Tratamento de Informações Redundantes e de Senso Comum

Vinícius R. Uzêda[1], Thiago A. S. Pardo[1], Sandra M. Aluísio[1]
[1]Instituto de Ciências Matemáticas e da Computação (ICMC), USP/São Carlos

1. Objetivos

Segundo pesquisa, em 2005, o conhecimento disponibilizado na Internet não podia ser apreendido por 14,9 milhões de brasileiros com 15 anos ou mais, consideradas analfabetas, algo em torno de 11% da população [1]. A UNESCO propôs o termo analfabeto funcional (AF) para denominar o conjunto de pessoas que, apesar de entenderem a maior parte dos símbolos lingüísticos, não são capazes de fazer uso da língua escrita satisfatoriamente em seu dia a dia. Essa definição abrange cerca de 25% da população.

O projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) propõe tornar acessível o conteúdo disponibilizado na Internet para um grupo maior de potenciais leitores, utilizando recursos de Simplificação Textual (ST) para tanto. Este trabalho, dentro do escopo desse projeto, ambiciona remover informações redundantes e de senso comum, desnecessárias para o leitor em geral, mas, em especial, complicadoras para a leitura por AFs, que já possuem uma memória de trabalho limitada [2].

2. Material e Métodos

Para se determinar quais sentenças serão consideradas redundantes, algumas métricas de semelhança sentencial foram propostas, essencialmente baseadas na medida cosseno [3], identificando quantas palavras idênticas ou sinônimas existem e quão estruturalmente semelhantes são duas sentenças de um texto.

Para detectar a presença de senso comum, utilizou-se o banco de dados OMCS-Br [4]. A proposta é identificar quanto das informações veiculadas por uma dada sentença é inovadora, o que justificaria mantê-la no texto.

Com este conjunto de medidas, construiu-se uma matriz de similaridade, utilizada em uma etapa de agrupamento de sentenças similares, potencialmente redundantes entre si [5]. Um posterior processo de seleção detecta quais sentenças deverão ser removidas.

3. Resultados

O sistema ainda está em etapa de finalização, mas alguns experimentos mostraram que a ferramenta é capaz de identificar sentenças anotadas como redundantes. Uma avaliação mais rigorosa ainda deverá ser realizada, mas os resultados parecem promissores.

4. Conclusões

O uso de medidas estatísticas simples parece ser suficiente para identificar informações redundantes e, com um recurso relativamente simples como um banco de informações de senso comum, pode-se tornar textos menos redundantes, tarefa considerada importante na disponibilização do conteúdo digital para AF.

5. Referências Bibliográficas

[1] IBGE, Estudos & Pesquisas - Informação Demográfica e Socioeconômica, V. 19, disponível em http://www.ibge.gov.br com acesso em Maio de 2009 (2006).

[2] A. Siddharthan, Syntactic Simplification and Text Cohesion. Tese de PhD, University of Cambridge (2003).

[3] G. Salton, Automatic text processing. Addison-Wesley Longman Publishing Co, Inc, Boston, MA (1988).

[4] J.C. Anacleto, H. Lieberman, M. Tsutsumi, V.P.A. Neris, A.F.P. de Carvalho, J. Espinosa, S. Zem-Mascarenhas, M.S. Godoi, Can common sense uncover cultural differences in computer applications? In BRAMER, M. (Org) Artificial Intelligence in Theory and Practice - WCC 2006, 1 ed. Berlin: Springer-Verlag, v. 217, p. 1-10 (2006).

[5] E.R.M. Seno e M.G.V. Nunes, Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In: *Proceedings of the International Conference on Computational Processing of Portuguese Language - PROPOR* (Lecture Notes in Artificial Intelligence, 5190), pp. 133-144. Aveiro, Portugal, September (2008).