Construção de um Córpus Paralelo Alinhado para a Tradução Automática Estatística

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. Introdução

A necessidade da tradução de grandes quantidades de textos em pouco tempo e a enorme disponibilidade de informação multilíngüe na internet, principalmente, estimulou a pesquisa de métodos de Tradução Automática (TA). Atualmente, pesquisam-se muito os modelos estatísticos de TA, para os quais pouco conhecimento lingüístico é necessário. Tais métodos extraem o conhecimento necessário para a tradução de córpus paralelos, que são conjuntos de textos em uma língua-fonte acompanhados de suas traduções em uma língua-alvo.

2. Objetivos

Este trabalho tem por objetivo a construção de um córpus paralelo para o par português-espanhol alinhado sentencialmente e revisado. Esta pesquisa é motivada pelo desempenho da abordagem estatística de tradução. A partir do córpus alinhado sentencialmente, obtém-se o modelo de tradução; o modelo de tradução é aquele que rege como sentenças na língua-fonte tornam-se sentenças na língua-alvo. Para o treinamento de um modelo robusto, com medidas que reflitam bem o processo de tradução, é necessário que o córpus possua traduções e alinhamentos corretos.

3. Material e Métodos

Textos sem formatação foram obtidos da edição *on-line* da Revista Pesquisa Fapesp para a constituição do córpus.

Para o propósito desse trabalho, o córpus deve ser alinhado sentencialmente e devem-se indicar as correspondências entre as sentenças do texto-fonte e as sentenças do texto-alvo, ou seja, estabelecer os alinhamentos, que podem ser de vários tipos. Um alinhamento do tipo substituição é um alinhamento no qual uma sentença do texto alvo corresponde a exatamente uma sentença do texto fonte. Outras formas de alinhamento são possíveis, tais como a remoção (uma sentença que existe no texto fonte não possui uma correspondente no texto alvo), a inserção (situação oposta à anterior) e a união (situação na qual um número *n* de sentenças do texto fonte está associado a um número *m* de sentenças do texto alvo).

Escolheu-se para a realização do alinhamento sentencial a ferramenta independente de língua TCAalign, pois, em avaliações anteriores a este trabalho, obteve desempenho médio de 95% para as línguas português, espanhol e inglês (Caseli, 2003). O método em que a ferramenta se baseia realiza o alinhamento segundo medidas e pistas extraídas do texto, tais como palavras cognatas (identificadas através da busca por subseqüências de caracteres comuns), pontuação, posição de nomes próprios e siglas, quantidade de palavras e de caracteres por sentença, etc.

A ferramenta exige que os textos sejam segmentados. A segmentação sentencial consiste em separar cada sentença do texto de maneira que estejam dispostas em linhas individuais do mesmo. Na segmentação dos

textos em português foi utilizada a ferramenta SENTER (Pardo, 2006); para a segmentação dos textos em espanhol, este trabalho propôs e implementou uma nova versão do SENTER apta à segmentação de tais textos. Obtidos os textos segmentados, produziu-se automaticamente via TCAalign uma versão alinhada do córpus.

A necessidade de produção de um córpus confiável exige que os alinhamentos estejam corretos; disso decorre a necessidade de revisão. Foi proposto neste trabalho um método de identificação automática de erros no alinhamento sentencial para posterior correção manual. O método é baseado na revisão dos alinhamentos distintos do alinhamento do tipo substituição. A revisão manual mostrou que 9,43% dos alinhamentos estavam incorretos.

Foi escolhida, para o treinamento do modelo de tradução, a ferramenta GIZA++ (Och e Ney, 2003) que tem como exigência o alinhamento sentencial linha a linha dos textos, a remoção de toda a pontuação e que todas as palavras estejam em letras minúsculas. Para adequar o córpus à ferramenta, produziu-se automaticamente uma versão do mesmo de acordo com tais especificações.

4. Resultados

Um córpus paralelo para o par português-espanhol alinhado sentencialmente e revisado composto por 645 pares de textos e contando com 908.533 palavras, sendo 65.086 distintas, é a principal contribuição deste trabalho até o presente momento, embora outros resultados tenham sido obtidos: o próprio córpus paralelo português-espanhol que pode ser utilizado por trabalhos em Processamento de Línguas Naturais que envolvam esse par de línguas e uma nova versão do segmentador sentencial SENTER adequado ao espanhol.

5. Conclusões

A revisão do alinhamento sentencial corrigiu 9,43% dos alinhamentos, o que representa um valor absoluto de 1.668 pares de sentenças que iriam incluir erros no modelo de tradução ou que deixariam de fazer parte do córpus implicando uma significativa perda de informação, dado o tamanho do córpus produzido. Este trabalho é parte crucial do futuro desenvolvimento de um tradutor estatístico a ser produzido, de sua avaliação e da comparação com outros tradutores, como o sistema Apertium (Corbí-Bellot et al., 2005).

6. Agradecimentos

Este trabalho contou com o apoio das agências de fomento à pesquisa FAPESP, CAPES e CNPq.

7. Referências Bibliográficas

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

Corbí-Bellot, A.M.; Forcada, M.L.; Ortiz-Rojas, S.; Pérez-Ortiz, J.A.; Ramírez-Sánchez, G.; Sánchez-Martínez, F.; Alegria, I.; Mayor, A.; Sarasola, K. (2005). An open-source shallow-transfer machine translation engine for the romance languages of Spain. In the *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, pp. 79-86.

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

Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.