneros: científico (resumos e *abstracts* de trabalhos acadêmicos desenvolvidos no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo em São Carlos), jurídico (textos oficiais da Área de Livre Comércio das Américas, a ALCA) e jornalístico (artigos do jornal "*The New York Times*"). Os textos paralelos desses córpus foram alinhados automaticamente por vários métodos computacionais de alinhamento sentencial e lexical e uma versão corretamente alinhada também foi gerada por um especialista humano para servir como referência na comparação dos resultados retornados pelos métodos automáticos.