Aprimoramento do sistema Apertium de tradução automática

Felipe Tassario Gomes, Thiago Alexandre Salgueiro Pardo
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

1. Objetivos

Nesse projeto, visou-se realizar um estudo de um sistema de Tradução Automática (TA) de código aberto, o Apertium [1]. O tradutor é um sistema superficial, com capacidades de análise morfológica e sintática, e de realização de transferências gramaticais através de regras de tradução. O tradutor, além de ter seu código fonte disponibilizado, também disponibiliza seus dados lingüísticos. Nesse trabalho, testou-se a possibilidade de utilização de dados lingüísticos produzidos automática e semi-automaticamente, utilizando-se os dados do projeto ReTraTos [2], capaz de inferir léxicos bilíngües e regras de tradução a partir da análise de um córpus paralelo.

2. Material e método

O aluno realizou um estudo sobre TA e Processamento de Línguas Naturais. Como um dos resultados do estudo, criou-se uma ferramenta visual on-line para realizar e visualizar alinhamentos sentenciais entre pares de textos, tarefa que consiste em identificar as correspondências entre um texto e sua tradução. Após isto, iniciou-se a compilação de recursos lingüístico-computacionais para serem utilizados no projeto, tais como córpus paralelos de textos em português e espanhol e léxicos bilíngües. Após a compilação dos recursos, procedeu-se à introdução no Apertium dos novos léxicos induzidos automaticamente. Criou-se três configurações do sistema de tradução: o sistema original, sem modificações (configuração C1); o sistema híbrido, mantendo seus dados originais e os novos léxicos (C2); apenas os novos léxicos (C3). Cada configuração foi testada com um córpus de 648 sentenças com a métrica BLEU [3]. Verificou-se também o número de palavras não traduzidas.

3. Resultados

Os resultados são apresentados na Tabela 1, em que uma maior nota representa uma melhor tradução (nota varia de 0 a 1).

Tabela 1. Resultados da avaliação

	C1	C2	C3
Nota BLEU	0.45	0.49	0.32
Palavras não traduzidas	1.911	817	1.210

4. Conclusões

Pode-se notar que a combinação dos dados manuais com os induzidos automaticamente produziu os melhores resultados. Com esses resultados, pode-se concluir que a indução automática de léxicos pode auxiliar a produção de um sistema de tradução, tendo em vista que ela ajudou a melhorar a nota do sistema (no caso C2), e que, apesar de não apresentar uma nota tão alta sozinha (caso C3), apresentou uma boa cobertura do léxico a ser traduzido.

5. Agradecimentos

Às agências de fomento à pesquisa FAPESP, CAPES e CNPq.

6. Referências bibliográficas

[1] 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.

[2] Caseli, H.M. (2007). *Indução de léxicos bilíngües e regras para a tradução automática.* Tese de Doutorado. ICMC-USP. Abril. 158 p.

[3] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W-J. (2002). BLEU: a method for automatic evaluation of ma-chine translation. In the *Proceedings of the 40th Annual Meeting of the ACL*, pp. 311-318.