PROPOSTA DE UM SISTEMA PARA TRATAMENTO DE INFORMAÇÕES REDUNDANTES E DE SENSO COMUM

Vinícius Rodrigues de Uzêda, Thiago Alexandre Salgueiro Pardo, Sandra Maria Aluísio
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos
Núcleo Interinstitucional de Lingüística Computacional (NILC)
vruzeda@grad.icmc.usp.br, taspardo@icmc.usp.br, sandra@icmc.usp.br

Resumo

Este trabalho visa tornar acessível conteúdo textual da Internet (como jornais eletrônicos) à analfabetos funcionais e pessoas com disfunções neurológicas que apresentem dificuldade de leitura. A proposta é remover informações redundantes ou de senso comum, bem como utilizar este mesmo senso comum no esclarecimento de significado de termos no texto.

1. Introdução

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.

2. Objetivos

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]. Ainda, inserindo-se fatos de senso comum, pretende-se auxiliar o leitor a captar qual o significado de determinados termos (ou trechos) do texto.

3. Materiais 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.

O processo de inserção de informações de senso comum com a finalidade de explicitar o significado de passagens do texto ainda está sendo estudado. A proposta é gerarem-se sentenças simples através do uso de moldes [6] simples para certos tipos de informações sobre os substantivos do texto, como definições (*is-a*) e sugestões de local onde encontrar (*location-of*).

4. Resultados e Experimentos

O sistema ainda não está completo, não sendo possível ainda ter-se uma avaliação completa da proposta. Esse estudo, na realidade, deverá contar ainda com uma avaliação com usuários reais, mas numa etapa ainda mais avançada da avaliação.

Experimentos iniciais revelaram que o sistema tem capacidade de encontrar sentenças rotuladas manualmente para remoção, mas a abordagem para detectar e remover tais sentenças ainda está sendo estudada. A Figura 1 mostra um texto exemplo, extraído da Folha de São Paulo online, utilizado

nesses primeiros testes. As sentenças marcadas em negrito foram consideradas redundantes entre si. A Figura 2 mostra as sentenças em ordem decrescente de redundância, ou seja, a primeira sentença exibida, a de número 3, seria a primeira a ser removida pelo sistema. Assim, logo em uma primeira iteração, o sistema resolveria o problema de redundância encontrado.

1: O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.
2: Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador -- Cahe descartou pancreatite ou úlcera.
3: "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".
4: Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.
5: Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.

Figura 1. Texto Exemplo

3 > 1 > 2 > 4 > 5

Figura 2. Sentenças em ordem de redundância

O processo de inserção de informações de senso comum, ainda em etapa muito primordial, ainda não possui quaisquer resultados.

5. Conclusões parciais

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.

Referências

[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).

[6] E. Reiter, Robert Dale, Building applied natural language generation systems. Natural Language Engineering, 3(1):57-87. (1997)