CONSTRUÇÃO DE UM CÓRPUS PARALELO ALINHADO PARA A TRADUÇÃO AUTOMÁTICA ESTATÍSTICA

Wilker Ferreira Aziz (ICMC-USP)
Thiago Alexandre Salgueiro Pardo (ICMC-USP)
Ivandré Paraboni (EACH-USP)

A necessidade da tradução de grandes quantidades de textos em pouco tempo estimulou, no passado, a pesquisa de métodos de Tradução Automática (TA). Atualmente, a Internet aprimorou o desafio devido a enorme quantidade de informação multilíngüe. A tradução entre pares de línguas distantes incentivou a busca por modelos estatísticos de TA, pois pouco conhecimento lingüístico é necessário nessa abordagem. Córpus paralelos são a base para o desenvolvimento desses modelos, pois, a partir deles é extraído automaticamente todo o conhecimento necessário para a tradução. Por córpus paralelo, entende-se um conjunto de textos em uma língua-fonte acompanhados das traduções na língua-alvo. Esses córpus podem ser alinhados lexicalmente e/ou sentencialmente, ou seja, são indicadas, em cada texto, as correspondências entre suas palavras e/ou sentenças com o texto na outra língua. Existem vários métodos para se produzir automaticamente o alinhamento entre textos paralelos, classificados como empíricos, lingüísticos ou híbridos, de acordo com o nível de conhecimento lingüístico que utilizam. Os métodos empíricos baseiam-se em medidas do texto, tais como número de palavras, quantidade de caracteres, ocorrência de caracteres especiais (letras maiúsculas e pontuação), similaridades e padrões, dispensando, assim, qualquer conhecimento lingüístico; os métodos lingüísticos, por sua vez, utilizam conhecimento sobre as línguas envolvidas, tais como glossários, léxicos, listas de elementos que freqüentemente são correspondentes entre as línguas e regras gramaticais, entre outros recursos; por fim, os métodos híbridos fazem uso de recursos de ambos os métodos anteriores para realizar o alinhamento. Há diversos córpus paralelos disponíveis para várias línguas, por exemplo, o córpus composto pelas cartas da ONU (Organização das Nações Unidas), traduzidas para os mais diversos idiomas. Da mesma forma, há diversas ferramentas de alinhamento textual, como o GIZA++ e o LIHLA, sendo que este último foi desenvolvido no NILC (Núcleo Interinstitucional de Lingüística Computacional), um dos maiores centros de pesquisa em Processamento de Línguas Naturais no Brasil. Para o desenvolvimento de bons tradutores automáticos estatísticos, faz-se necessário um grande córpus paralelo, em geral, com mais de 200 milhões de palavras, que, idealmente, deve conter alinhamentos corretos. Propõe-se, neste trabalho, a construção de um córpus paralelo alinhado representativo para o par de línguas português-espanhol, com a finalidade de desenvolver um tradutor automático para tais línguas. Posteriormente, pretende-se estender esse córpus para a língua inglesa. A metodologia para a construção desse córpus é: coleta dos textos e suas traduções a partir de fonte confiável, se possível, revisados por humanos; escolha de uma ferramenta de alinhamento sentencial, adequação dos textos ao formato de entrada exigido pela ferramenta e geração dos alinhamentos sentenciais; produzido o alinhamento sentencial, faz-se necessária, para garantia de resultados consistentes, a pós-edição/revisão dos alinhamentos (manualmente, em geral); escolha de uma ferramenta de alinhamento lexical, adequação dos textos alinhados sentencialmente ao formato de entrada exigido pela ferramenta e geração dos alinhamentos lexicais; pós-edição/revisão dos alinhamentos lexicais. Inicialmente, pretende-se construir o córpus com artigos da Revista FAPESP (publicada em português, espanhol e inglês). As ferramentas de alinhamento TCAalign e LIHLA deverão ser utilizadas neste trabalho. Ferramentas visuais podem ser usadas para facilitar o processo de revisão, tal como o VisualTCA, também desenvolvido no NILC. É importante dizer que, em princípio, o alinhamento lexical não se faz necessário para o desenvolvimento de tradutores estatísticos, mas é sabido que este recurso pode aprimorar os resultados obtidos. Com a disponibilização do córpus paralelo alinhado, há a possibilidade de se desenvolver tradutores estatísticos propriamente ditos, via o uso de toolkits disponíveis para a comunidade de pesquisa, por exemplo, o GIZA++ e o sistema Rewriter. Os córpus e os tradutores a serem construídos são a grande contribuição deste trabalho.